Évaluation adaptative des systèmes de transcription en contexte applicatif

Mohamed Amer Ben Jannet

NNT :2015SACLS041

THESE DE DOCTORAT
DE
L'UNIVERSITE PARIS-SACLAY
PREPAREE A
"LABORATOIRE NATIONALE DE METROLOGIE ET D'ESSAIS"

ECOLE DOCTORALEN° 580

ED STIC : Sciences et technologies de l'information et de la communication

Doctorat en informatique

Par

Mohamed Ameur Ben Jannet

Titre de la thèse
Evaluation adaptative des systèmes de transcription en contexte applicatif

Thèse présentée et soutenue à la salle de conférence du LIMSI-CNRS , le 14 Octobre 2015
Composition du Jury :

M. Bechet Frédéric	Professeur à l'université Aix Marseille, LIF	Président
M. Besacier Laurent	Professeur à l'université J. Fourier, LIG	Rapporteur
M. Estève Yannick	Professeur à l'université du Maine, LIUM	Rapporteur
M. Zweigenbaum Pierre	Directeur de recherche CNRS, LIMSI	Examinateur
M^{me}. Rosset Sophie	Directrice de recherche CNRS, LIMSI	Directeur de thèse
M. Galibert Olivier	Ingénieur-Docteur, LNE	Co-directeur de thèse
M^{me}. Adda-Decker Martine	Directrice de recherche CNRS, LPP	Invité

Résumé

Il est important d'évaluer régulièrement les produits de l'innovation technologique afin d'estimer le niveau de maturité atteint par les technologies et d'étudier les cadres applicatifs dans lesquels elles pourront être exploitées. Le traitement automatique des langues (TAL) relève à la fois de la recherche et de l'innovation technologique et a pour but la modélisation et le développement d'outils permettant de traiter automatiquement le langage naturel. Pendant longtemps, les différentes briques technologiques issues du TAL étaient développées séparément. Par conséquent, les méthodes d'évaluation existantes sont dans la plupart modulaires et ne permettent d'évaluer qu'un seul module à la fois, alors qu'aujourd'hui nombreuses applications nécessitent de combiner plusieurs modules de TAL pour résoudre des tâches complexes. Le nouveau défi en terme d'évaluation est alors de pouvoir évaluer les différents modules (ou briques) tout en prenant en compte le contexte applicatif. Notre travail porte sur l'évaluation des systèmes de reconnaissance automatique de la parole (RAP) en contexte applicatif, en particulier, celui de la reconnaissance d'entités nommées (REN). En première partie, nous abordons la problématique de l'évaluation des systèmes de RAP en contexte applicatif à travers une étude de l'état de l'art. Nous y décrivons les tâche de RAP et de REN proposées dans les campagnes d'évaluation ainsi que les protocoles mis en place pour leurs évaluation. Nous y discutons également les limites des approches d'évaluations modulaires et nous y exposons les mesures alternatives proposées dans la littératures. En deuxième partie, nous décrivons la tâche de détection, classification et décomposition d'entités nommées étudiée et nous proposons une nouvelle métriques ETER (Entity Tree Error Rate) permettant de prendre en compte les spécificité de cette tâche et le contexte applicatif lors de l'évaluation. ETER permet également de supprimer les biais observés avec les métriques existantes. En troisième partie, nous définissons une nouvelle mesure ATENE (Automatic Transcriptions Evaluation for Named Entities) qui permet d'évaluer la qualité des systèmes de RAP et l'impact de leurs erreurs pour des systèmes de REN appliqués en aval. ATENE consiste à comparer les probabilités de présence d'entités sur les transcriptions de référence et d'hypothèse plutôt qu'une comparaison directe des graphèmes. Elle est composée de deux mesures élémentaires. Une première permettant l'évaluation de risque d'erreur d'omission et de substitution d'entités et une seconde permettant d'évaluer le risque d'erreur d'insertion d'entités causé par

les erreurs de RAP. Nos expériences de validation montrent que les mesures données par ATENE corrèlent mieux que les autres mesures de l'état de l'art avec les performances des systèmes de REN.

Mots-clefs : Reconnaissance automatique de la parole, évaluation, ETER, ATENE

Abstract

Title : Titre en anglais

It is important to regularly assess the technological innovation products in order to estimate the level of maturity reached by the technology and study the applications frameworks in which they can be used. Natural language processing (NLP) aims at developing modules and applications that automatically process the human language. That makes the field relevant to beth research and technological innovation. For years, the different technological modules from the NLP were developed separately. Therefore, the existing evaluation methods are in most modular. They allow to evaluate only one module at a time, while today, many applications need to combine several NLP modules to solve complex tasks. The new challenge in terms of evaluation is then to evaluate the different modules while taking into account the applicative context. Our work addresses the evaluation of Automatic Speech Recognition (ASR) systems according to the applicative context. We will focus on the case of Named Entities Recognition (NER) from spoken documents transcriped automatically. In the first part, we address the issue of evaluating ASR systems according to the application context through a study of the state of the art. We describes the tasks of ASR and NER proposed during several evalution campaigns and we discuss the protocols established for their evaluation. We also point the limitations of modular evaluation approaches and we expose the alternatives measures proposed in the literature. In the second part we describe the studied task of named entities detection, classification and decomposition and we propose a new metric ETER (Entity Tree Error Rate) which allows to take into account the specificity of the task and the applicative context during the evaluation. ETER also eliminates the biases observed with the existing metrics. In the third part, we define a new measure ATENE (Automatic Transcriptions Evaluation for Named Entities) that evaluates the quality of ASR systems and the impact of their errors for REN systems applied downstream. Rather than directly comparing reference and hypothesis transcriptions, ATENE measure how harder it becames to identify entities given the differences between hypothesis and reference by comparing an estimated likelihood of presence of entities. It is composed of two elementary measurements. The first aims to assess the risk of entities dele-

tions and substitutions and the second aims to assess the risk of entities insertions caused by ASR errors. Our validation experiments show that the measurements given by ATENE correlate better than other measures from the state of the art with the performance of REN systems.

Keywords : Automatic Speech Recognition, evaluation, ETER, ATENE.

Remerciements

Je voudrais remercier les membres de mon jury qui m'ont fait l'honneur d'accepter d'examiner ce travail. Je tiens à remercier Laurent Besacier et Yannick Estève qui ont accepté d'être rapporteurs. Merci pour votre lecture attentive et vos suggestions constructives. Je remercie Pierre Zweigenbaum et Frédéric Béchet pour avoir accepté de faire partie de jury et de consacrer du temps pour examiner ce manuscrit.

Merci à Sophie Rosset et Olivier Galibert qui ont toujours été disponibles pour m'aider et me conseiller. Merci pour votre rigueur scientifique et pour avoir lu et corrigé ce manuscrit avec beaucoup d'attention. Ces trois années ont été riches de travaux et d'émotions et c'est ce qui les a rendu aussi précieuses. Merci à Martine Adda-Decker pour ses encouragements et pour ses conseils très précieux.
Merci à Juliette Kahn pour ses conseils et pour les discussions passionnantes que nous avons pu avoir.

Je tiens à remercier l'ANRT et le LNE pour avoir financé ce travail de thèse et m'avoir donné l'occasion de me consacrer pleinement à ma thèse.
Merci à Bernard Pique et Stéphane Jourdin pour le soutien qu'ils m'ont accordé tout au long de ma thèse.
Merci à Patrick Paroubek et Guillaume Bernard pour leurs conseils et encouragements dans les moments difficiles.
Je tiens à remercier les membres du LIMSI et du LNE pour leur accueil et leur bonne humeur. Merci à Brigitte, Anne, Vincent, Munshi, Camille, Sami, Marine, Rémy...

Enfin, merci à tous ceux qui, de près ou de loin, m'ont encouragé et soutenu tout au long de ce travail. Une pensée pour tous, et un mot pour ma famille, ma mère, mon frère, mes soeurs et tous mes amis.

Table des matières

Liste des tableaux

Table des figures

Introduction générale

Dans ce travail de thèse, nous nous sommes intéressé à la thématique de l'évaluation des systèmes issus de la recherche technologique dans le domaine du traitement automatique des langues (TAL). Ce document met l'accent sur l'importance du processus d'évaluation dans le cycle de production des produits innovants et discute, à travers une lecture de l'état de l'art, des besoins émergents dans ce domaine. Le document présente également de nouvelles mesures d'évaluation qui visent à répondre aux besoins identifiés.

L'évaluation pour promouvoir l'innovation et la recherche technologique

La recherche technologique est le moteur de l'innovation. Elle permet de générer les connaissances théoriques et pratiques offrant la possibilité de développer des solutions technologiques répondant à des besoins sociétaux, économiques et industriels. Ces produits technologiques innovants sont essentiels pour le développement d'une industrie compétitive. Ils lui apportent de la valeur, ouvrent de nouveaux marchés et participent à la croissance économique. Il est donc très important d'évaluer régulièrement les produits de l'innovation technologique, afin d'estimer le niveau de maturité atteint par les technologies et d'étudier les cadres applicatifs dans lesquels elles pourront être exploitées, et avec quelle efficacité. Ceci permet de mettre en place une stratégie de développement efficace à long terme. .

L'intérêt de l'évaluation dépend du rôle joué par l'évaluateur dans le cycle de vie de la technologie. Ainsi, l'évaluateur peut être chercheur, développeur, consommateur, entrepreneur..., avec, dans chaque cas, des attentes différentes.

Pour les chercheurs et les développeurs, l'évaluation est un outil essentiel qui leur permet d'observer l'évolution des performances de leurs systèmes en fonction des adaptations et des modifications qu'ils y apportent. L'évaluation guide ainsi les développeurs dans leur quête d'optimisation et d'amélioration.

Pour les consommateurs, les mesures de performance affichées par les métriques permettent d'identifier les produits qui satisferont au mieux leurs besoins.

Paradigmes d'évaluation

Nous pouvons distinguer trois types de paradigmes d'évaluation selon le besoin de l'évaluateur et l'étape dans laquelle se trouve le produit par rapport à son cycle de développement.

Évaluation diagnostique : cette méthode d'évaluation consiste à soumettre un produit à un ensemble de tests bien ciblés dans le but d'identifier ses points forts et ses points faibles. Elle est principalement utilisée par les développeurs qui cherchent à identifier et à corriger les failles de leurs produits, mais peut également intéresser les utilisateurs, particulièrement dans les cas ou des caractéristiques particulières sont requises. Dans le cas des technologies de traitement automatique des langues, ce paradigme nécessite généralement la construction de données de test. Ces données doivent être représentatives de la tâche et serviront à tester les systèmes dans les conditions les plus proches de ces cas d'utilisation.

Évaluation comparative : appelée aussi étalonnage de performances ou encore *benchmarking*, cette méthode consiste à comparer un ensemble d'acteurs d'un domaine sur la base de critères fixés à l'avance. Elle est aujourd'hui largement utilisée que ce soit dans le secteur privé ou dans le secteur public, et ce, dans tous les domaines : monde de l'entreprise, de l'éducation, de la santé, de la recherche... Le *benchmarking* peut être considéré comme un outil d'auto-évaluation et d'aide à la décision qui permet de sélectionner parmi les pratiques et les méthodologies d'un domaine celles qui sont les plus performantes et les plus efficaces. Les meilleures solutions sont ainsi considérées comme des étalons de référence. En se comparant à ces étalons, les autres acteurs peuvent combler leurs retard et se rattraper [Bruno, 2009].

Évaluation de la pertinence : ce type de paradigme d'évaluation a pour but de déterminer l'adéquation d'un produit technologique avec son but usager. Est-ce qu'il répond à des besoins ? Avec quelle efficacité et à quel prix ? Ce type d'évaluation s'adresse principalement à des utilisateurs potentiels et demande une étude préalable afin d'identifier les besoins des consommateurs.

La métrologie, pilier de l'évaluation

Quel que soit le paradigme utilisé, le résultat de l'évaluation dépend nécessairement du choix des critères et des outils sur lequel se fonde l'évaluation. Un choix inadapté des critères ou des outils aura des conséquences graves sur les résultats de l'évaluation et sur leur interprétation. Dans le cas de la recherche technologique, un mauvais choix peut conduire à l'adoption d'une mauvaise stratégie ou à la suspension du financement d'un domaine de recherche comme cela a été le cas pour la traduction automatique dans les années soixante. Pour que les différents acteurs de la recherche technologique puissent profiter des bienfaits de l'évaluation, il est d'abord très important de s'assurer que les outils d'évaluation utilisés

sont bien adaptés aux cadres des évaluations. La métrologie est la science qui permet de maîtriser les processus des mesures en étudiant les aspects théoriques et pratiques d'un domaine d'application. Elle permet la mise en place et la diffusion des protocoles et des pratiques à suivre pour élaborer une évaluation. Elle permet aussi de développer et de fournir les outils (métriques) adaptés suivant le cadre et les objectifs de l'évaluation. La métrologie joue ainsi un rôle clé pour la recherche technologique et constitue un levier pour l'innovation, qu'il faut soutenir et renforcer.

L'évaluation en traitement automatique des langues

Le traitement automatique des langues est un domaine interdisciplinaire à l'intersection de l'informatique, de la linguistique et de l'intelligence artificielle. Il relève à la fois de la recherche et de l'innovation technologique et a pour but la modélisation et le développement d'outils permettant de traiter automatiquement le langage naturel. Parmi les principales briques technologiques émergentes on trouve :

- la traduction automatique ;
- la reconnaissance automatique de la parole ;
- la reconnaissance automatique du locuteur ;
- l'extraction d'information ;
- la compréhension automatique de texte ;
- la génération automatique de texte.

L'évaluation n'a pas toujours été bien perçue au sein de la communauté du TAL, notamment à cause du rapport ALPAC de 1966 [Nirenburg *et al.*, 2003] qui a jugé que les systèmes de traduction automatique, développés à l'époque, étaient très coûteux et peu performants [Paroubek, 2013]. Les critiques de ce rapport, estimé très sévère, [King, 1984] ont eu comme impact le gel des financements de la recherche en TAL aux États-Unis et ailleurs dans le monde pendant un certain nombre d'années. Il faut attendre 1987 [Pallett, 2003] pour voir se profiler un revirement qui arrivera encore une fois des États-Unis. En cette année-là, le NIST avait organisé une campagne d'évaluation ouverte à tous les acteurs de la recherche, autour de la reconnaissance automatique de la parole. Cette campagne d'évaluation avait connu un grand succès et de nombreux organismes de recherche de différents pays avaient participé. Depuis, de nombreuses campagnes d'évaluation dans différents domaines du TAL, se sont succédées. Aujourd'hui, les bienfaits de l'évaluation sur l'évolution du domaine du traitement automatique des langues sont bien reconnus. Les campagnes d'évaluation constituent une occasion qui permet de réunir les chercheurs de la communauté autour d'une tâche (ou d'une problématique) pour discuter et comparer leurs méthodes afin de sélectionner les plus avantageuses. Elles permettent aussi la mise à disposition et la création de

précieuses ressources (données d'entraînement et de test) nécessaires au développement de la recherche, ainsi que des paradigmes et des outils d'évaluation permettant d'analyser les sorties des systèmes afin de dégager les forces et les faiblesses de chacun. En outre, organiser régulièrement des campagnes d'évaluation offre la possibilité d'établir un suivi sur l'évolution des technologies, d'identifier les nouveaux enjeux et ainsi de mettre en place des stratégies de recherche à long terme.

Contexte et motivations

Le TAL n'est pas un domaine de recherche récent. Amorcé depuis les années cinquante, notamment aux États-Unis, il s'est peu à peu élargi, aidé notamment par l'évolution des puissances de calcul des machines et des capacités de stockage des données. Pendant longtemps, le développement des différentes briques technologiques issues du TAL se faisait indépendamment les unes des autres. Par conséquent, les outils et les mesures d'évaluation existants sont modulaires dans la plupart des cas (ils ne permettent d'évaluer qu'un seul module à la fois). Les évaluations modulaires sont particulièrement importantes pendant les premières phases du cycle de développement des technologies, puisque elles offrent des mesures spécifiques permettant de diagnostiquer et d'améliorer les fonctionnalités, poussant ainsi les produits vers la maturité. Aujourd'hui, de nombreux systèmes issus des recherches menées en TAL ont apporté la preuve de leur efficacité. Ces systèmes, même s'ils ne sont pas encore totalement parfaits, peuvent être intégrés dans des systèmes plus complexes où plusieurs niveaux de traitement et d'analyse doivent être enchaînés. Par exemple, les systèmes de dialogue oral homme-machine [Bellegarda, 2014] intègrent la reconnaissance de la parole, la compréhension de la parole, la gestion de dialogues, etc. De même, ceux pour la traduction automatique de la parole intègrent un traitement de la reconnaissance automatique de la parole et de la traduction automatique. Le développement de telles applications, que ce soit dans le cadre de défis lancés par les campagnes d'évaluation ou pour le développement d'applications destinées à différents marchés, a fait apparaître un nouveau besoin en termes d'évaluation qui est celui de l'évaluation selon le contexte applicatif.

En effet, chaque module utilisé a un impact sur le module suivant, une erreur pouvant en entraîner une autre. Ainsi, il ne s'agit pas de développer un système pour qu'il atteigne le meilleur score dans l'absolu, mais bel et bien de développer un module qui va contribuer à la performance globale du système complexe. Dans ce contexte, disposer d'outils d'évaluation permettant d'évaluer l'impact de chaque brique sur les briques suivantes et sur les fonctionnements de l'application globalement est très important. Que ce soit pour des utilisateurs désirant s'approprier la meilleure combinaison de modules pour leurs applications ou pour des déve-

loppeurs motivés par l'optimisation de leurs systèmes selon les cas d'utilisation réels.

En tant qu'organisme d'évaluation, le LNE (Laboratoire national de métrologie et d'essais) s'intéresse à cet axe de recherche, motivé notamment par la grande expansion que connaît le développement de la robotique et des systèmes intelligents complexes. C'est ainsi que les travaux de cette thèse s'articulent autour de la problématique de l'évaluation en contexte applicatif, et plus particulièrement, autour de l'évaluation des systèmes de reconnaissance automatique de la parole dans le cadre d'une application de reconnaissance d'entités nommées à partir de la parole.

Organisation du document

Ce document de thèse comporte trois parties. La première partie dresse un panorama historique des tâches de reconnaissance d'entités nommées (REN) et de la reconnaissance automatique de la parole (RAP), ainsi que de leurs méthodes d'évaluation à travers les principales campagnes d'évaluation. Tout d'abord, nous commençons par décrire les campagnes d'évaluation de la tâche de REN et les mesures utilisées pour son évaluation. Ensuite, nous exposons l'évolution des défis liés à la RAP dans les campagnes d'évaluation pour aborder la problématique de la RAP. Enfin, nous concluons cette partie par un état de l'art sur les mesures proposées pour l'évaluation des systèmes de RAP en contexte applicatif.

En deuxième partie, nous commençons par décrire la tâche de reconnaissance d'entités nommées structurées et compositionnelles étudiées et ses enjeux. Nous montrons ensuite que les métriques d'évaluation existantes ne permettent pas d'évaluer efficacement cette tâche, et nous introduisons une nouvelle métrique (ETER) qui permet de gérer la complexité générée par la structure hiérarchique et compositionnelle de la tâche de REN étudiée. Nous concluons cette partie par une expérience permettant de valider la métrique proposée.

Dans la dernière partie, nous abordons la problématique de l'évaluation des systèmes de RAP pour la REN. Nous commençons par montrer que les mesures d'évaluation classiquement utilisées pour évaluer les systèmes de RAP en isolation, ne sont pas adaptées pour l'évaluation en contexte de REN. Nous montrons également que les erreurs de RAP peuvent avoir des impacts différents sur le fonctionnement des systèmes de REN appliqués en aval, en faisant baisser soit le rappel soit la précision de ces derniers. En s'appuyant sur les résultats de nos analyses et de notre étude de la littérature, nous avons défini une nouvelle mesure qui permet d'évaluer la qualité des systèmes de RAP et l'impact de leurs erreurs pour les systèmes de REN appliqués en aval. Notre mesure vise à mesurer la perte d'information causée par les erreurs de RAP en ciblant les passages de textes importants

pour les systèmes de REN. Enfin, nous validons notre mesure en la comparant aux mesures de l'état de l'art sur la base de données réelles issues de campagnes d'évaluation.

Première partie

Cadre et états de l'art

Chapitre 1

Historique des campagnes d'évaluation en reconnaissance d'entités nommées

1.1 Introduction

Dans le cadre de cette thèse, nous nous intéressons au cadre applicatif de reconnaissance d'entités nommées à partir de la parole. L'une des deux briques technologiques nécessaire pour l'aboutissement de cette application est la reconnaissance d'entités nommées (REN). Cette brique technologique relève du traitement automatique des langues (TAL) et peut être inscrite dans la branche des traitements relevant de la dimension sémantique.
Ce chapitre introductif a pour but de présenter la tâche de la REN et ses perspectives d'utilisation. Dans un premier temps, nous donnerons une définition générale de cette tâche. Ensuite, nous dresserons un historique des principales campagnes d'évaluation l'ayant traitée, pour nous arrêter sur les changements ayant touché à la définition de cette tâche, ainsi que les types de données traitées et les méthodes mises en place pour son évaluation. Enfin, nous clôturerons ce chapitre introductif en présentant quelques cadres applicatifs exploitant la REN à travers une petite discussion sur les enjeux relatifs au traitement des données orales.

1.2 Définitions

Il n'existe pas de définition unique pour désigner les entités nommées. Ainsi, nous pouvons trouver dans la littérature des définitions variées suivant la période et le point de vue adoptés. Pour plus de détails concernant les différentes approches de définition des entités nommées nous renvoyons le lecteur à la thèse de Maud Ehrmann [Ehrmann, 2008] portant sur ce sujet. Nous abordons ici la

question d'un point de vue plus général. L'évolution de la définition des entities nommées et son impact seront discutés à travers l'étude de l'historique des campagnes d'évaluation, dans la section suivante.

La REN s'inscrit parmi les tâches du TAL qui visent le traitement sémantique du texte. Elle a pour but d'identifier les éléments porteurs de sens et de les classer en fonction de catégories sémantiques bien définies. Parmi les catégories les plus utilisées dans le traitement des textes journalistiques et médiatiques, nous pouvons citer, les Personnes, les Lieux et les Organisations, auxquelles s'ajoutent les Quantités et les Expressions temporelles. Aujourd'hui, la reconnaissance d'entités nommées est également utilisée pour le traitement des données dans des domaines spécifiques tels que la biologie [Fukuda *et al.*, 1998] et la médecine [Bodenreider et Zweigenbaum, 2000], ainsi, dans ces cas les catégories sémantiques visées concernent plutôt les Gènes, les Protéines, les Maladies, etc.
Même s'il semble y avoir un accord sur le fait que les entités nommées sont des éléments riches sémantiquement qui désignent des référents uniques. Des cas de désaccords peuvent être observés concernant leur nature, leurs catégories et les règles d'annotations à adopter pour les traiter.
Pour mieux comprendre les enjeux relatifs à la tâche de REN, nous commençons par parcourir son historique à travers les principales campagnes d'évaluation américaines et internationales.

1.3 Historique des campagnes d'évaluation en reconnaissance d'entités nommées

Les campagnes d'évaluation, (appelées aussi *evaluation conferences* en anglais) sont des événements qui permettent d'encourager la recherche dans un domaine particulier. Elles offrent la possibilité de réunir différents acteurs autour d'une problématique, afin qu'ils puissent proposer des solutions permettant de la résoudre. Elles permettent aussi de fournir des données, des outils d'évaluation et de mesurer la maturité de la technologie.

Des nombreuses campagnes d'évaluation ont été financées dans le but d'encourager la recherche en extraction d'informations et en compréhension de textes. La tendance a commencé aux États-Unis avec les campagnes MUC (*Messages Understinding Conference*), puis des campagnes d'évaluation similaires ont été organisées un peu partout dans le monde.
Nous parcourons ici brièvement l'historique des principales campagnes d'évaluation qui se sont intéressées à la problématique de la REN, afin de mieux comprendre les raisons de son succès, puis nous nous arrêterons sur les modifications

ayant touché à la définition de la tâche et aux méthodes utilisées pour son évaluation.

1.3.1 Les campagnes d'évaluation MUC

L'histoire des campagnes d'évaluation en extraction d'informations a commencé avec les campagnes MUC (*Message Understanding Conference*) en 1987. Ces campagnes se sont déroulées entre 1987 et 1997. Elles ont été financées par la DARPA (*Defense Advanced Research Project Agency*) dans le but d'encourager et de promouvoir l'analyse automatique des messages militaires contenant des informations textuelles. La série de conférences MUC a permis de définir un programme de recherche et de développement qui a aidé les chercheurs et les développeurs à mieux comprendre les enjeux de la tâche. Cela a aussi été l'occasion de mettre en place et de développer des procédures et des métriques d'évaluation nécessaires pour évaluer cette technologie. C'est aussi durant les cycles de cette conférence que la tâche de REN a vu le jour. Nous pouvons distinguer trois phases différentes dans le déroulement de la conférence MUC selon la nature de la tâche visée.

1.3.1.1 MUC-1 et MUC-2 : phase exploratoire

Cette première phase se caractérise par l'utilisation exclusive de données provenant des messages de la marine américaine (*US Navy*), et par l'absence de tâches et de procédures d'évaluation bien définies.

MUC-1 (1987) : la première édition a été une phase exploratoire dont le but était de faire un état de l'art des systèmes de compréhension de textes et de population de base de connaissances. Lors de cette première édition les participants choisissaient le format des sorties de leurs systèmes, aucune évaluation formelle n'ayant été mise en place. Cette conférence a le mérite d'avoir réuni des chercheurs et des développeurs pour discuter de la nature des informations utiles à extraire à partir des messages de la marine américaine et de développer les premières approches pour aborder une telle problématique.

MUC-2 (1989) : durant cette deuxième édition, les participants ont exigé d'avoir une définition de la tâche d'extraction précise avec un format bien défini afin de pouvoir, par la suite, évaluer la qualité des systèmes abordant une même tâche. Cette tâche consistait à repérer des événements dans des textes et à remplir, pour chaque événement, un formulaire (*template*). Chaque formulaire contenait des champs relatifs à un incident (l'opération) qui étaitent les acteurs, la date, l'heure, le lieu et les résultats (d'autres informations à extraire existaient mais n'étaient pas bien définies). La précision (P) et le rappel (R) étaient les métriques adoptées pour mesurer les performances des systèmes, dans MUC-2 P et R étaient définis comme suit :

$$R = \frac{\text{Nombre de réponses corrects}}{\text{Nombre total de réponses recherchées}}$$

$$P = \frac{\text{Nombre de réponses corrects}}{\text{Nombre de réponses corrects} + \text{Nombre de réponses incorrects}}$$

L'inconvénient majeur de cette conférence était l'absence d'une procédure d'évaluation automatisée, les participants étaient alors amenés à évaluer eux-mêmes leurs sorties manuellement. Ceci a conduit à une grande variation dans la manière dont chaque participant évaluait les sorties de son système. Ceci a mis en évidence la nécessité de mettre en place un paradigme d'évaluation unifié et automatisé.

1.3.1.2 MUC-3, MUC-4 et MUC-5 : remplir des formulaires

Cette deuxième phase des campagnes MUC s'est distinguée de la précédente par l'utilisation de données plus diversifiés et par l'adoption d'une tâche axée sur la compréhension du texte pour remplir des formulaires (*Fill in Template*). Des procédures d'évaluation automatisées ont été mises en place et de nouvelles métriques d'évaluation fondées sur le taux d'erreur ont été introduites.

MUC-3 (1991) : la troisième édition de la série MUC s'est distinguée des deux versions précédentes essentiellement par l'utilisation d'une plus grande variété de données. Cette fois les données ne se limitaient plus aux messages de la marine américaine, elles incluaient une diversité d'articles journalistiques portant sur les activités terroristes en Amérique latine. Ceci avait permis d'avoir plus de données pour les tests et pour l'entraînement des systèmes. La tâche consistait à détecter les événements se référant à des actes terroristes en Amérique latine et il y avait dix-huit champs à remplir par formulaire. Ainsi, cette tâche est devenue un peu plus complexe, avec plus d'informations à extraire (dix-huit champs par formulaire) et des données plus riches et plus diversifiées à analyser. Il est important de remarquer aussi que les informations à extraire nécessitaient plus d'analyse et de compréhension du texte ce qui a augmenté la complexité de la tâche [Sundheim, 1991]. Voilà une liste non exhaustive des informations à extraire :

- déterminer si c'est un acte criminel ou terroriste ;
- déterminer si les informations (dans le texte) sont précises ou vagues ;
- déterminer s'il s'agit d'une menace ou d'un véritable acte ;
- déterminer la cause de l'acte ;
- déterminer l'origine des acteurs ;
- déterminer le lieu, la cible, la date, les instruments utilisés, etc.

MUC-3 s'est aussi distingué par la mise en place d'une vraie procédure d'évaluation automatisée qui consistait à comparer les formulaires remplis par les systèmes aux formulaires de référence remplis manuellement, puis, ensuite, à comptabiliser

les bonnes et les mauvaises réponses pour calculer le rappel et la précision pour chaque système. Il s'est avéré, durant cette campagne, que l'utilisation de deux métriques d'évaluation (P et R) créait une certaine confusion quant à la classification des systèmes, du fait que, dans certains cas, il était difficile de décider quel système avait de meilleures performances par rapport aux autres.

MUC-4 (1992) : la même tâche que celle de MUC-3 a été proposée dans MUC-4, avec, toutefois, quelques améliorations apportées aux définitions de certains *slots* (champs à remplir dans le formulaire), qui, étant très générales, étaient à l'origine de nombreux cas de confusion. Ainsi, des définitions plus précises des *slots* ont été mises en place et le nombre des slots à extraire est passé de dix-huit (dans MUC-3) à vingt-quatre (dans MUC-4) [Sundheim, 1992]. MUC-4 s'est distingué des conférences précédentes par son utilisation d'une nouvelle métrique d'évaluation la « F-mesure » qui était définie comme la moyenne harmonique entre P et R [Chinchor et Sundheim, 1993]. La F-mesure a donné la possibilité de classer les systèmes de REN selon une mesure unique.

MUC-5 (1993) : durant cette cinquième édition, il y avait deux types d'événements à traiter : les coentreprises internationales et la fabrication de circuits électroniques, et ce, dans deux langues, l'anglais et le japonais. Contrairement aux autres conférences où il n'y avait qu'un seul type de formulaire, dans MUC-5, il y en avait onze avec un total de quarante-sept types de champs distincts [Chinchor et Sundheim, 1993]. Ainsi, la tâche avait encore gagné en complexité par rapport aux éditions précédentes.

Cette conférence a permis d'enregistrer l'utilisation d'une nouvelle métrique qui permettait d'évaluer les performances des systèmes en termes de taux d'erreur. Cette métrique était appelée « *the error per response* » (ERR) et elle était la mesure officielle durant MUC-5 [Chinchor et Sundheim, 1993].

$$ERR = \frac{\text{Nombre de fausses réponses}}{\text{Nombre total des réponses + Nombre d'omissions}}$$

1.3.1.3 MUC-6 et MUC-7 : entités nommées

Durant les cinq premières campagnes (MUC-1 à MUC-5) qui se sont déroulées entre 1987 et 1993, la tâche mise en place a consisté à extraire des informations se trouvant dans des documents textuels afin de remplir des formulaires. Même si les performances obtenues par les systèmes étaient encourageantes, 57 % pour le rappel et 64 % pour la précision sur l'ensemble des données dans MUC-5, la tâche de remplissage des formulaires n'a pas cessé de gagner en complexité d'une campagne à l'autre. Elle nécessitait des niveaux d'analyse et de compréhension de textes de plus en plus avancés. En effet, pour remplir un formulaire, le système doit effectuer des traitements à plusieurs niveaux : détecter les entités, catégoriser les entités détectées, extraire les événements, repérer les relations pouvant exister entre les entités ou entre les entités et les événements et déterminer la nature des

relations. La complexité de cette tâche rend les diagnostics et la compréhension des provenances des erreurs très difficiles, puisque tous les modules sont étroitement liés les uns aux autres.

MUC-6 (1995) : les objectifs dans MUC-6 (1995) étaient d'améliorer les performances des systèmes d'extraction d'informations. L'hypothèse était que ceci n'était possible que si l'on améliorait l'analyse sémantique effectuée par les systèmes. Ainsi, une nouvelle tâche de coréférence fut introduite pour encourager l'amélioration du traitement sémantique. Elle a consisté à marquer des relations comme l'anaphore, la métonymie, et d'autres relations.

Mais, comme il s'agissait de l'avant-dernière édition de la série, il y avait dans MUC-6 une volonté de prouver que la technologie d'extraction d'informations existante était exploitable rapidement avec des performances élevées et qu'elle pouvait être indépendante du domaine. Pour atteindre ce but, l'idée était d'identifier, parmi la technologie développée, la brique (le composant) qui satisfaisait au mieux ces critères. C'est dans cette logique que la tâche (*Named Entity*) de reconnaissance d'entités nommées a été introduite pour la première fois durant MUC-6.

La tâche de la REN a consisté à utiliser des marqueurs SGML pour identifier des noms propres dans les textes (noms de personnes, noms d'organisations ou noms de lieux), des expressions temporelles et des expressions numériques (monétaires ou pourcentages) [Sundheim, 1996]. Trois classes d'entités ont été définies comme suit :

- **ENAMEX** : (Personnes, Organisations, Lieux) limité aux noms propres et acronymes ;
- **NUMEX** : limité aux expressions monétaires et de pourcentages. Toute expression qui n'utilise pas des termes monétaires ou des pourcentages n'est pas annotée ;
- **TIMEX** : il inclut seulement les expressions temporelles absolues donnant une information précise sur l'heure ou la date. Les expressions relatives ne sont pas annotées.

Nous présentons ici quelques exemples d'entités nommées appartenant aux différentes classes adoptées. Les exemples sont tirés du guide d'annotation des campagnes d'évaluation. Nous avons choisi de garder les exemples dans la langue de leur campagne d'évaluation afin de conserver les ambiguïtés lexicales ou/et sémantiques qu'ils contiennent et qui pourraient être modifiées par une éventuelle traduction. Voici quelques exemples d'annotations de la classe ENAMEX :

- <ENAMEX TYPE="ORGANIZATION"> DARPA </ENAMEX>
- Mr. <ENAMEX TYPE="PERSON"> Harry Schearer </ENAMEX>
- <ENAMEX TYPE="LOCATION"> Los Angeles </ENAMEX>

1.1.3 Historique des campagnes d'évaluation en reconnaissance d'entités nommées

Pour la classe ENAMEX, les expressions complexes et les groupes nominaux se référant à des entités sans les nommer ne sont pas annotés, ainsi :

- vice president
- prime minister
- <ENAMEX TYPE="ORGANIZATION"> Machinists </ENAMEX> union
- the <ENAMEX TYPE="PERSON"> Kennedy </ENAMEX> family
- <ENAMEX TYPE="LOCATION"> U.S </ENAMEX> exporters

Les règles d'annotations mises en place à l'occasion de MUC-6 excluent l'annotation de certains noms propres comme les noms de groupes politiques, les noms de lois, de prix, de maladies, de saints, ou encore les noms de rues portant des noms de personnes : *The Republicans, Alzheimer's, the Nobel prize, St. Michael, Gatchell Road*.

Exemples d'expressions appartenant à la classe NUMEX :

- <NUMEX TYPE="MONEY">20 millions New Pesos</NUMEX> ;
- <NUMEX TYPE="PERCENT">15 pct</NUMEX> ;
- about <NUMEX TYPE="PERCENT">5%</NUMEX> ;
- over <NUMEX TYPE="MONEY">$90,000</NUMEX> ;

Les exemples suivants ne sont pas annotés : 12 points, 1.5 times, priced at 99, 1/4.
Exemples d'expressions appartenant à la classe TIMEX :

- <TIMEX TYPE="TIME">twelve o'clock noon</TIMEX> ;
- <TIMEX TYPE="DATE">January 1990</TIMEX> ;
- <TIMEX TYPE="DATE">July</TIMEX> last year ;

L'expréssion *last night* n'est pas annotée.

La tâche de REN a été un vrai succès. Les résultats des évaluations ont été très satisfaisants. Presque la moitié des participants ont obtenu des valeurs de rappel et de précision supérieures à 90 %. Le meilleur système avait 96 % de rappel, 97 % de précision. La métrique officielle adoptée pour classer les systèmes était ERR et le meilleur système affichait 5 % d'erreurs [Sundheim, 1996].

MUC 7 (1997) : après son succès dans la version précédente, la tâche de REN a été également adoptée dans cette septième et dernière édition de MUC. La définition de la tâche était la même que celle introduite dans MUC-6, avec cependant de légères différences. En effet, dans MUC-7 il était question d'annoter certaines expressions temporelles relatives [Chinchor, 1998], des règles d'annotations distinctes de celles de MUC-6 ayant été mises en place pour traiter les cas des conjonctions de coordination.

Cette édition s'est distinguée des précédentes par le fait que les données d'apprentissage et de test ne portaient pas sur le même domaine (apprentissage : crash des avions, test : lancement de fusée). Ceci avait pour but d'encourager le développement de systèmes flexibles et génériques.

1.3.1.4 Résumé des campagnes MUC

MUC avait pour but d'encourager la compréhension de textes et l'extraction d'informations. Après une phase exploratoire, la tâche de remplissage de formulaires a vu le jour à l'occasion de la troisième édition. Cette tâche s'est avérée trop complexe, et il a été décidé de se concentrer sur le développement de briques plus élémentaires, d'où l'introduction de la tâche de reconnaissance d'entités nommées à l'occasion de MUC-6.
Un résumé des résultats des campagnes MUC est affiché dans le tableau 1.1. Nous pouvons voir que les performances sur la tâche de remplissage des formulaires étaient encourageantes malgré la complexité qui n'a pas cessé de croître d'une campagne à une autre. Toutefois, les performances n'ont pas dépassé 60% de F-mesure durant tout MUC.

	REN	Fill in Template	Multilangue
MUC-3		R $\simeq$ 50 % P $\simeq$ 70 %	
MUC-4		F $\simeq$ 56 %	
MUC-5		anglais F $\simeq$50 %	japonais F $\simeq$ 57 %
MUC-6	F$\simeq$97 %	F$\simeq$57 %	
MUC-7	F$\simeq$94 %	F$\simeq$51 %	

TABLEAU 1.1 – Tableau récapitulatif des résultats (seulement les meilleurs résultats sont affichés) des conférences MUC, tirés de [Chinchor, 1998].

Les résultats pour la tâche de REN étaient très satisfaisants (> 90%). Ceci a fait le succès de cette tâche et a encouragé son développement pour viser l'extraction de plus d'informations en élargissant le concept d'entité nommée. C'est ainsi que la définition de cette tâche a évolué dans la nouvelle série des campagnes d'évaluation ACE (*Automatic Content Extraction*) qui a pris la relève des campagnes MUC en 1999.

1.3.2 Les campagnes d'évaluation ACE

La série ACE, oraganisée par le NIST, avait pour but de promouvoir l'analyse sémantique, en développant la tâche de la REN introduite dans MUC. En effet, les

bons résultats affichés par les systèmes de REN permettent d'envisager de pousser le traitement pour couvrir plus d'éléments riches sémantiquement. ACE visait également à encourager le développement de systèmes plus génériques pouvant couvrir des domaines plus larges et plus robustes comme le traitement des données bruitées provenant de systèmes de reconnaissance automatique de la parole (RAP) ou ayant été numérisées par OCR (*Optical Character Recognition*).

1.3.2.1 Déroulement

ACE (1999-2000) : la première version de ACE (1999-2000) s'est focalisée sur la détection et la classification des entités « *Entity Detection and Tracking* ». Quatre tâches ont alors été définies :

- la détection d'entités nommées ;
- la classification des entités ;
- la détection des mentions qui consiste à repérer dans le texte les expressions nominales ou pronominales se référant à une même entité (toutes les occurrences d'une entité) ;
- la reconnaissance des extensions des mentions qui consiste à extraire les syntagmes décrivant des mentions d'entités.

Comme nous pouvons le remarquer dans cette première édition, la détection et la classification des entités étaient séparées en deux tâches et ont été évaluées séparément. Les deux tâches sur les détections des mentions d'entités visaient à pousser un peu plus l'analyse sémantique et l'exploitation des sorties des systèmes de REN. Il y avait aussi une volonté de redéfinir les classes d'entités pour offrir une classification plus précise [NIST, 2000]. Ainsi cinq classes d'entités ont été définies comme suit :

- la classe des Personnes : toute personne ou tout groupe de personnes mentionnés par un nom, un groupe nominal ou un pronom. Une personne peut être désignée par son nom (*John Smith*), sa fonction (*the butcher*), sa relation familiale (*dad*), un pronom (*he*)... Les groupes de personnes ne faisant pas partie d'une organisation structurée font aussi partie de la classe Personnes, par exemple : *the family, the house painters, the linguist...*
- la classe des Organisations : toute organisation ou ensemble d'organisations ayant une structuration bien définie. Ceci inclut les entreprises, les unités gouvernementales, les équipes sportives, les groupes musicaux... ;
- la classe GSP (*Geographical- Social-Political entities*) : toutes les régions géographiques définies sur des bases politiques ou sociales comme *France* ou *Boston*, y compris des occurrences faisant référence à différents concepts. cela peut référer à une région géographique comme dans *France has a temperate climate*, un gouvernement comme dans *France signed a treaty last year*, les citoyens de la région comme dans *France elected a new president* ;

- la classe des Lieux : elle désigne un lieu défini sur des bases géographiques ou astronomiques et ne constituant pas une entité politique. Ceci inclut par exemple : *solar system, Mars, the continents, the Mideast , Mt-Everest*.... Un Lieu peut être une partie d'un GSP comme dans *the center of the city* ou *the south of France* ;
- la classe des Bâtiments : elle inclut les infrastructures et les bâtiments construits par l'homme comme les habitations, les usines, les stades, les prisons, les musées, les parkings, les routes, les ponts...

Dans ACE, les entités nommées ne se limitaient plus aux seuls noms propres, mais couvraient aussi les pronoms et les différents types de syntagmes nominaux.

Afin d'encourager le traitement de différents types de données (écrites, orales et données bruitées), les corpus d'entraînement et de test étaient composés d'un tiers d'articles de journaux venant du Web (*newswire*), d'un tiers d'émissions de journaux télévisés (*broadcast news*) transcrites manuellement et d'un tiers de journaux papier (*newspapers*) numérisés par OCR. Les participants étaient également amenés à traiter des données de type journaux télévisés transcrites automatiquement par des systèmes de RAP.

La métrique d'évaluation, qui a été utilisée pour la tâche de REN, est le *Slot Error Rate* (SER) [Makhoul *et al.*, 1999] qui est une adaptation de ERR (utilisée dans MUC) et qui se calcule comme suit :

$$SER = \frac{I + D + S}{\text{Nombre total des slots dans la référence}} \tag{1.1}$$

Avec

- I : le nombre des insertions, ou fausses alarmes, lorsqu'une entité se trouve dans l'hypothèse mais n'existe pas dans la référence ;
- D : le nombre des omissions, lorsqu'une entité se trouve dans la référence mais n'existe pas dans l'hypothèse ;
- S : le nombre des substitutions, lorsqu'une entité est correctement détectée mais mal classifiée ou que les frontières de début et/ou de fin sont mal placées.

ACE (2001-2002) Cette édition de la série ACE s'est distinguée par l'introduction d'une nouvelle tâche qui consiste à détecter les relations pouvant relier des entités. Par exemple, une personne A travaille dans l'organisation B, ou bien une personne C se trouve dans le lieu D. La détection de ce type de relation est très utile pour de nombreuses applications comme la compréhension de textes et l'indexation de documents, et illustre la volonté de pousser encore un peu plus loin le traitement sémantique en exploitant les sorties des systèmes de REN.
Plus de données que dans la version précédente ont été collectées à l'occasion de cette campagne d'évaluation. Toutefois, les mêmes types de données que ACE

(1999-2000) (*Newswire, broadcast news et newspapers*) et avec la même distribution (un tiers de chaque type) ont été rassemblés pour l'entraînement et le test.

Étant donné que la technologie de REN intéresse des personnes ayant des visées applicatives différentes, un nouveau modèle d'évaluation dit « modélisation du coût » (*cost model*) a été proposé pour prendre en compte l'application visée lors de l'évaluation. L'idée est de créer un modèle statistique qui permette de quantifier le coût des erreurs d'omission et d'insertion étant donnée l'application visée[NIST, 2002]. La métrique résultante est appelée *Cost Detection* (c_{det}), est définie comme suit :

$c_{det} = (\text{coût estimé pour une omission /unité}) + (\text{coût estimé pour une insertion / unité})$

L'utilisation d'un modèle de prédiction de coût permet de changer facilement le modèle pour s'adapter à des cadres applicatifs différents. En revanche, on rencontre un inconvénient qui est que les mesures offertes peuvent ne pas être intuitives, et que la validité des modèles de prédiction n'a pas été prouvée [NIST, 2002].

ACE (2003) Cette édition a gardé les mêmes consignes d'annotations et les mêmes métriques d'évaluation que l'édition précédente. Elle s'est toutefois distinguée par l'exportation de la tâche à deux autres langues, le chinois et l'arabe. Ceci a permis de rassembler des données pour ces deux langues pour encourager le développement d'outils permettant de les traiter. C'était aussi une occasion pour évaluer la portabilité de la technologie de REN vers des langues étrangères.

ACE (2004) Nous pouvons distinguer deux grandes innovations, par rapport aux éditions précédentes :

- l'introduction de deux nouvelles tâches :
  - la détection des événements, sachant que cinq types d'événements ont été définis : destruction, création, transfert des biens, déplacement, interaction entre agents,
  - la détection des expressions temporelles ;
- la définition d'une nouvelle typologie, dans le but d'affiner le tri des entités. Deux nouveaux types ont été introduits (*Vehicle* et *Weapon*) et une structure hiérarchique avec des sous-types à l'intérieur de chaque type a été mise en place.

Le retour à la tâche de détection d'événements, qui était introduite au début durant les conférences MUC, reflète une volonté d'évaluer la possibilité de résoudre cette tâche qui était à l'époque de MUC un peu trop complexe. Toutefois, il est à noter que la tâche de détection d'événements qui a été mise en place dans le cadre de cette édition de ACE est plus avancée que celle de MUC et couvre beaucoup plus de catégories d'événements.
L'élargissement de la typologie d'entités et l'introduction des sous-types reflètent, quant à eux, la volonté de couvrir plus de domaines et de montrer que la REN peut être généralisée vers certains domaines spécifiques.

Comme les deux versions précédentes de ACE, les données d'entraînement et de test étaient du type journaux papier, journaux du Web et émissions radio- et télédiffusées.

Cependant, la procédure d'évaluation a encore changé durant cette édition car les deux tâches de détection d'entités nommées et de détection des mentions ne sont plus évaluées indépendamment l'une de l'autre. Une procédure d'évaluation et une nouvelle métrique, qui permettent d'évaluer la combinaison des deux, ont été mises en place. Ainsi, une nouvelle métrique d'évaluation a été définie pour évaluer la tâche globale complexe en résultant. Cette métrique s'appelait EDT-value (*Entity Detection and Tracking Value*) [Doddington *et al.*, 2004]. Elle consistait à calculer un score pour chaque entité. Ainsi, la somme des scores de toutes les entités de l'hypothèse était le score affecté au système. Le score calculé au niveau des entités comportait une composante pour la détection et la classification d'entités nommées et une autre composante pour la détection des occurrences.

$$EDT\_value_{sys} = \sum_i value\_of\_sys\_entity_i$$

Ainsi, puisque chaque entité pouvait avoir plusieurs mentions, le score relatif à l'annotation en entités nommées ($Entity\_Value$) était multiplié par la somme des scores de toutes ses mentions $(Mention\_value)$.

$$Value\_of\_sys\_entity = Entity\_Value(entity).\sum Mention\_value(entity)$$

La métrique EDT_value permettait d'affecter des poids différents aux entités selon leurs types et leurs classes. Ainsi, pour les paires d'entités (hypothèse, référence) issues de la phase d'alignement, le score relatif à l'annotation en entités nommées ($Entity\_Value$) était calculée en multipliant le poids relatif au type et à la classe de l'entité (si le type ou la classe de référence et l'hypothèse étaient différents, le poids le plus petit était sélectionné) par le poids des erreurs. Plus les erreurs étaient graves plus le poids s'approchait de zéro réduisant ainsi le score obtenu pour l'annotation de l'entité.

$$\begin{aligned} Entity\_Value = \min \begin{pmatrix} ETypeValue(hyp).EClasseValue(hyp) \\ ETypeValue(ref).EClasseValue(ref) \end{pmatrix} \\ \times (W_{Err-type}.W_{Err-sousType}.W_{Err-classe}) \end{aligned} \tag{1.2}$$

Dans le cas des omissions, un score correspondant à la valeur affectée au type et à la classe de l'entité était retiré, et, dans le cas des fausses alarmes, un score de pénalité était affecté à $Entity\_Value$ (la pénalité dépendait également du type et de la classe de l'entité insérée). $Mention\_value$ était le score correspondant à l'annotation des mentions. Il était calculé de façon similaire à $Entity\_Value$ en comparant cette fois les annotations relatives au type, rôle et style de la mention.

1.1.3 Historique des campagnes d'évaluation en reconnaissance d'entités nommées

ACE (2005) Il s'agit d'une campagne d'évaluation multilingue (anglais, chinois, arabe) durant laquelle il y a eu l'introduction d'une nouvelle notion, celle de « classe ». Une révision de la typologie des catégories d'entités a été également effectuée. Ainsi, de nouvelles sous-catégories ont été introduites et d'autres (datant de ACE-2004) ont été supprimées. Chaque entité possède une catégorie, une sous-catégorie et une classe. Le tableau 1.2 présente les quatre classes d'entités définies dans ACE 2005 et leurs définitions.

Classes	Définitions
SPC	Les entités à référent unique, particulier ou spécifique
GEN	Entités génériques plutôt qu'une entité spécifique
NEG	Entités génériques, négativement quantifiées
USP	Non spécifié, classe incertaine

TABLEAU 1.2 – Les classes d'entités introduites durant ACE 2005 [NIST, 2005].

Les mêmes procédures et métriques d'évaluation que ACE 2004 ont été utilisées. Toutefois, les données à traiter comportaient en plus des trois types habituels (journaux papier, journaux du Web et émissions radio- et télédiffusées) des données de type paroles conversationnelles dans des émissions télévisées, des conversations téléphoniques et des données collectées sur des blogs et des forums de discussion. L'introduction de ces nouveaux types de données témoigne de l'intérêt que suscitait l'application de la technologie de la REN sur la parole.

ACE 2007 Durant cette édition, il a été décidé de garder la même tâche pour la REN et les mêmes méthodes d'évaluation [NIST, 2007] que ACE-2005. Cette édition s'est distinguée, essentiellement, à travers l'exportation de la tâche de la REN à une nouvelle langue, l'espagnol, en plus de l'arabe et du chinois, et par l'introduction d'une nouvelle tâche qui est celle de la traduction des entités nommées. L'introduction de cette nouvelle tâche était motivée par les résultats des expériences menées dans [Babych et Hartley, 2003] qui montraient que l'intégration d'un système d'extraction d'entités nommées dans un système de traduction pouvait améliorer les performances du système de traduction de 20 %. En effet, les entités sont des mots porteurs de sens, et pouvoir traduire ces expressions est essentiel pour améliorer les performances des systèmes multilingues. La tâche de traduction d'entités consistait à extraire, à partir de documents en langues arabe et chinoise, les entités nommées pour donner en sortie un catalogue contenant leur traduction et leur description en anglais.

Les données étaient du même type que celles ACE-2005, mais elles comportaient en plus la traduction du corpus test vers les différentes langues (espagnol, arabe et chinois).

ACE 2008 Le défi, dans cette dernière édition des conférences ACE, consistait à reconnaître toutes les mentions d'une même entité dans le même document et dans des documents différents.
Deux processus d'évaluation intra et inter-documents ont été mis en place pour les tâches de détection d'entités nommées et de détection des relations. La métrique d'évaluation qui a été utilisée durant ACE 2008 est LEDR-value (*Local Entity Detection and Recognition value*) [NIST, 2008] qui consiste à normaliser le score EDT_{value} (utilisé dans ACE-2004) [Doddington *et al.*, 2004] calculé à partir des entités de l'hypothèse par la somme des scores calculés à partir des entités de la référence.

$$LEDR\_value_{sys} = \frac{\sum_i value\_of\_sys\_token_i}{\sum_i value\_of\_ref\_token_i}$$

1.3.2.2 Résumé des campagnes ACE

La série des campagnes d'évaluation ACE, qui a commencé en 1999, avait pour but de promouvoir la tâche de la REN introduite dans MUC et de la développer afin de pousser plus en avant le traitement sémantique de la langue écrite et parlée. Ainsi, les campagnes d'évaluation incluaient diverses tâches permettant d'exploiter les annotations en entités nommées et d'améliorer le traitement sémantique tel que la détection d'occurrences, la reconnaissance des relations, la détection des événements et la traduction des entités, afin d'encourager le développement de systèmes robustes par rapport aux ambiguïtés de la langue. Les corpus (d'entraînement et de test) créés provenaient de sources différentes, journaux papier, journaux du Web et émissions radio- et télédiffusées, conversations radio- et télédiffusées, conversations téléphoniques, *web blog*, etc. Ces corpus incluaient des données transcrites automatiquement par des systèmes de RAP, ainsi que des données numérisées avec OCR. Dans le but d'exporter la technologie à d'autres langues, certaines campagnes de la série étaient multilingues, anglais, arabe, chinois et espagnol (pour ACE-2005). La série ACE a ainsi contribué énormément au développement de la REN et du traitement sémantique de la langue. Elle a aussi servi de modèle à de nombreuses autres campagnes d'évaluation à travers le monde.

1.3.3 Autres campagnes d'évaluation en extraction d'entités nommées

En parallèle aux deux séries MUC et ACE, qui étaient les conférences fondatrices pour de nombreux aspects de la tâche de la REN et de ses méthodes d'évaluation, il y a eu de nombreuses autres conférences, qui ont suivi la ligne de MUC et ACE avec plus ou moins d'innovation qui touche principalement les langues

traitées et la typologie des entités. Nous citons ici brièvement une sélection non exhaustive de ces conférences.

En parallèle à MUC-6 et MUC-7, il y a eu les campagnes MET (*Multilingual Entity task)* [Merchant *et al.*, 1996] dans lesquelles les définitions d'entités adoptées étaient similaires à celles de MUC. Les langues traitée durant MET étaient l'espagnol, le chinois et le japonais.
La campagne HUB-4 [Przybocki *et al.*, 1999] était la première conférence qui s'est intéressée à l'extraction d'entités à partir de données de type *broadcast news* transcrites automatiquement par des systèmes de RAP. Cette conférence s'est déroulée en 1998 avant la première conférence ACE et a adopté une définition des entités similaire à celle de MUC. Juste après, il y a eu le projet d'évaluation IREX au Japon qui s'est intéressé à l'évaluation de la tâche de la REN en japonais. La définition des entités nommées dans la campagne IREX [Sekine et Isahara, 2000] était proche de celle de MUC et incluait une nouvelle catégorie d'entités « miscellaneous ». Cette nouvelle catégorie regroupe toutes les entités qui ne sont pas des Personnes, des Organisations ou des Lieux.
CONLL (*Conference On Natural Language Learning*) s'est aussi intéressé à la tâche de REN dans le but d'encourager le développement de méthodes fondées sur des approches statistiques. Ainsi, deux campagnes d'évaluation ont été organisées pour cette tâche en 2002 et 2003. La première version [Sang, 2002] traitait l'espagnol et le néerlandais, alors que la deuxième [Tjong Kim *et al.*, 2003] traitait l'anglais et l'allemand. En 2006, la campagne d'évaluation HAREM [Santos *et al.*, 2006] aborde le portugais et adopte une définition des entités proche de celle des campagnes ACE, avec une typologie de catégories et de sous-catégories beaucoup plus large. En Italie, la série de conférences EVALITA traite la tâche de REN depuis 2007 [Speranza, 2007] pour l'italien. Quatre campagnes d'évaluation EVALITA ont été organisées en 2007, 2009, 2011 et 2014. Elles s'occupaient en plus de REN des tâches telles que la reconnaissance d'occurrences d'entités, la détection des relations et la détection des événements.

C'est ainsi que, grâce aux différentes campagnes d'évaluation, la tâche de la REN s'est généralisée et s'est exportée vers différentes langues et différents types de données. Même si le principe de base reste le même, quelques variations peuvent être notées dans la manière de définir les entités nommées, leurs catégories ou les métriques d'évaluation utilisées. Par conséquent, il est difficile de comparer les performances réalisées dans chacune des conférences.

1.3.4 Les campagnes d'évaluation de la REN en langue française

Il faut attendre 2005 pour voir la première campagne d'évaluation en reconnaissance d'entités nommées qui traite la langue française. Il s'agit de la campagne d'évaluation ESTER-1 (évaluation des systèmes de transcription enrichie d'emissions radiophoniques). ESTER-1 traite des données radiophoniques, l'idée était d'enrichir les transcriptions fournies par les systèmes de RAP par d'autres informations dont les entités nommées. Ainsi, une tâche d'extraction d'entités très proche de celle de MUC-6 (les entités se limitaient uniquement aux noms propres) a été mise en place [Le Meur *et al.*, 2004]. En 2007, une tâche d'extraction d'entités nommées proche de celle définie dans ACE avec une définition des entités nommées, qui ne se limite plus aux seules noms propres, a été mise en place durant la campagne d'évaluation ESTER-2 [Galliano *et al.*, 2009a].

À l'occasion du projet QUAERO [QUAERO, 2008] une nouvelle tâche de détection, de classification et de décomposition d'entités a été définie, et des règles d'annotations originales ont été mises en place dans le but d'extraire des informations plus riches. Cette nouvelle tâche a fait l'objet de deux campagnes d'évaluation : la campagne d'évaluation ETAPE en 2011 (évaluation en traitement automatique de la parole) et la campagne QUAERO en 2010-2011. Nous décrivons plus en détail la définition de cette tâche et les règles d'annotations utilisées dans le chapitre 3. Les données traitées dans les deux campagnes d'évaluation ETAPE [Galibert *et al.*, 2014] et QUAERO [Galibert *et al.*, 2010] étaient des données radiophoniques incluant de la parole spontanée et de la parole conversationnelle, les données QUAERO incluant également des données numérisées par OCR.

Ainsi, les campagnes d'évaluation françaises se sont intéressées, dès le début, principalement aux données orales étant donné les enjeux et les perspectives d'applications qu'offrait le traitement de ce type de données.

1.4 Applications

Le grand intérêt qu'a suscité la tâche de REN un peu partout dans le monde, est principalement motivé par les nombreuses applications dans lesquelles cette brique technologique peut être utilisée. En effet, la REN peut fournir des informations qui permettent d'améliorer des performances dans le cas d'applications complexes et peut également avoir des cas d'application directes dans lesquels elle constitue la brique principale.

1.4.1 Quelques applications indirectes

C'est en tant qu'information sémantique que les entités nommées peuvent être utilisées pour l'améliorations de certaines applications en TAL.
Dans [Brun et Hagege, 2004] et [Osenova et Kolkovska, 2002], les auteurs montrent que l'intégration des sorties d'un système de REN dans un analyseur syntaxique permet d'améliorer les performances de ce dernier, notamment grâce aux informations apportées par la classification.
La REN peut également contribuer dans le processus de désambiguïsation lexicale, notamment pour la désambiguïsation de paraphrases et de métonymies [M., 2003].
Les systèmes de traduction automatiques peuvent eux aussi exploiter les annotations en entités nommées pour repérer ces éléments qui font l'objet d'une translitération (traduction autonymique comme London qui devient Londres en français) plutôt que d'une traduction proprement dite.

1.4.2 Quelques applications directes

La REN constitue la brique principale des applications fondées sur l'extraction d'informations. Nous avons vu, par ailleurs, qu'historiquement la REN est apparue dans MUC parce qu'elle était le composant de base pour la tâche de remplissage de formulaires qui était la tâche proposée initialement. En effet, la REN est principalement employée dans les applications de constitution de bases de connaissances. Dans ces cas applicatifs le but est de constituer une base de données relative à un domaine ou à un sujet prédéfinis, en repérant ses acteurs et les éléments qui permettent de le caractériser. Un cas particulier de la constitution de bases de connaissances est la veille documentaire, qui consiste à surveiller et actualiser les connaissances sur un domaine donné afin de prédire son évolution. Ainsi, la REN constitue un outil intéressant qui permet d'aider les experts à fouiller de grandes bases de documents pour repérer et détecter les entités reliées au domaine ciblé [Poibeau, 1999].

Une autre application directe de REN est l'anonymisation des documents. L'anonymisation consiste à identifier et neutraliser certaines informations considérées comme confidentielles dans un ou plusieurs documents. Cette tâche est particulièrement importante dans les cas de documents médicaux et juridiques. Dans ces domaines, il est intéressant, pour les spécialistes, de partager les informations relatives à un cas ou à un patient avec d'autres experts (ou étudiants), afin de discuter avec eux et de leur demander leur avis sur un cas concret. Pour ce faire, il faut tout d'abord que les documents soient rendus anonymes pour rendre irrécupérable, c'est-à-dire non identifiable et non traçable, l'identité des personnes impliquées. Ce type d'informations concerne principalement des entités nommées qui peuvent être détectées et traitées par des systèmes de REN. Ainsi, la REN a été utilisée dans le traitement des documents cliniques [Grouin, 2013] et juridiques

[Plamondon *et al.*, 2004].

1.5 Conclusion

Dans ce chapitre nous avons dressé un bref historique des principales campagnes d'évaluation en REN. Ceci nous a permis de tracer l'origine de la tâche et de suivre les évolutions qu'a connues la définition de cette tâche, ainsi que ces méthodes d'évaluation. Les nombreuses campagnes d'évaluation qui ont été dédiées au traitement de la REN témoignent du grand succès de cette tâche. De fait, si la REN a connu ce grand succès c'est parce qu'elle constitue un enjeu important pour de nombreuses applications, que ce soit en tant que composante qui permet d'améliorer la chaîne de traitement globale, comme dans le cas de l'analyse syntaxique, de la désambiguïsation lexicale ou de la traduction automatique, ou en tant que brique principale, comme dans le cas de constitution de bases de connaissances, de veille technologique ou d'anonymisation. Elle est également utilisée pour traiter différents types de données dans des domaines distincts. Ainsi, la REN peut être utilisée dans des domaines assez généraux comme les textes journalistiques, et dans d'autres très spécifiques tels que le domaine médical, le domaine juridique, le domaine biologique, etc.

Il est malheureusement difficile de comparer les performances obtenues dans les différentes campagnes d'évaluation puisque les définitions de la tâche et les méthodes d'évaluation varient d'une campagne à une autre. Toutefois, les performances affichées sont très satisfaisantes notamment pour le traitement de données journalistiques. Le développement de systèmes robustes pouvant traiter des données bruitées a suscité très tôt un vif intérêt. En effet, une grande quantité de données se trouvant au format audio ou audiovisuel est intéressante à traiter, et ce, dans de nombreux domaines. Étant donné que la REN ne peut traiter actuellement que des données qui sont sous format texte, il est alors obligatoire de passer par des systèmes de RAP pour transcrire ces documents. Mise à part la différence entre la langue écrite et la langue parlée, laquelle introduit une difficulté supplémentaire, les systèmes de RAP font des erreurs de transcription qui font baisser les performances des systèmes de REN. C'est pourquoi, comme il a été dit *supra*, le développement de systèmes de REN robustes au bruit a suscité un vif intérêt très tôt. La première campagne d'évaluation ayant traité cette problématique est la campagne HUB-4 qui a encouragé le développement de système de REN pouvant traiter des données radio- et télédiffusées. Par la suite, de nombreuses campagnes d'évaluation ont intégré le traitement de ce type de données. Ainsi, toutes les campagnes de la série ACE incluaient le traitement de données radio- et télédiffusées transcrites automatiquement et manuellement, certaines campagnes incluant éga-

lement le traitement de données de type conversations radio- et télédiffusées et téléphoniques. Dans les campagnes d'évaluation françaises, il était question de traiter principalement des données radiophoniques, avec une proportion de paroles conversationnelles et de l'OCR dans QUAERO. L'intérêt qu'a suscité la REN de la langue parlée, dans toutes ces campagnes d'évaluation, illustre l'importance des enjeux reliés à cette application (reconnaissance d'entités nommées à partir de la parole) qui combine les systèmes de REN avec les systèmes de RAP et qui dépend donc nécessairement des performances de ces deux briques technologiques. Dans le chapitre suivant, nous allons nous intéresser à la brique de RAP, qui est la deuxième brique technologique traitée dans le cadre de cette thèse, afin d'en suivre l'évolution et d'en comprendre les enjeux applicatifs.

Chapitre 2

Historique et état de l'art en évaluation de la reconnaissance automatique de la parole

2.1 Introduction

L'invention de machines maîtrisant des capacités humaines, en particulier la maîtrise du langage, passionne depuis longtemps les chercheurs et les ingénieurs. Les premières expériences de reconnaissance automatique de la parole (RAP) remontent à 1930 [Dudley, 1939]. Depuis, la problématique de RAP a été abordée progressivement. Tout d'abord, en partant de la reconnaissance de quelques mots isolés pour ensuite aborder celle de la parole continue à grand vocabulaire en prenant en compte les différentes sources de variations pouvant toucher le signal de la parole, comme les variations liées aux changements de locuteur et aux conditions d'enregistrement.
Les grandes avancées qui ont été réalisées dans le domaine ont permis d'envisager l'intégration de RAP dans différents cadres applicatifs. Dans ce chapitre, après une brève définition de la tâche de RAP, nous passerons en revue les défis majeurs ayant été proposés dans les campagnes d'évaluation pour promouvoir le développement de cette technologie. Puis, nous discuterons de l'évolution des perspectives d'utilisation de la RAP et, par la suite, des défis organisés par les campagnes d'évaluation qui ont conduit à l'apparition de nouveaux besoins en termes d'évaluation. Nous passerons ensuite en revue les principales mesures d'évaluation proposées dans la littérature et visant à répondre à ces besoins.

2.2 Définition

La reconnaissance automatique de la parole (RAP) consiste à fournir une transcription textuelle et orthographique, à partir de la parole véhiculée par un signal audio. Il s'agit d'une brique technologique importante puisque elle permet d'avoir accès à de grande quantité de données sous format audio et audiovisuel ce qui offre des perspectives d'utilisation intéressantes. Ainsi, de nombreuses compagnes d'évaluation ont été organisées dans le but de promouvoir le développement de la RAP, pour mieux comprendre ses forces et ses faiblesses et cerner les applications possibles compatibles avec son niveau de maturité.

Les tâches, ou les défis, et les conditions d'enregistrement du signal de la parole ont varié d'une campagne à l'autre suivant les performances des systèmes de RAP de l'époque et les attentes des utilisateurs potentiels. La section suivante parcourt l'historique des principales campagnes d'évaluation, pour s'arrêter sur les étapes importantes de l'évolution de cette tâche de reconnaissance automatique de la parole.

2.3 La reconnaissance automatique de la parole : évolution et maturité vues à travers les principales campagnes d'évaluation

2.3.1 La reconnaissance de mots isolés et les premières métriques d'évaluation

En 1981, aux États-Unis, George Doddington et Thomas B. shalk ont organisé une évaluation qui s'intéressait pour la première fois à la technologie de la RAP [Doddington et Schalk, 1981]. Il s'agissait à l'époque d'évaluer les performances en reconnaissance de mots isolés sur des données multi-locuteurs. Le test comportait la prononciation des lettres de l'alphabet (de a à z) et de vingt mots différents :

- Les mots de zéro à neuf en anglais ;
- les mots : *start, stop, yes, no, go, help, erase, rubout, repeat, enter*.

Le choix des mots était guidé par des statistiques sur leur fréquence d'utilisation dans la langue. L'ensemble a été prononcé par seize locuteurs différents (huit hommes et huit femmes). Chaque locuteur a participé à neuf sessions d'enregistrement différentes, qui s'étalaient approximativement sur deux mois, et durant lesquelles il produisait en tout seize réalisations de chaque mot. Le corpus de test final comportait alors 5 120 prononciations de mots [Doddington et Schalk, 1981] plus les prononciations des lettres de l'alphabet. Sept participants ont soumis les sorties de leur système pour l'évaluation. La performance des systèmes était mesurée en taux d'erreur. Deux types d'erreurs ont été considérés :

- les substitutions : quand un mot est reconnu à la place d'un autre ;
- les omissions (rejet ou oubli) : quand le système ne sort pas d'hypothèse.

Cette campagne pionnière en évaluation de la reconnaissance automatique de la parole a permis de faire un constat sur les critères à prendre en compte dans l'évaluation des systèmes de RAP. Ainsi, les discussions dans les groupes de travail ont permis d'étudier les facteurs qui influencent les performances des systèmes de RAP et de décrire les procédures d'évaluation à mettre en place, ainsi que de possibles métriques permettant d'évaluer les performances [Pallett, 1985]. Les principales métriques proposées étaient les suivantes :

$$Taux\ de\ reconnaissance = 100 \times \frac{\text{Nombre de mots correctement reconnus}}{\text{Nombre de mots dans la référence}}$$

$$Pourcentage\ de\ substitution = 100 \times \frac{\text{Nombre de mots substitués}}{\text{Nombre de mots dans la référence}}$$

$$Pourcentage\ d'omission = 100 \times \frac{\text{Nombre de mots supprimés}}{\text{Nombre de mots dans la référence}}$$

$$Pourcentage\ d'insertion = 100 \times \frac{\text{Nombre des mots insérés}}{\text{Nombre de mots dans la référence}}$$

2.3.2 La campagne *Resource Management* et la reconnaissance de la parole continue

En 1987, le NIST (*National Institute of Standards and Technology*) organisait une campagne d'évaluation *The Resource Management Test* (RM). Cette campagne financée par l'ARPA (*Advanced Research Projects Agency*), avait pour but d'évaluer la possibilité d'utiliser la technologie de RAP pour gérer certaines tâches de la marine américaine avec des commandes vocales. Le test mis en place était limité à un lexique composé de 997 mots jugés utiles pour les applications visées [Pallett, 2003]. Les participants avaient pour tâche de développer des systèmes capables de traiter de la parole continue afin de fournir les transcriptions des différentes commandes vocales. Quatre évaluations ont eu lieu entre 1989 et 1992 dans le cadre de *Resource Management* et ont permis la mise en place de procédures d'évaluation automatisées. Le NIST était chargé de la construction des données et de la mise en place d'une procédure d'évaluation incluant l'alignement entre l'hypothèse et la référence et le calcul de scores. Durant cette campagne le WER (*Word Error Rate*, ou taux d'erreur mots en français) était la métrique mise en

place pour l'évaluation des performances des systèmes. Le WER est défini comme suit :

$$WER = \frac{(I + D + S) \times 100}{N}$$

Avec :

- D : le nombre d'erreurs d'omission, les mots qui existent dans la référence et ne s'alignent avec aucun mot de l'hypothèse ;
- I : le nombre d'erreurs d'insertion, les mots qui existent dans l'hypothèse et ne s'alignent avec aucun mot de la référence ;
- S : le nombre d'erreurs de substitution, les mots de la référence substitués par d'autres mots dans l'hypothèse ;
- N : le nombre de mots dans la référence.

La campagne RM a permis de mettre en place des plateformes et des outils d'évaluation pour tester la technologie de RAP dans le contexte d'une application réelle. Il était question de traiter de la parole continue, mais le vocabulaire était limité à mille mots connus à l'avance. Le meilleur système affichait un WER de 5 % lors de la dernière évaluation en 1991.

2.3.3 La tâche WSJ/NAB et la reconnaissance de la parole continue à grand vocabulaire

En 1992, la DARPA (*Defense Advanced Research Projects Agency*) finance le développement des systèmes de RAP à grand vocabulaire dans le cadre du projet *Human Language Technology*. Les données étaient issues de lecture à voix haute d'articles de journaux, plus précisément le *Wall Street Journal* (WSJ) et le *North American Business* (NAB). Ceci a donné lieu à la tâche connue sous le nom de WSJ/NAB [Young et Chase, 1998]. Le test incluait des mots hors vocabulaire, de la parole non native, et des données enregistrées avec différents types de microphones. Ainsi, il y avait plusieurs tâches ou sessions suivant la qualité des données à traiter. Quatre campagnes d'évaluation ont eu lieu entre 1992 et 1995. Certaines ont été des campagnes internationales et ont enregistré la participation de laboratoires français, anglais, allemands et canadiens [Pallett *et al.*, 1995]. Les performances obtenues oscillaient entre 22 % et 7,2 % d'erreurs [Pallett *et al.*, 1995]. Durant ces conférences, des procédures ont été mises en place pour mesurer l'impact des mots hors vocabulaire sur les performances des systèmes de RAP.

2.3.4 La campagne ATIS et la reconnaissance de la parole spontanée

La série de campagnes d'évaluation ATIS (*Air Travel Information Systems*), qui a été lancée au début des années quatre-vingt-dix, avait pour but d'encourager le dé-

veloppement de la technologie de la compréhension de la parole et de l'analyse sémantique de la langue parlée. En effet, les performances encourageantes affichées par les systèmes de RAP durant les évaluations précédentes (RM et WSJ/NAB) permettaient d'envisager la possibilité d'utiliser les sorties des systèmes de RAP pour accomplir des tâches plus complexes. Ainsi, le défi dans ATIS consistait à développer des systèmes intelligents qui possédaient la capacité de comprendre la requête d'un client (interlocuteur) voulant réserver un billet d'avion, et de lui fournir les bons renseignements. En plus de la reconnaissance de la parole, ATIS s'intéressait également à des technologies intégrant l'analyse syntaxique, l'analyse sémantique, la génération de textes et la synthèse de la parole. Pour la communauté de la RAP cette tâche a constitué un nouveau défi, qui est celui du traitement de la parole spontanée avec toutes les variations liées aux locuteurs, aux types d'élocutions et aux accents.

ATIS a représenté également un défi pour les évaluateurs, car il fallait constituer pour le test une base de requêtes correspondant à ce que des utilisateurs réels pourraient faire en interaction avec un système automatique. Ceci a été fait grâce à une simulation réalisée par des sujets humains [Hirschman, 1998].

Bien que les performances aient été faibles durant la première édition de ATIS (plus de 50 % d'erreurs), les performances se sont améliorées lors des éditions suivantes pour atteindre des taux d'erreur inférieurs à 10 % en compréhension de la parole. Il a été noté, à l'occasion des campagnes ATIS, que les systèmes de compréhension affichaient des taux d'erreur inférieurs à ceux des transcriptions automatiques sur lesquelles ils étaient appliqués. Il en a été déduit qu'il n'est pas nécessaire de reconnaître tous les mots d'un énoncé pour parvenir à en extraire son sens [Hirschman, 1998].

2.3.5 L'évaluation de la reconnaissance multilingue en Europe : le projet SQALE

Le projet SQALE (*Speech Quality Assessment for Linguistic Engineering*) a été financé par la Commission européenne dans le but d'adapter les paradigmes d'évaluation, développés dans les campagnes d'évaluation financées par l'ARPA, aux besoins européens. Le projet SQALE s'est déroulé entre 1993 et 1995 et a été le premier à s'attaquer à la problématique de la reconnaissance de la parole multilingue. Les langues traitées étaient l'anglais britannique, le français et l'allemand. Les données à traiter pour l'anglais et le français provenaient principalement de deux journaux, le *Wall Street Journal* et *Le Monde*. Pour l'allemand le corpus PHONDAT (un mélange d'énoncés et de requêtes en allemand) a été utilisé.
Les outils du NIST ont été utilisés pour l'évaluation avec des modifications mineures pour prendre en compte les caractères accentués qui existent en français et en allemand, mais pas en anglais. Les performances des systèmes de recon-

naissance obtenus étaient aux alentours de 15% d'erreurs pour les trois langues [Young *et al.*, 1997].
À l'issue de ce projet, il a été conclu que les méthodes et les outils d'évaluation développés par le NIST étaient fiables et facilement utilisables dans le cas des langues étudiées. Les méthodes de RAP, utilisées pour la modélisation et la reconnaissance de l'anglais américain, restent valide pour les langues étudiées mis à part les problèmes liés aux spécificités des langues tels que la présence d'un grand nombre d'homophones et de la liaison dans le cas du français.

2.3.6 La reconnaissance de la parole dans les données radiophoniques

Avec le développement rapide des capacités de stockage, de la puissance de calcul des machines, ainsi que les performances très satisfaisantes en RAP affichées par les approches fondées sur les apprentissages statistiques, les systèmes de RAP sont devenus très gourmands en termes de données. En 1995, et afin de répondre à ce besoin qui grandissait de plus en plus, l'intention du NIST a été attirée par les journaux radiophoniques (*Broadcast News*) qui étaient relativement faciles à collecter (par rapport à un enregistrement de corpus classique). C'est ainsi que le LDC (*Linguistic Data Consortium*) s'est chargé d'obtenir les droits pour enregistrer, transcrire et distribuer ces données à la communauté du traitement automatique des langues. Les données de type *Broadcast News* présentaient énormément de diversités. Les enregistrements obtenus pouvaient contenir de la parole préparée, de la lecture, de la parole spontanée, des disfluences et des niveaux très variables de bruits suivant le type d'émission et les locuteurs qui intervenaient dedans. Afin de prendre en compte cette variabilité lors de l'évaluation, il a été décidé de mesurer les performances des systèmes de RAP indépendamment suivant la nature de la parole enregistrée en plus d'une évaluation globale faite sur l'ensemble des données. Ainsi, dans une même évaluation sont présentes différentes sessions (appelées *Hub ou Spoke*) visant chacune à promouvoir l'étude et le traitement d'un aspect particulier [Young et Chase, 1998]. Voici ci-dessous quelques exemples de sessions :

- adaptations des modèles de langage ;
- adaptations aux variations liées aux changements de locuteur ;
- adaptations aux variations liées aux changements de microphone ;
- reconnaissance de la parole en milieu bruité.

2.3.7 La reconnaissance de la parole spontanée et conversationnelle

Le traitement de la parole spontanée et conversationnelle était une des principales motivations dans les programmes visant le développement de systèmes de

RAP à large vocabulaire. C'est ainsi que, en 1992, un corpus de conversations téléphoniques, appelé *Switchboard Corpus*, a été enregistré [Godfrey *et al.*, 1992] dans le but d'étudier ce type de parole qui s'est avéré plus difficile à traiter que la parole préparée [Young et Chase, 1998]. Ainsi, 2 500 conversations et 500 locuteurs ont était ainsi enregistrés. Les conversations portaient sur divers sujets imposés à l'avance à des locuteurs qui ne se connaissaient pas. Le *Switchboard Corpus* a servi à étudier et analyser les aspects de la parole conversationnelle. Mais il s'est avéré que ces conversations n'étaient pas naturelles puis qu'elles étaient enregistrées dans le cadre d'une expérience où les sujets des conversations étaient imposés et où les locuteurs n'étaient pas forcément très à l'aise en discutant avec un interlocuteur inconnu [Young et Chase, 1998].
Dans le but d'obtenir des données provenant de conversations spontanées, un programme d'enregistrement d'appels téléphoniques réels a été mis en place. La procédure consistait à offrir des appels gratuits aux personnes qui acceptaient que leur conversation soit enregistrée anonymement. Ceci a permis de construire, en 1997, ce qui a été appelé le *Callhome* corpus [Canavan *et al.*, 1997], qui contenait des sous-corpus en anglais, espagnol (dialectes variés), arabe (égyptien), mandarin, japonais et allemand.
Des taux d'erreur élevés par rapport à la reconnaissance de la parole préparée ont été enregistrés (40 % de WER pour l'anglais). Ceci a reflété la difficulté que représentait le traitement de la parole spontanée et la nécessité de mettre en place des méthodes et des outils permettant de gérer des phénomènes tels que les disfluences, les variations phonétiques, l'utilisation de mots que l'on ne retrouve pas forcément dans le langage écrit, etc.

2.3.8 Maturité de la technologie

Grâce aux différents programmes visant à promouvoir les travaux de recherche sur la RAP, les taux d'erreur ont baissé progressivement, reflétant les progrès réalisés. Ceci peut être vérifié en retraçant la courbe d'évolution du WER à travers les cycles des campagnes d'évaluation comme dans la figure 2.1 extraite de [Pallett, 2003]. Cette figure affiche l'évolution des performances des systèmes de RAP, enregistrée dans les campagnes d'évaluation organisées par le NIST entre 1988 et 2003, sur laquelle nous pouvons voir une chute du WER qui traduit l'amélioration des performances des systèmes de RAP sur différents types de données (parole préparée, spontanée, en milieu bruité, etc.). Même si la technologie n'était pas encore parfaite et que beaucoup de travail restait à faire pour la RAP en milieu bruités ou pour le traitement des langues peu dotées, les performances atteintes par la technologie RAP permettaient déjà d'envisager son intégration dans différents types d'applications.

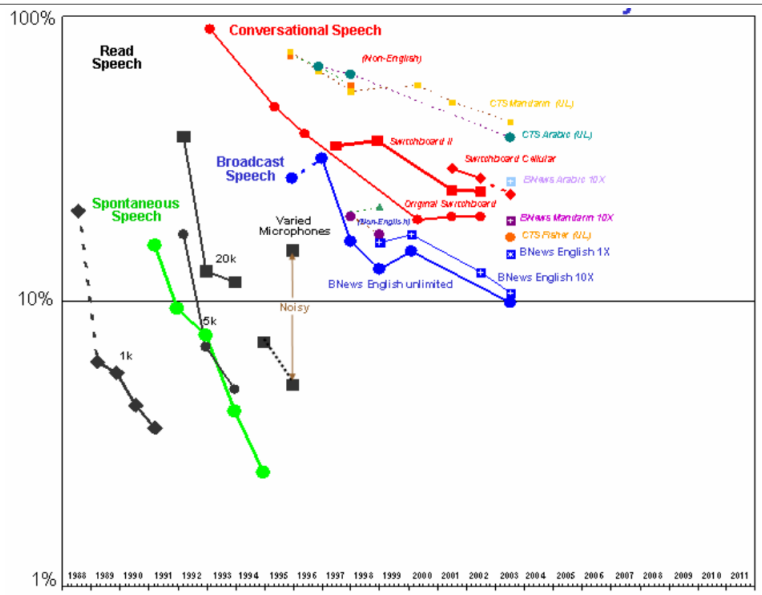

FIGURE 2.1 – Évolution des performances en reconnaissance automatique de la parole à travers les campagnes d'évaluation organisées par le NIST, entre 1988 et 2003, extraite de [Pallett, 2003].

2.4 De la reconnaissance automatique de la parole au traitement automatique de la parole

Les performances atteintes par les systèmes de RAP dans les campagnes d'évaluation ont montré que la technologie pouvait traiter de la parole continue à grand vocabulaire. Ainsi, les cas d'application de la RAP ne se limitent plus à la reconnaissance de quelques commandes vocales ou à une dictée automatique. Il est devenu possible d'envisager d'utiliser cette technologie pour traiter les grandes quantités de données se trouvant sous format audio ou audiovisuel. De plus, l'extension rapide d'Internet a également ouvert l'accès à de grandes quantités de données, dont une grande quantité en format audio ou vidéo, auxquelles s'ajoutent d'autres bases de données pouvant provenir des médias (télés, radios) ou des archives. Des applications visant à extraire des informations à partir de ces données, à les archiver, à les traduire ou à les résumer automatiquement nécessitent d'abord une phase de transcription dans laquelle interviennent les systèmes de RAP.

Dans le but de proposer des tâches qui reflètent les cas d'utilisation réels de la technologie, les campagnes d'évaluation se sont tournées vers des défis dans lesquels les systèmes de RAP sont couplés avec d'autres briques technologiques pour accomplir des tâches complexes telles que :

- la compréhension de la parole (*spoken language understanding*) qui consiste à analyser et à interpréter les sorties des systèmes RAP. Cette tâche a été

initiée par la campagne ATIS [Hirschman, 1998] comme nous l'avons mentionné dans la section 2.3.4 ;
- l'extraction d'informations à partir de la parole qui peut concerner plusieurs tâches avec des définitions plus ou moins différentes. Ainsi, nous pouvons mentionner la détection d'entités nommées à partir de la parole promue par la série de campagnes d'évaluation ACE qui se sont déroulées entre 2000 et 2008. Nous pouvons également retrouver l'extraction de mots-clés à partir de la parole qui a été promue également par la série des campagnes d'évaluation TREC (1997-2002) et, par la suite, par la série TRECVID (2000-2008) ;
- la transcription enrichie qui consiste à fournir des métadonnées liées aux sorties des systèmes de RAP. Les métadonnées peuvent concerner des informations de nature différente telles que la date, l'heure, l'origine du document, et peuvent inclure également des annotations en entités nommées. Les campagnes d'évaluation ESTER1 (2003-2005) [Galliano *et al.*, 2006], ESTER2 (2008-2009) [Galliano *et al.*, 2009b], QUAERO (2008-2014) [QUAERO, 2008] et ETAPE (2011-2012) [Gravier *et al.*, 2012] ont soutenu cette tâche ;
- la traduction de la parole qui consiste à traduire automatiquement les sorties des systèmes de RAP d'une langue source en une langue cible. Cette tâche a notamment été soutenue par les campagnes IWSLT (*International Workshop on Spoken Language Translation*) entre 2004 et 2014 et par GALE (*Global Autonomous Language Exploitation*) entre 2009 et 2014.

Même si les systèmes de RAP ne sont pas parfaits et produisent encore des erreurs, le grand succès de ces campagnes d'évaluation témoigne que la RAP est devenue aujourd'hui une brique clé pour de nombreuses applications. Étant donné cette évolution qui concerne les cas d'utilisation des systèmes de RAP, un certain nombre de questions se posent :
- Quel est l'impact des erreurs de transcription sur les applications qui les utilisent ?
- Est-ce que les mesures d'évaluation quantitatives telles que le WER, qui étaient introduites à l'époque où la RAP était une fin en soi, restent adaptées dans le cas des nouveaux cas d'utilisation des systèmes de RAP ?
- Comment peut-on mesurer l'impact de la qualité des sorties ASR dans le cas d'applications complexes ?

2.5 Problématique

2.5.1 Introduction

Les procédures d'évaluation automatiques et les métriques d'évaluation sont des outils très importants pour l'orientation de la recherche et le développement des technologies.
Simple et facile à interpréter, le WER a été longtemps utilisé pour mesurer et

optimiser les performances des systèmes de RAP. Comme nous l'avons mentionné, le WER consiste à énumérer les erreurs de transcription et à les normaliser par le nombre total de mots dans la référence pour fournir un pourcentage d'erreurs.

$$WER = 100 \times \frac{(I + D + S)}{N} \tag{2.1}$$

Avec, I, D et S qui, comme il a été dit *supra*, sont respectivement les nombres d'erreurs d'insertion, d'omission et de substitution déterminés par un alignement de type Levenshtein [Levenshtein, 1966] et N qui représente le nombre de mots dans la référence.
Ainsi, le WER affecte un poids égal pour toutes les erreurs, alors qu'il est raisonnable de considérer que le poids de chaque erreur va dépendre du cadre applicatif dans lequel le système de RAP sera utilisé. Dans le cas de la dictée automatique, par exemple, l'objectif est de réduire le nombre de corrections que l'utilisateur doit faire sur les sorties du système de RAP pour obtenir le résultat souhaité. Dans ce type d'application, le WER constitue une métrique pertinente puisque toutes les erreurs nécessitent une intervention manuelle de l'utilisateur. En revanche, dans d'autres applications, comme la recherche de mots-clés, il n'est pas impératif d'avoir une transcription parfaite si les informations recherchées sont elles-mêmes bien transcrites.
Pour illustrer cela avec un exemple concret, nous allons considérer deux systèmes de RAP, le premier qui donne en sortie le même nombre de mots que la référence, mais tous ces mots sont faux, et le second qui donne en sortie le double des mots de la référence, tous faux également. Ainsi, le premier système affiche un WER de 100 % et le deuxième un WER de 200 %. Dans ces cas les deux systèmes communiquent simplement zéro information et ne permettront pas de trouver les mots-clés recherchés. La différence en termes de WER est donc insignifiante.

2.5.2 Le WER et l'estimation des performances en traitement automatique de la parole

Une question que l'on peut légitimement se poser est la suivante :
Est-ce que le WER, qui a été initialement introduit pour évaluer les performances en reconnaissance automatique de la parole, reste fiable pour l'évaluation des systèmes de RAP dans le cas d'applications de traitement automatique de la parole ?
Dans la littérature nous pouvons trouver des points de vue divergents sur ce sujet. Certains pensent que les mesures offertes par le WER restent suffisamment informatives. D'autres déplorent des cas de non-pertinence liée au WER et réclament le développement de mesures plus adaptées pour l'évaluation des systèmes de RAP pour diverses applications en traitement automatique de la parole. Nous citons *infra* quelques études, représentant les deux points de vue et dans lesquelles cette

problématique a été abordée.

Dans [Munteanu *et al.*, 2006], les auteurs étudient l'impact de la dégradation de la qualité des transcriptions automatiques (mesurée en WER) sur les performances d'utilisateurs humains en archivage de données Web. Leur expérience a consisté à fournir des transcriptions de cours en ligne à des taux d'erreur différents à des sujets (48 personnes). Les sujets devaient ensuite répondre à un ensemble de questions visant à savoir s'ils avaient bien compris le thème du cours. Trois transcriptions ont été fournies avec des taux d'erreur de 0 % (manuelles), 25 % et 45 %. Les auteurs ont observé une relation linéaire entre le WER et le nombre de réponses correctes. Ainsi, une augmentation du WER entraîne une baisse du nombre des réponses correctes données par les sujets. Cette étude montre surtout qu'en cas de grande différence en termes de taux d'erreur le, WER reste une métrique fiable.
Une autre étude [Przybocki *et al.*, 1999] faite sur les données de la campagne d'évaluation HUB-4 portant sur l'extraction d'entités nommées à partir de la parole transcrite automatiquement, a montré qu'il y a une forte corrélation entre le WER et les performances en extraction d'entités nommées. Toutefois, les auteurs ont remarqué que le système de RAP qui était classé premier en termes de WER passe cinquième dans le classement fondé sur les performances des systèmes de REN.

À l'occasion de la conférence TREC (*Text REtrieval Conference*), le NIST a mis en place, une tâche de recherche d'information, qui consiste à sélectionner à partir de requêtes des documents oraux transcrits automatiquement par des systèmes de RAP ayant des performances différentes en termes de WER [Garofolo *et al.*, 1999].

Les évaluateurs se sont alors intéressés à mesurer l'évolution des performances en recherche d'informations en fonction de l'augmentation du taux d'erreur de la transcription automatique. Dans TREC les performances en recherche d'informations étaient mesurées en MAP (*Mean Average Precision*) qui était la moyenne des précisions calculée sur toutes les histoires du test. La performance en RAP était mesurée en SWER (*Story Word Error Rate*) qui est la moyenne des WER calculée sur toutes les histoires du test. La figure 2.2 tirée de [Garofolo *et al.*, 1999] montre l'évolution des performances des systèmes d'recherche d'informations en termes de MAP en fonction du SWER. Les courbes obtenues permettent de voir que les performances en recherche d'informations ont tendance à baisser quand le WER augmente. En revanche, cette tendance n'est pas absolue et nous pouvons voir que les meilleures performances en recherche d'informations ne sont pas forcément obtenues avec les transcriptions ayant le WER le moins élevé. Ceci conduit à l'obtention de classements différents par rapport à des systèmes de RAP selon que le classement est fondé sur le WER ou sur les performances des systèmes de recherche d'information. Afin d'obtenir des mesures qui permettent une meilleure

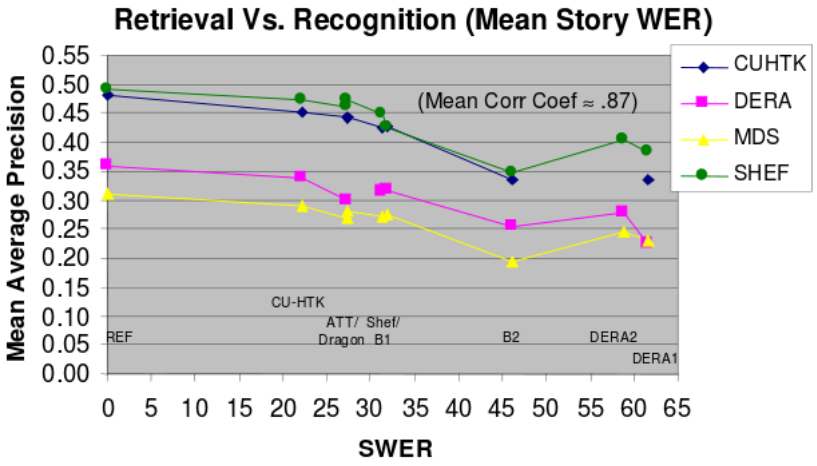

FIGURE 2.2 – Évolution des performances en recherche d'informations en fonction du SWER (*Story Word Error Rate*), conférence TREC, tirée de [Garofolo *et al.*, 2000].

prédiction que le WER pour les performances en recherche d'informations, les auteurs de [Garofolo *et al.*, 1999] ont essayé d'autres mesures que nous discuterons plus tard (section 2.6).

Nous terminons cette série d'exemples par la présentation d'une expérience menée par Y. Wang [Wang *et al.*, 2003] sur les données de la campagne ATIS qui portait sur la compréhension de la parole. Afin de mettre en évidence le fait que le WER n'était pas la métrique idéale pour mesurer la qualité des transcriptions automatiques pour la compréhension de la parole, Y. Wang avait proposé de simuler la tâche ATIS et de mesurer la variation des performances sur la tâche globale en fonction du WER. Pour ce faire, il avait développé trois systèmes de RAP ayant tous le même modèle acoustique et le même vocabulaire mais possédant des modèles de langue différents. Les trois systèmes de RAP ainsi obtenus affichaient des performances différentes en termes de WER. Sur les sorties de ces trois systèmes de RAP, les auteurs ont appliqué le même système de compréhension de la parole afin d'extraire les informations demandées. Enfin, les auteurs ont mesuré les performances sur la tâche finale en termes de Slot ID (équivalent au *Slot Error Rate*) [Makhoul *et al.*, 1999] et Task ID. Le Slot ID consiste à relever toutes les informations extraites par l'arbre de décision sémantique du système et de les comparer avec les informations annotées manuellement dans la référence. Le Task ID est un taux d'erreur de plus haut niveau que le Slot ID. Il est calculé par rapport aux erreurs d'extraction de thématique des documents. Les résultats de l'expérience sont affichés dans le tableau 2.1.

Ces résultats montrent que les transcriptions ayant le WER le plus bas 8,2 % entraînent les plus mauvaises performances pour la compréhension de la parole

	RAP-1	RAP-2	RAP-3
WER	8,2 %	12,3 %	12 %
Slot ID	11,6	11,1	9,8
Task ID	7,9	7,1	5,6

TABLEAU 2.1 – Résultats de l'expérience de Y. Wang [Wang *et al.*, 2003] sur la compréhension de la parole.

(Slot ID = 11,6 %). Les meilleures performances sont obtenues sur la transcription classée deuxième en termes de WER (12 %). Les auteurs de cette expérience ont déploré la divergence qui existait entre les mesures de qualité des transcriptions fournies par le WER et les performances en compréhension de la parole. Ils ont suggéré que les systèmes de RAP ne soient plus optimisés selon le WER mais selon le but applicatif final visé.
D'autres auteurs ont signalé des observations similaires sur les tâches de productions par des humains de résumés de séminaires préalablement transcrits automatiquement [Favre *et al.*, 2013] et sur la compréhension automatique de la parole [Riccardi et Gorin, 1998]. Ces auteurs ont déploré la non-pertinence des mesures offertes par le WER quand il s'agissait d'évaluer la qualité des sorties RAP dans un cadre spécifique et notamment dans le cadre de l'extraction d'informations à partir de documents transcrits automatiquement. De fait, ils demandent la mise en place de mesures plus adaptées à leurs besoins, permettant une meilleure prédiction de la dégradation des performances induite par les erreurs des transcripteurs automatiques. En effet, les erreurs de transcription peuvent avoir un impact important si elles modifient des informations pertinentes pour la tâche globale. En revanche, certaines erreurs ont une influence marginale et sont sans conséquence pour le résultat final. Nous discuterons dans la section suivante des principales mesures constituant une alternative par rapport au WER et qui sont proposées dans la littérature.

2.6 État de l'art des mesures d'évaluation de la qualité des transcriptions automatiques

Suite au besoin exprimé par la communauté du traitement automatique de la parole pour avoir des mesures d'évaluation permettant de dépasser les limitations du WER, un certain nombre de mesures alternatives ont été proposées dans la littérature. Nous présentons ici une liste non exhaustive de ces propositions.

Dans [Miller, 1955], les auteurs proposent de mesurer la perte d'information causée par les erreurs des systèmes de RAP. La mesure qu'ils proposent, le RIL (*Relative Information Loss*, perte relative d'information en français), est fondée sur le principe de l'information mutuelle (IM) [Papoulis et Pillai, 2002]. Étant donnée X l'ensemble des mots de la référence $x_1..x_n$ et Y l'ensemble des mots de l'hypothèse $y_1..y_n$, IM permet d'obtenir une mesure de la dépendance statistique entre le vocabulaire de la référence X et celui de l'hypothèse Y. Elle est représentée en termes d'entropie de Shannon telle que décrit dans l'équation 2.2.

$$IM(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{2.2}$$

Avec H, l'entropie de Shannon donnée par les équations 2.3 et 2.4.

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log P(x_i) \tag{2.3}$$

$$H(Y|X) = -\sum_{i,j} P(x_i|y_j) \log P(x_i|y_j) \tag{2.4}$$

Ainsi, le RIL est défini comme suit :

$$RIL = 1 - \frac{IM(X,Y)}{H(Y)} = \frac{H(Y|X)}{H(Y)} \tag{2.5}$$

Les probabilités dans les équations (2.3) et (2.4) sont estimées à partir d'une matrice de confusion déduite d'un alignement des mots de la référence et de l'hypothèse. La matrice est incrémentée en comptant les erreurs d'omission et d'insertion.
Mis à part le fait que le RIL est une métrique assez lourde à implémenter (notamment l'estimation de $H(Y|X)$), il ne permet pas de prendre en compte les erreurs systématiques (qui se répètent toujours de la même façon) puisque $P(x_i|y_j)$ converge vers 1 si un mot donné de la référence est toujours remplacé par le même mot dans l'hypothèse, ce qui entraîne que son terme dans la somme H(X|Y) sera proche de zéro. Ce cas peut être particulièrement accentué pour les mots peu fréquents (dans le test), pouvant être des mots hors vocabulaire et potentiellement des entités nommées (informations riches sémantiquement). Le RIL est également une métrique très dépendante de l'alignement comme le montre l'étude de [Maier, 2002] dans laquelle, l'auteur teste plusieurs algorithmes d'alignement et montre que les mesures données par le RIL changent significativement avec le type d'alignement.
Une approximation du RIL, le WIL (*Word Information Lost*, perte d'information liée à des erreurs de transcription de mots), a été proposée dans [Morris *et al.*, 2004]. Le WIL est fondé sur une matrice de confusion des mots de la référence et de l'hypothèse comme le RIL. Mais, contrairement au RIL, la matrice prend en compte les mots corrects et les substitutions. Il a été montré dans les études [Morris, 2002] et

[Morris *et al.*, 2004] qu'une telle matrice est dominée par la contribution des mots corrects. Ainsi, l'inverse du RIL, le WIP (*Word Information Preserved*), peut s'écrire comme suit :

$$WIP = \frac{C}{N1}.\frac{C}{N2} \cong \frac{IM(X,Y)}{H(Y)} \tag{2.6}$$

$$WIL = 1 - WIP \tag{2.7}$$

avec C, le nombre de mots correctement transcrits, N1 et N2 qui sont respectivement les nombres de mots dans la référence et le nombre des mots dans l'hypothèse.

Les expériences menées par [Morris *et al.*, 2004] et [McCowan *et al.*, 2004] ont montré que le RIL et le WIL peuvent être des mesures intéressantes lorsque les systèmes de RAP possèdent des taux d'erreur élevés.

D'autres mesures [1] d'évaluation inspirées par le RIL ont été proposées dans [McCowan *et al.*, 2004] où les auteurs proposent d'adapter les métriques utilisées pour l'évaluation des systèmes d'extraction d'informations (précision, rappel et F-mesure) pour mesurer la perte d'information causée par les erreurs de RAP. L'idée consiste à mesurer le rappel, la précision et la F-mesure en se basant sur le résultat de l'alignement entre les transcriptions de l'hypothèse et les transcriptions de la référence. Ainsi, la RAP est plutôt vue comme un problème d'extraction d'information, dans lequel chaque mot constitue une information à trouver. Par la suite, le rappel R et la précision P sont données par les équations 2.8 et 2.9.

$$R = \frac{C}{N_R} \tag{2.8}$$

$$P = \frac{C}{N_H} \tag{2.9}$$

avec C, le nombre de mots correctement transcrits (déterminés à partir de l'alignement), et N_R et N_H qui sont respectivement les nombres de mots dans la référence et dans l'hypothèse. Par la suite, la F-mesure est donnée par une moyenne harmonique entre R et P.

Suite aux incohérences entre les classements donnés par le WER et ceux donnés par les performances des systèmes d'extraction d'informations durant la campagne TREC (voir figure 2.2), une série de mesures alternatives a été proposée dans [Garofolo *et al.*, 1999] dans le but chercher une meilleure corrélation entre les performances des systèmes de RAP et les performances finales en extraction

1. Nous appelons mesures toutes les méthodes d'évaluation ne mesurant pas une simple distance entre un échantillon (sortie de systèmes) et un étalon (une référence respectant les normes).

d'informations. Dans cette étude les auteurs proposent de mesurer un taux d'erreur qui donne plus de valeur à des mots considérés comme importants dans le cadre applicatif visé. Les mesures testées incluaient le NE-WER, qui est un taux d'erreur calculé uniquement sur les entités nommées présentes dans le corpus test, et de nombreux autres taux d'erreur fondés sur l'utilisation d'anti-dictionnaires (stop-lists) qui permettaient à chaque fois de ne pas tenir compte des erreurs commises pour une catégorie de mots, telle que les prépositions, les articles ou les conjonctions. Ce qui distingue les travaux de [Garofolo *et al.*, 1999] des travaux présentés précédemment, c'est le fait d'avoir testé les mesures proposées sur des données réelles issues de campagnes d'évaluation [Garofolo *et al.*, 2000]. C'est le NE-WER, donné par l'équation 2.10, qui a permis d'obtenir les meilleures corrélations avec les performances des systèmes d'extraction d'informations.

$$\text{NE-WER} = \frac{D_{NE} + I_{NE} + S_{NE}}{N_{NE}} \tag{2.10}$$

avec, D_{NE}, I_{NE} et S_{NE} qui sont respectivement les nombres d'erreurs d'omission, d'insertion et de substitution de mots appartenant à des entités nommées. N_{NE} est le nombre de mots appartenant à des entités nommées dans les transcriptions de référence.

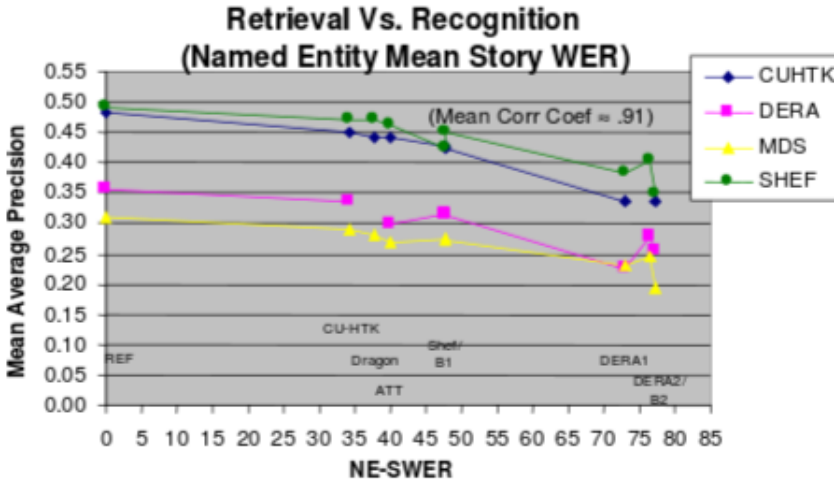

FIGURE 2.3 – Évolution des performance en extraction d'information en fonction du NE-SWER (*Named Entities Story Word Error Rate*), conférence TREC, tirée de [Garofolo *et al.*, 2000].

La figure 2.3 montre l'évolution des performances des systèmes d'extraction d'informations en fonction du NE-WER. En comparant ces résultats avec ceux obtenus en utilisant le WER (figure 2.2), nous pouvons remarquer que le NE-WER affiche une meilleure corrélation (corrélation de Pearson) que le WER avec les performances des systèmes de REN (0,87 *VS* 0,91). Ceci se manifeste par un redressement des courbes des performances des systèmes d'extraction d'informations. En

revanche, le problème du classement des systèmes de RAP observé dans le cas du WER persiste avec NE-WER.

2.6.1 Discussion

À travers cette lecture de l'état de l'art (section 2.5 et section 2.6), nous pouvons voir que le WER n'est pas toujours la meilleure mesure à utiliser pour l'évaluation des systèmes de RAP. Ainsi, une meilleure compréhension de l'impact des erreurs de transcription automatique passe forcément par le développement de meilleures mesures (ou métriques) d'évaluation permettant de prendre en compte le contexte applicatif dans lequel les systèmes de RAP sont utilisés. Quelques mesures alternatives par rapport au WER peuvent être trouvées dans la littérature. Nous trouvons le principe de mesure de la perte d'information utilisé notamment dans le RIL et le WIL très intéressant. Le calcul du taux d'erreur ciblé sur les zones de textes porteuses de sens pour l'application visée, comme dans le cas du NE-WER et d'anti-dictionnaire, semble aussi être une piste à creuser. Toutefois, la plupart des méthodes restent très proches du WER puisqu'elles sont fondées sur le même principe de comparaison entre la référence et l'hypothèse *via* un alignement qui détermine les erreurs d'omission, d'insertion et de substitution de mots. Prouver la validité de ces mesures en les testant sur des données réelles est également une phase nécessaire pour confirmer leur robustesse.

2.7 Conclusion

Dans ce chapitre nous avons dressé un bref panorama de l'évolution des défis en RAP dans les campagnes d'évaluation. Ces défis sont passés de la reconnaissance de mots isolés à la reconnaissance de la parole continue à grand vocabulaire, ouvrant ainsi la voie à un large spectre de perspectives d'applications. Ainsi, même si la technologie de RAP n'est pas parfaite et que beaucoup de travail reste à faire pour prendre en compte les diverses sources des variations (bruit, parole spontanée, parole non native, langues peu dotées...), les campagnes d'évaluation ont commencé à s'intéresser à des tâches tournées vers le traitement automatique de la parole. Dans celles-ci la RAP est utilisée comme une brique qui permet d'accéder au contenu des documents audio (ou multimédias) en fournissant les transcriptions automatiques correspondantes. Ces transcriptions constituent, en effet, une précieuse ressource pour d'autres briques technologiques qui appliquent dessus des traitements tels que l'extraction d'informations, la traduction, l'indexation des documents... Bien évidemment, en parallèle à ces évaluations, la RAP a continué de progresser. Mais, malgré les progrès, les erreurs de transcription persistent. Ces erreurs influencent le fonctionnement des briques utilisant en entrée les transcrip-

tions automatiques, faisant ainsi baisser les performances de l'application finale.

Il est donc important de pouvoir évaluer la qualité des transcriptions automatiques selon le cadre applicatif visé. Le WER, qui est la métrique classiquement utilisée pour l'évaluation des systèmes de RAP quand ils sont utilisés seuls, dans le cas de la dictée automatique, n'est pas toujours la meilleure mesure à utiliser dans le cas d'applications complexes du traitement automatique de la parole. Ainsi, de nombreux cas d'incohérence ont été signalés entre les mesures données par le WER et celles obtenues pour l'application finale. Ceci a exacerbé le besoin de nouvelles mesures d'évaluation permettant d'évaluer la qualité des transcriptions automatiques selon le cadre applicatif visé.

D'autres métriques visant à mesurer la perte d'information causée par les erreurs de transcription ont été proposées. Mais la plupart de ces métriques n'ont pas été testées sur des données réelles provenant de campagnes d'évaluation, ni beaucoup utilisées par la communauté du traitement automatique de la parole.

La problématique de l'évaluation des briques technologiques imbriquées dans des applications complexes et qui tient compte du cadre applicatif est une problématique qui intéresse le Laboratoire national de métrologie et d'essais. Il la perçoit comme un enjeux majeur pour l'évaluation de nombreux produits issus de la recherche technologique. C'est aussi un enjeu important qui permettra au LNE de proposer à ses clients un étalonnage correspondant à leurs besoins applicatifs.

Dans le cadre de cette thèse, nous nous focalisons sur le cas de la reconnaissance d'entités nommées à partir de la parole. C'est est une tâche pour laquelle le LNE a été chargé de l'organisation des campagnes d'évaluation ETAPE (2011-2012) [Gravier *et al.*, 2012] et QUAERO (2008-2014) [QUAERO, 2008]. Après une description de la tâche de reconnaissance d'entités nommées étudiée et des outils permettant de l'évaluer dans la partie II, nous présentons dans la partie III une nouvelle mesure pour l'évaluation de la qualité des transcriptions automatiques dans le cadre de la reconnaissance d'entités nommées et nous la comparons à celle de l'état de l'art. Pour cela, nous nous appuyons sur des données réelles issues des campagnes d'évaluation.

Deuxième partie

Définition et évaluation des entités nommées structurées et compositionnelles

Chapitre 3

Les enjeux de la modélisation de la tâche de reconnaissance d'entités nommées

3.1 Introduction

Nous avons vu dans le chapitre 1 que la tâche de REN (reconnaissance d'entités nommées) introduite pour la première fois durant la campagne MUC (*Message Understanding Conference*) [Sundheim, 1996] a été un grand succès et qu'elle a été ensuite adoptée dans différentes campagnes d'évaluation telles que ACE (*Automatic Content Extraction*) [NIST, 2000], HAREM [Santos *et al.*, 2006], EVALITA [Speranza, 2007], ESTER [Galliano *et al.*, 2009a], avec, à chaque fois, des changements qui touchent les règles d'annotations et la définition des entités nommées. Ces changements étaient justifiés par la nécessité de résoudre les cas d'ambiguïtés et par la volonté de satisfaire un public aux attentes variées en termes de besoin applicatif.

Tout d'abord, nous discutons dans ce chapitre des principales difficultés auxquelles sont confrontés les concepteurs de la tâche et qui ont conduit à l'évolution et à la diversification des règles d'annotations en REN. Nous présentons, ensuite, la typologie des entités nommées et les règles d'annotations mises en place dans le cadre du projet QUAERO, puis nous décrivons la structure des entités nommées structurées et compositionnelles qui en résulte. Enfin, nous donnons un descriptif des corpus ETAPE [Gravier *et al.*, 2012] et QUAERO [Galibert *et al.*, 2010] annotés selon les nouvelles règles d'annotations mises en place dans le cadre du projet QUAERO.

Dans le cadre de cette thèse, nous allons nous intéresser particulièrement aux entités nommées structurées et compositionnelles dont la tâche résultante a fait l'objet de deux campagnes d'évaluation en français QUAERO [Galibert *et al.*, 2010] et ETAPE [Galibert *et al.*, 2014]. Les deux campagnes d'évaluation incluaient la re-

connaissance d'entités nommées à partir de transcriptions automatiques dont les données nous sont accessibles.

3.2 La modélisation de la tâche de reconnaissance d'entités nommées

L'extraction d'entités nommées vise à dégager, à partir de données textuelles, les unités porteuses de sens dans le but de les exploiter dans divers types d'applications comme l'analyse sémantique, l'indexation de documents ou la création de bases de données. La mise en place d'une telle tâche nécessite au moins deux étapes importantes :

- la définition des éléments à reconnaître ;
- la mise en place de règles d'annotations permettant de respecter la définition des entités et de gérer les exceptions et les cas d'ambiguïtés.

Les règles d'annotations résultantes serviront à mettre en place un guide d'annotation qui servira de référence pour les annotateurs pendant la phase de création de corpus et qui permettra également d'éclairer les développeurs des systèmes de REN sur la tâche demandée.

3.2.1 La complexité de la définition des typologies

La première question qui se pose lorsque l'on s'intéresse à la formalisation de cette tâche, que ce soit pour faire de l'annotation de données ou pour développer des systèmes de REN, est celle-ci : que veut-on reconnaître précisément ? Cherche-t-on à reconnaître des noms de personnes, des noms de plantes, des noms de pièces de théâtre, ou encore des noms de médicaments ? Les critères de classification dans des tâches similaires, comme celle de la classification de noms propres, divergent. Certains défendent une catégorisation fondée sur des critères linguistiques et pragmatiques [Daille et Morin, 2000], d'autres penchent pour des typologies morphosyntaxiques et sémantiques [Friburger, 2002].
Définir une typologie consiste à affecter pour une même classe un ensemble d'objets qui possèdent un certain nombre de critères en commun. Étant donné la nature de la tâche d'annotation en entités nommées, les critères communs les plus utilisés, lors de la mise en place d'un guide d'annotation pour la tâche de REN, sont de nature sémantique.
Converger vers une définition précise de ce que l'on veut reconnaître dépend fortement de l'existence d'un cadre applicatif précis, alors que l'objectif sous-jacent dans de nombreuses campagnes d'évaluation, est de montrer que la technologie de REN possède une large couverture et peut être utilisée dans un vaste éventail de domaines, dans le but de toucher le plus large public possible.
C'est dans la recherche de cet équilibre entre généricité et couverture de domaines

spécifiques que réside la complexité du travail de catégorisation.

3.2.1.1 La catégorisation et les problèmes de couverture et de pertinence

Nous avons vu dans le chapitre 1 que le choix des catégories ainsi que leur nombre ont beaucoup changé à travers les campagnes d'évaluation. Les définitions de la tâche de REN données sont plus ou moins différentes les unes des autres et permettent de voir à quel point il est difficile de se mettre d'accord sur une typologie universelle. Des similarités entre les typologies adoptées dans différentes campagnes d'évaluation existent du fait que l'on s'inspire souvent de travaux antérieurs. Ainsi, on retrouve toujours la triade Personnes, Organisations et Lieux, adoptée lors de la campagne MUC. Hormis ces trois catégories habituelles, les choix des typologies divergent. Ainsi, nous pouvons rencontrer des catégories allant des Armes et des Véhicules (ACE) [NIST, 2005], en passant par des Travaux artistiques et des Évènements (IREX) [Sekine et Isahara, 2000], jusqu'à des Oeuvres, des Abstractions et des Choses (HAREM) [Santos *et al.*, 2006], par exemple. De fait, il est difficile d'expliquer pourquoi on choisit de traiter telle catégorie d'entités et non telle autre. Certaines catégories peuvent être imposées par les financeurs des campagnes d'évaluation. Ainsi, dans la campagne d'évaluation ACE-05 [NIST, 2005], il fallait faire la distinction entre plusieurs types d'armes différentes (biologiques, armes blanches, armes chimiques, explosifs, armes de tir...). S. Sakine considère que dans le cas d'applications en domaine ouvert tel que les systèmes de questions-réponses et certains systèmes d'extraction d'informations, l'utilisation de sept ou huit catégories n'est pas suffisante. Il a ainsi construit une typologie composée de cent cinquante classes différentes [Sekine *et al.*, 2002], qu'il a ensuite étendue à deux cents classes [Sekine et Nobata, 2004] et ce, dans le but d'offrir une typologie générale qui permette de couvrir au mieux le domaine des articles journalistiques.
Les divergences dans les stratégies de choix des classes à adopter sont aussi motivées par des visions différentes du problème ou/et par les difficultés rencontrées lors de la création des corpus annotés. Ainsi, la confrontation à la réalité des données lors des annotations a fait apparaître des classes complexes, comme la classe GSP (*Geographical- Social-Political entities*) introduite durant la campagne d'évaluation ACE. Cette classe inclut des entités désignant des emplacements géographiques ayant des aspects sociaux et politiques, comme par exemple, dansles phrases :

- *La France a signé l'accord*
- *Aujourd'hui la Suisse est en grêve*

Les entités GSP réfèrent simultanément à plusieurs types : un lieu géographique (un pays), une organisation (un État), une population (les citoyens), par exemple. De telles catégories sont apparues afin de faciliter le travail d'annotation, de réduire les hésitations et d'améliorer les accords inter-annotateurs [NIST, 2000].
La difficulté de créer une typologie à large couverture à conduit lors des conférences CoNLL 2002 et 2003 [Tjong Kim *et al.*, 2003] à l'apparition d'une nouvelle classe « Miscellaneous » qui renferme toutes les entités détectées qui ne rentrent pas dans les autres classes. Ce type de classe a été ensuite adopté dans plusieurs projets qui ont suivi. Ainsi, nous trouvons la classe « Artifact » dans IREX [Sekine et Isahara, 2000] et la classe « Variado » dans HAREM [Santos *et al.*, 2006], lesquelles servent à aspirer les entités qui n'appartiennent a aucune classe dans la typologie.
Toutes ces divergences peuvent être expliquées par l'absence d'un cadre applicatif précis visé et par la volonté de couvrir plus de catégories afin de toucher un public plus large. Elles témoignent de la difficulté du travail liée aux définitions de catégories d'entités.

3.2.1.2 La catégorisation et les difficultés de spécialisation et de précision

Mis à part les difficultés liées à la couverture et à la catégorisation des entités, le problème de spécialisation constitue aussi un enjeu important, car même si les entités d'une même classe possèdent des caractéristiques communes, elles ne sont pas forcément homogènes. Prenons par exemple la catégorie Personnes, nous pourrions trouver dans cette catégorie des entités telles que :

- *François Hollande*
- *l'épouse Chirac*
- *Zizou*
- *les démocrates*
- *les Italiens*
- *Zorro*
- *Mickey*
- *saint. Nicolas*
- *Odin*

Dans MUC, il a été tout simplement décidé de n'annoter que les prénoms et les noms de famille. D'autres projets ont choisi de mettre en place des sous-catégories. Afin de faire des classifications plus précises, le nombre des sous-classes considérées a augmenté rapidement. Ainsi pour la classe des personnes, on trouve dans ACE [NIST, 2000] les sous-catégories Individus, Groupes et Indéfinis, et dans ESTER les sous-catégories Humain, Animal et Imaginaire. Il en va de même pour les autres classes, que ce soit les organisations où nous pouvons trouver privée,

publique, à but non lucratif, religieux, politique... ou encore les lieux qui peuvent être classés en ville, pays, continent, adresse postale, montagne...
Opter pour une typologie fine et multiplier le nombre de classes demandent des efforts supplémentaires de la part des concepteurs de typologies et des développeurs de systèmes. Mais cette stratégie présente l'avantage d'être facilement adaptable à des tâches moins précises. En revanche, une typologie très spécialisée devient difficile à adapter. Toutes ces difficultés mettent en évidence la délicatesse de la tâche de classification et expliquent les divergences que nous pouvons observer dans les guides d'annotation d'un projet à une autre.

3.2.2 Règles d'annotations et ambiguïtés linguistiques

Une fois la typologie définie, il reste à définir les règles d'annotations qui permettent de spécifier les pratiques à adopter pour annoter les parties du texte qui constituent des entités. La mise en place de ces règles se confronte aux ambiguïtés liées à la langue et qui ne sont pas forcément prévisibles pendant la phase de modélisation. La mise en place des règles d'annotations obéit à des démarches expérimentales et à des phases de validation que nous n'allons pas aborder dans cette thèse. Nous allons plutôt nous intéresser aux ambiguïtés linguistiques et syntaxiques à l'origine des divergences dans la définition des règles d'annotations.

3.2.2.1 Les ambiguïtés sémantiques

Les mots ou les expressions possédant plusieurs sens différents constituent une source d'ambiguïtés pour plusieurs applications de traitement automatique des langues. La reconnaissance d'entités nommées n'est pas épargnée et nous pouvons distinguer deux phénomènes linguistiques particulièrement importants dans le cadre de la REN :

- l'homonymie ;
- la métonymie.

L'homonymie : Nous parlons d'homonymie lorsque des mots d'une même langue possèdent la même forme écrite ou orale mais avec des sens différents. Les mots qui possèdent la même forme orale sont des homophones. Ils posent des problèmes plutôt pour le traitement de la parole que de l'écrit. Les mots qui possèdent la même forme écrite sont des homographes et ils sont à l'origine de confusions pour la tâche de reconnaissance d'entités nommées. Prenons comme exemple d'homographie le mot « orange » qui peut désigner le fruit (J'ai mangé une *orange*), la ville (*Orange* est située à 118 km de Marseille) ou encore l'organisation (Il travaille chez *Orange*). Un autre exemple, « Charles de Gaulle » qui peut désigner la personne, le porte-avions, l'aéroport, une rue... À notre connaissance, pour ce phénomène,

il n'y a pas eu de règles d'annotations spécifiques, l'interprétation de l'entité et sa classification demandent une désambiguïsation en fonction du contexte d'apparition, comme c'est le cas pour de nombreuses tâches en traitement automatique des langues.

La métonymie : La métonymie est une figure de style dans laquelle un mot lié à une entité est utilisé pour en désigner une autre, une relation quelconque relie les deux entités (cause à effet, partie de tout, contenant pour contenu, le lieu ou le producteur pour la production, etc.). Ainsi, le sens du mot peut renvoyer simultanément vers deux entités différentes. Par exemples « l'Élysée » peut désigner le bâtiment (le palais) et/ou l'organisation (le gouvernement), « Allemagne » peut désigner le pays et/ou l'organisation (le gouvernement).
Les règles de gestion de la métonymie diffèrent d'une campagne d'évaluation à une autre. Alors que durant MUC ce phénomène n'était pas pris en compte, durant le cycle des campagnes ACE des règles spécifiques ont été mises en place pour le gérer et, nous pouvons ainsi distinguer principalement trois démarches.
La première démarche a été utilisée durant les premières versions de la campagnes ACE. Elle consistait à adopter des définitions par défaut. Ainsi, par exemple, tous les pays sont considérés comme des lieux, tous les bâtiments religieux ou rendant des services sont considérés comme des organisations, etc. Cette solution a permis un accord sur la façon d'annoter ce type d'entités et, ainsi, a facilité le travail des annotateurs. Néanmoins, elle a fait émerger beaucoup de questionnements quant à la pertinence de la classe par défaut avec le rôle effectif de l'entité dans le texte.

La deuxième démarche, introduite à l'occasion d'ACE-03, consistait à regrouper certains cas de métonymie dans la classe GSP que nous avons présentée précédemment et qui concerne les emplacements géographiques ayant des aspects sociopolitiques. Ceci a surtout permis de faire la distinction entre les entités à caractère purement géographique et les entités qui désignent des villes ou des États et qui sont souvent utilisées pour faire référence aux organes politiques qui les dirigent.
La troisièmes démarche, introduite également plus tard durant ACE, consistait à affecter deux étiquettes (labels) aux entités de la classe GSP, avec une première étiquette qui désigne le type explicite (sens premier) de l'entité et une seconde qui désigne le type le plus en accord avec le contexte. Cette solution a permis de disposer de plus d'informations pour mieux désambiguïser les entités de la classe GSP.
La plupart des campagnes d'évaluation en extraction d'entités nommées qui ont suivi, ont adopté l'une de ces trois démarches pour traiter le problème de la métonymie. Dans IREX [Sekine et Isahara, 2000] une étiquette optionnelle « Optional » est ajoutée en cas de confusion. Une solution similaire a été adoptée dans HAREM

[Santos *et al.*, 2006], elle consiste à utiliser l'étiquette « Alt » (alternative) pour les entités pouvant appartenir à différentes classes. Dans EVALITA [Speranza, 2007] et ESTER [Le Meur *et al.*, 2004] les concepteurs ont opté pour l'utilisation de la classe GSP pour gérer les métonymies.

3.2.2.2 La difficulté du choix des frontières

Mis à part les ambiguïtés sémantiques, l'annotation des entités nommées est confrontée à d'autres ambiguïtés liées à la structure syntaxique des phrases. En effet, des cas de coordination ou d'imbrication entre entités ou avec d'autres composants de la phrase peuvent créer des confusions, notamment sur la délimitation des frontières de début et fin de chaque entité. Ces types d'ambiguïtés doivent également être pris en compte dans les règles d'annotations.

3.2.2.2.1 Les mots-outils et les descripteurs Les entités nommées apparaissent souvent entourées de mots pouvant être des déterminants, des articles, des titres ou des descripteurs qui sont liés sémantiquement aux entités. Prenons, par exemple, les syntagmes suivants :

- *le Premier ministre Manuel Valls*
- *Monsieur Dupont*
- *le psychiatre Charles*
- *Louis XIV*
- *abbé Pierre*
- *Le Mans*

Durant la campagne MUC les entités nommées concernaient uniquement les noms propres. Ainsi, les titres et les descripteurs n'étaient pas inclus dans ce qu'il fallait annoter. Avec l'extension du concept d'entité nommée, qui ne se limite plus depuis ACE aux noms propres, les avis sur ce qu'il faut annoter divergent. Certain ont décidé d'annoter les titres, comme « *Monsieur* », « *Madame* » et ne pas annoter les descripteurs comme « *Premier ministre* » ou « *psychiatre* », d'autres annotent les deux. Là encore, la réponse à cette question dépend du contexte applicatif visé.

3.2.2.2.2 La coordination Deux ou plusieurs entités peuvent être liées par une coordination et avoir ainsi des éléments en commun. Par exemple :

- *M. et Mme Chirac*
- *les entreprises françaises et allemandes*
- *le nord et le sud de la France*
- *des billets de 10, 20 et 50 euros*

- *ministre de l'Économie et des Finances*

Les avis sur la manière de procéder pour annoter les cas de coordination divergent. Faut-il considérer qu'il s'agit de deux entités différentes et dans ce cas quelle frontière choisir ? Ou bien faut-il considérer qu'il s'agit d'une seule entité, pour faciliter l'annotation ? Déjà, dans MUC, il y a eu des divergences. Alors que dans MUC-6 [Grishman et Sundheim, 1996] il avait été décidé d'annoter séparément les entités liées par une coordination, dans MUC-7 [SAIC, 1998], les entités liées par une coordination ont été considérées comme une seule et unique entité. Dans les campagnes ACE [NIST, 1999] l'indécision persiste. Ainsi, nous pouvons observer une fluctuation entre la première et la deuxième manière de procéder selon l'avis des concepteurs sur la question, sans toutefois vraiment donner de vrais justificatifs sur leurs choix.

3.2.2.2.3 Les imbrications Il y a imbrication quand une entité nommée en englobe une autre plus petite. Les entités nommées imbriquées constituent une autre source d'ambiguïtés. Les cas d'imbrications peuvent exister entre des entités de même types ou de type différents, par exemple :

- le comité d'entreprise de Renault (un « comité » est une organisation et « Renault » est une organisation aussi) ;
- la reine d'Angleterre (le mot « reine » désigne une fonction et « Angleterre » un pays) ;
- l'université de Grenoble (une « université » est une organisation et « Grenoble » est une ville) ;
- l'entraîneur du club de Paris (le mot « entraîneur » désigne une fonction, « club » une organisation et « Paris » une ville).

Séparer les entités imbriquées et les annoter indépendamment les unes des autres engendrent une perte d'information, considérer qu'il s'agit d'une seule entité et opter pour une annotation globale posent la question de savoir quelle entité favoriser et pour quelles raisons, alors que opter pour une annotation récursive (annoter les entités qui se trouvent à l'intérieur d'autres entités) augmente la complexité de la tâche d'annotation.

3.3 Les entités nommées structurées et compositionnelles

Les entités nommées étendues, mises en place dans le cadre du projet QUAERO [Grouin *et al.*, 2011], se distinguent par leur structure hiérarchique et compositionnelle. La tâche de reconnaissance d'entités nommées qui en découle est en cohérence avec les tâches définies dans les campagnes antérieures MUC, ACE et ESTER, mais le traitement est plus poussé, afin d'extraire des informations plus

fines.
Cette nouvelle définition se distingue de celles des campagnes antérieures par l'adoption d'une annotation à deux niveaux. Le premier niveau d'annotation consiste à détecter les entités et à les classifier selon une typologie comprenant des catégories et des sous-catégories proches de celles de la campagne ACE. En plus de ce niveau d'annotation habituel, un deuxième niveau de traitement consiste à annoter les différents constituants composant les entités détectées, selon une autre typologie fondée sur le rôle de ces constituants dans l'entité. L'adoption de ce nouveau concept a donné naissance à la tâche de « détection, classification et décomposition d'entités nommées » qui offre une annotation plus riche que celle des campagnes MUC et ACE. Toutefois, les règles d'annotations et la structure des entités nommées en résultant sont plus complexes.

3.3.1 Définitions

3.3.1.1 Définition des entitées nommées structurées et compositionnelles

La définition des entitées nommées structurées et compositionnelles est inspirée de celle de la thèse de Maud Ehrmann [Ehrmann, 2008] : «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». Cette définition est donc plus générale que celle de MUC, elle ne se limite pas aux noms propres, mais elle couvre un large éventail d'expressions linguistiques. Ainsi, nous trouvons en plus des noms propres, les descripteurs comme les noms de métier ou de rôle et les gentilés. La définition permet aussi d'inclure les qualificatifs et les titres quand ils apparaissent liés à une entité nommée.
Exemples d'expressions constituant des entités nommées selon le guide d'annotation QUAERO [Rosset *et al.*, 2011] :

- *président de la République*
- *parents d'élèves*
- *Italiens*
- *Mme Chirac*
- *chef d'entreprise*
- *l'État français*

Les expressions temporelles et les quantités sont des éléments riches sémantiquement. Elles sont également nécessaires pour de nombreuses applications. Le traitement de ces expressions a été séparé de la tâche de REN durant ACE. Des tâches spécifiques comme TIDES (Translingual Information Detection, Extraction and Summarization) [Ferro *et al.*, 2001] et TimeML [Pustejovsky *et al.*, 2003] ont été proposées pour traiter ce type d'expression.

Les concepteurs des entités nommées structurées et compositionnelles définies dans QUAERO, quant à eux, incluent ce type d'expression dans la tâche de reconnaissance d'entités nommées. Les règles d'annotations du guide QUAERO prévoient l'annotation des expressions temporelles absolues et relatives (relatives à une date faisant référence).

- *lundi 19 mars* (absolue)
- *l'année 1945* (absolue)
- *hier matin* (relative)
- *il y a quelques années* (relative)

La définition donnée des quantités dans le guide d'annotation est la suivante : « Une quantité est composée d'une valeur assortie éventuellement d'une unité de mesure, suivie éventuellement d'un objet » [Rosset *et al.*, 2011]. Les durées sont considérées comme étant des quantités temporelles et font ainsi partie de la classe Quantité. Ci-après, quelques exemples d'entités faisant partie de la classe Quantités :

- *trois pompiers*
- *une dizaine de kilomètres*
- *entre deux et trois kilomètres*
- *pendant trois ans* (durée)
- *une semaine entière* (durée)

3.3.2 Typologies

La tâche de détection, classification et décomposition ainsi définie comprend deux sous-tâches d'annotation, la classification et la décomposition. Ces deux sous-tâches complémentaires apportent chacune des informations de nature différente. La classification fournit la catégorie des entités détectées, alors que la décomposition vise à décrire les constituants des entités. Ainsi, deux types d'étiquetage différents ont été conçus pour chacune des sous-tâches impliquant des typologies différentes. Les constituants d'une entité peuvent être d'autres entités plus petites. Pour la suite de cette thèse nous allons adopter la terminologie suivante :

- un « constituant » désigne toute partie d'une entité nommée pouvant être à son tour une entité nommée ou un composant ;
- un « composant » désigne toute partie d'une entité nommée qui n'est pas une entité nommée.

3.3.2.1 La typologie des entités nommées : les types d'entités

La typologie adoptée pour la classification des entités nommées comporte sept catégories principales et trente-deux sous-catégories. Parmi les sept catégories principales nous trouvons les trois classes habituelles, Personnes, Organisation et Lieux, plus les classes Fonction, Produit, Quantité et Expression temporelle. Les étiquettes qui désignent les catégories des entités sont appelées des étiquettes types. Elles sont composées du nom de la catégorie principale (ou des premières lettres du nom de la catégorie) concaténé au nom de la sous-catégorie. Les étiquettes types sont ainsi utilisées pour classifier les entités nommées. Elles constituent le premier niveau d'étiquetage d'une entité. La hiérarchie des catégories et des sous-catégories est présentée dans le tableau 3.1. Pour des informations détaillées sur la définition de chaque catégorie et/ou sous-catégorie, nous renvoyons le lecteur vers le guide d'annotation [Rosset *et al.*, 2011].

Types	**Sous-types**
Personne	pers.ind, pers.coll
Fonction (professions)	func.ind, func.coll
Organisation	org.adm, org.ent
Lieu	loc.adm.town, loc.adm.reg, loc.adm.nat, loc.adm.sup, loc.add.phy, loc.add.elec, loc.oro, loc.fac, loc.phys.geo, loc.phys.hydro, loc.phys.astro
Produit	prod.object, prod.art, prod.media, prod.fin, prod.soft, prod.award, prod.serv, prod.doctr, prod.rule
Expression temporelle	time.date.abs, time.date.rel, time.hour.abs, time.hour.rel
Quantité	(pas de sous-catégories)

TABLEAU 3.1 – Hiérarchie des catégories d'entités nommées établie dans le projet QUAERO

Le tableau 3.2 présente des exemples d'entités nommées appartenant aux différentes classes d'entités. Les exemples sont tirés du guide d'annotation.

3.3.2.2 La décomposition et la typologie des composants

La définition d'une entité adoptée dans le projet QUAERO ne se limite pas seulement aux noms propres. Les gentilés, les titres, les descripteurs et les unités de mesure de temps ou de quantité font également partie de ce qu'il faut annoter. La sous-tâche de décomposition vise à annoter les différents constituants d'une entité, car ceci permet surtout d'apprendre des informations plus fines sur les entités détectées et ainsi de mieux les caractériser. Ce niveau d'annotation peut

Classes	Exemples
Personne	*socialiste Bertrand Delanoë, Beatles, Astérix, diaspora argentine*
Fonction	*ministre de l'Intérieur, chercheur CNRS, pompiers de Paris, miss Italie*
Organisation	*police française, société Peugeot, syndicat FSU*
Lieu	*ville de Paris, région Atlas, canal Saint-Martin, Lune, autoroute A6, 9 place de Rungis*
Produit	*Ak 47, Renault Espace, Le malade imaginaire, Firefox 3.5, Palme d'or*
Expression temporelle	*lundi 19 mars, l'année 1945, hier matin, il y a quelques années*
Quantité	*trois pompiers, une dizaine de kilomètres, entre deux et trois kilomètres, pendant trois ans, une semaine entière*

TABLEAU 3.2 – Exemples d'expressions linguistiques qui constituent des entités nommées, tirés du guide d'annotation QUAERO

être particulièrement utile pour certaines applications qui requièrent une analyse plus poussée de ce type de données. La typologie adoptée pour la sous-tâche de décomposition est donc différente de celle utilisée pour la classification. Certains composants sont communs à plusieurs catégories d'entités, alors que d'autres sont spécifiques à des catégories particulières. Ainsi, deux groupes d'étiquettes composants ont été distingués, les étiquettes composants transversales (communes à plusieurs catégories d'entités) et les étiquettes composants spécifiques (ne pouvant être utilisées que dans l'annotation des composants appartenant à des classes particulières). La typologie des constituants est fondée sur le rôle qu'ils jouent dans l'entité nommée. La décomposition vise à répondre à des questions comme :

- Est-ce le nom de l'entité ?
- Est-ce un descripteur ou un titre ?
- Est-ce une unité de mesure ?
- Est-ce une valeur ?

Le tableau 3.3 présente la liste des étiquettes spécifiques pour chaque classe d'entités.

Voici des exemples d'entités nommées contenant des composants spécifiques :

- **<pers.ind>**
  <name.first> Laurent </name.first> <name.last> Blanc </name.last>
  </pers.ind>

Types	Les composants spécifiques
Personne	name.last, name.first
Lieu	address-number, po-box, zip-code, other-adress-component
Produit	award-cat
Expression temporelle	day, week, month, year, century, millenium, reference-era, time-modifier
Quantité	object

TABLEAU 3.3 – Les composants spécifiques

- **<prod.award>** <name> Prix Nobel </name> de <award-cat> chimie </award-cat>
  </prod.award>
- **<time.date.abs>**
  <day> 25 </day> <month> janvier </month> <year> 2013 </year>
  </time.date.abs>

Les informations fournies par les composants spécifiques permettent de désambiguïser les entités, et offrent une description précise de leur contenu.
Les composants transversaux peuvent être utilisées pour annoter des composants à l'intérieur de n'importe quel type d'entité. Ils sont au nombre de dix : *name, name.nickname, kind, extractor, demonym, demonym.nickname, qualifier, val, unit, range-mark*. Voici des exemples de classification et de décomposition d'entités avec des composants transversaux :

- **<org.ent>**
  <kind> hôpital </kind> <name> Pitié-Salpêtrière </name>
  </org.ent>

- **<org.ent>**
  <kind> syndicat </kind> <name> FSU </name>
  </org.ent>

- **<func.ind>**
  <kind> ministre </kind> <demonym> américain </demonym>
  </func.ind>

- **<prod.rule>**
  <qualifier> dernier </qualifier> <kind> plan de paix </kind> <qualifier> international </qualifier>

<prod.rule>

- **<amount>**
  <val> trois </val> <unit> kilomètres </unit>
  </amount>

Les composants transversaux donnent aux entités structurées et compositionnelles leur caractéristique générique. Elles offrent la possibilité d'affiner la classification des entités facilement et sans avoir besoin de redéfinir à chaque fois une nouvelle typologie. Il est possible d'affiner la classification (ou d'effectuer une recherche spécifique) des Organisations, des Fonctions ou des Produits, en se fondant sur les composants portant l'étiquettes *kind* par exemple.

3.3.3 Les annotations et la gestion des exceptions

L'annotation utilise un format similaire à XML. Une balise formée par deux chevrons contenant le nom de l'étiquette **<labelA>** indique la frontière de début de l'entité ou du constituant et une deuxième balise contenant la même étiquette précédée par un slash **</labelA>** indique la frontière de fin d'entité. Le chevauchement est interdit, ainsi les balises fermantes doivent toujours apparaître dans l'ordre inverse des balises ouvrantes respectives. L'étiquetage en composants est interdit à l'extérieur des entités, ainsi les balises contenant des étiquettes composants doivent être utilisées seulement entre des balises ouvrantes et fermantes contenant une étiquette type qui délimite une entité nommée.
Exemple d'annotation :
La<pers.coll><kind>diaspora</kind><demonym>argentine</demonym>
</pers.coll>s'est réunie à<loc.adm.town><name>Paris</name></loc.adm.town>.

Des règles d'annotations ont été mises en place pour gérer les exceptions et les cas d'ambiguïtés. Ces règles d'annotation ont un impact important sur la structure de l'annotation résultante et, par la suite, sur les procédures d'évaluation qui doivent prendre en compte toutes les particularités de la tâche. Dans ce qui suit, nous allons nous intéresser à quelques exemples pour comprendre les difficultés que peuvent engendrer certaines règles d'annotations lors de l'évaluation de cette tâche.

3.3.3.1 Gestion des métonymies

Nous avons vu dans la section 3.2.2.1 que la métonymie est une des causes d'ambiguïté pour la reconnaissance d'entités nommées. La solution qui a été proposée durant ACE était la mise en place de la classe GSP. Cette solution a été adoptée ensuite dans d'autres projets. Toutefois, l'utilisation de la classe GSP pose un certain nombre de problèmes. Tout d'abord, elle ne traite que des métonymies

qui touchent des lieux (typiquement les États et les villes) alors que les autres types d'entités sont également l'objet de métonymies. Nous pouvons ainsi trouver des cas de métonymies entre un Produit et son Lieu de fabrication, entre une Personne et un Produit ayant le même nom ou entre une Organisation et un Lieu (bâtiment)... De plus, la classe GSP a tendance à aspirer un grand nombre d'entités, puisque cette classe a été originalement créée pour faciliter le travail des annotateurs et réduire les hésitations. Elle est très souvent utilisée en cas de doute. Les règles d'annotations mises en place dans QUAERO prévoient l'utilisation d'une double annotation plutôt que de classes fourre-tout comme la classe GSP. La double annotation permet de traiter tous les cas de métonymies et ne se limite pas seulement aux entités géopolitiques.

Ainsi, les directives du guide d'annotation QUAERO autorisent l'affectation de deux étiquettes types à une seule entité dans le cas d'une ambiguïté causée par une métonymie. Voici quelques exemples d'annotation de cas de métonymie :

- La <org.adm><loc.adm.nat> <name> France </name> </loc.adm.nat> </org.adm> a signé la convention.
- <org.ent><prod.media> <name> France-Inter </name> </prod.media> </org.ent>.
- La <org.ent> <pers.coll> <name> mafia </name> </pers.coll> </org.ent> a fait plusieurs victimes.

3.3.3.2 Gestion des coordinations

Les règles d'annotations QUAERO exigent que les entités liées par une coordination soient considérées comme des entités indépendantes. Elles doivent donc être annotées séparément. Voici quelques exemple :
<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> et <loc.phys.geo> de la <name> Bourgogne </name> </loc.phys.geo>
La coordination des titres fait exception, ainsi la séquence *« M. et Mme Chirac »* est considérée comme une seule entité de type *pers.coll*.

3.3.3.3 Gestion des imbrications

Lorsque il y a imbrication de deux entités, la plus petite entité joue le rôle de constituant pour l'entité qui l'englobe. Les règles d'annotations QUAERO autorisent l'utilisation d'étiquettes types pour annoter les constituants qui sont eux-mêmes des entités.
Ainsi, voici un exemple :
Le <func.ind>
<kind> maire </kind> de <loc.adm.town> <name> Paris </name> <loc.adm.town>
</func.ind>

Dans cet exemple un des constituants de l'entité de type *func.ind* (maire de Paris) est une entité de type *loc.adm.town* (Paris). Le mots Paris est tout d'abord annoté avec l'étiquette type correspondante, ensuite il est décomposé de la façon suivante. Son composant reçoit l'étiquette *name*. Il faut noter ici que le composant *name* ne dépend pas de l'entité de type *func.ind* puisque c'est un nom de Lieu et non pas un nom de Fonction. Dans les cas d'imbrications, les constituants de chaque entité nommée sont ceux qui se trouvent à un seul niveau en dessous de l'étiquette type correspondante. Ainsi dans notre exemple :

- les constituants de « func.ind » sont « kind » et « loc.adm.town » ;
- le constituant de « loc.adm.town » est « name ».

3.3.4 La structure des entités structurées et compositionnelles

Les entités structurées et compositionnelles se distinguent par l'utilisation de deux niveaux d'annotations. Un premier niveau utilise les étiquettes types pour classifier les entités, et un second niveau utilise les étiquettes composants pour décomposer les entités. Ceci donne aux entités structurées et compositionnelles une structure qui peut être assimilée à un arbre dans lequel l'étiquette type joue le rôle du noeud principal et les étiquettes composants jouent le rôle des feuilles.

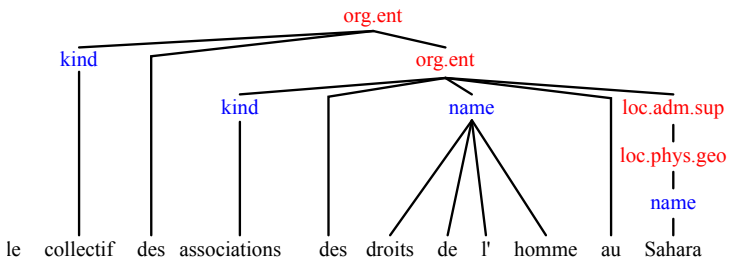

FIGURE 3.1 – Exemple d'annotation hiérarchique contenant des cas de métonymies et d'imbrications, avec étiquettes composants en bleu et les étiquettes types en rouge.

Les règles de gestion des exceptions (imbrications et métonymies) ont contribué à rendre la structure de ces entités encore plus complexe. Ainsi, nous pouvons trouver une structure avec plusieurs sous-arbres dans le cas d'imbrications. La figure 3.1 présente un exemple d'entité contenant un cas d'imbrications et de métonymies conduisant à une telle structure.

	Train-ETAPE	Test-ETAPE
Nombre de mots	335 387	115 803
Nombre d'entités	190 270	5 933
Nombre de composants	35 988	8554

TABLEAU 3.4 – Description du corpus ETAPE en termes de nombre de mots, d'entités et de composants

3.4 Les données

Le guide d'annotation QUAERO a servi comme modèle pour l'annotation de deux grand corpus en langue française, les corpus ETAPE [Gravier *et al.*, 2012] et QUAERO [Galliano *et al.*, 2009a]. Nous décrivons ci-dessous ces deux corpus.

3.4.1 Le corpus ETAPE

Les données ETAPE sont constituées de 41 heures et 20 minutes d'enregistrement d'émissions radiophoniques et télévisées, sélectionnées de manière à contenir principalement de la parole non planifiée et une portion raisonnable de parole superposée (1,5 heures). Le jeu d'entraînement ETAPE comporte 33 heures, alors que le jeu de test comporte 8 heures et 20 minutes [Gravier *et al.*, 2012]. Toutes les données ont été transcrites manuellement et annotées en entités nommées par des experts. Les annotations ont été réalisées conformément au guide d'annotation développé durant le projet QUAERO [Rosset *et al.*, 2011]. Le tableau 3.4 contient une description des corpus en termes de nombre de mots et d'entités nommées.

3.4.2 Le corpus QUAERO

Le corpus QUAERO est constitué de 100 heures de parole type broadcast news provenant de différents types d'émissions radiophoniques et télévisuelles portant sur divers sujets. Tout le corpus a été transcrit manuellement et annoté en entités nommées conformément au guide d'annotation QUAERO. Le tableau 3.5 affiche une description du corpus obtenu.

3.5 Conclusion

Dans le cadre du projet QUAERO une nouvelle tâche de détection, classification et décomposition d'entités nommées a été mise en place. Les règles d'annotations correspondantes exigent une annotation à deux niveaux, un premier niveau pour

	Train-QUAERO	Test-QUAERO
Nombre de mots	1 251 586	97 871
Nombre d'entités	113 885	5 523
Nombre de composants	146 405	8 902

TABLEAU 3.5 – Description du corpus QUAERO en termes de nombre de mots, d'entités et de composants

la classification et un second pour la décomposition. Ceci offre aux entités structurées et compositionnelles l'avantage d'être à la fois précises et génériques. Toutefois, la structure d'annotation qui en résulte est une structure complexe qui peut être assimilée à un arbre dans lequel l'étiquette type joue le rôle du noeud principal et les étiquettes composants jouent le rôle des feuilles. Avec l'introduction de la sous-tâche de décomposition et devant la nouvelle structure arborescente des annotations QUAERO, se pose la question de la compatibilité des métriques d'évaluation existantes avec cette nouvelle tâche ainsi définie. Dans le chapitre suivant nous discutons de l'impact de la structure d'annotation arborescente mise en place dans le cadre du projet QUAERO sur les métriques classiquement utilisées pour évaluer la tâche de REN. Nous proposons une nouvelle métrique qui permet de prendre en compte toutes les spécificités de la tâche.

Chapitre 4

Nouvelle métrique pour l'évaluation des entités structurées et compositionnelles

4.1 Introduction

Nous avons décrit dans le chapitre précédent la tâche de détection, classification et décomposition d'entités nommées mise en place dans le cadre du projet QUAERO, ainsi que la structure d'annotation arborescente résultante. Cette tâche inclut deux niveaux d'annotations. Le premier niveau consiste à détecter et classifier les entités et le second niveau à décomposer les entités détectées. Les concepteurs de la tâche indiquent que les méthodes d'évaluation classiquement utilisées pour évaluer la qualité de la REN (reconnaissance d'entité nommée) ne prennent pas en compte toutes les caractéristiques de cette nouvelle tâche [Grouin *et al.*, 2011].

Dans ce chapitre, nous décrivons les différentes étapes qui constituent le processus d'évaluation de la tâche de REN. Ensuite, nous discutons les points faibles et les points forts des métriques d'évaluation classiquement utilisées pour l'évaluation de la tâche de REN et nous argumentons sur le fait qu'elles ne sont pas adaptées pour l'évaluation de tâches complexes. Nous présentons par la suite la nouvelle métrique ETER (*Entity Tree Error Rate*) que nous proposons pour l'évaluation de la tâche de reconnaissance, classification et décomposition d'entités. Cette métrique prend en compte les caractéristiques de la tâche et la structure d'annotation complexe résultante et permet de prendre en considération le contexte applicatif lors de l'évaluation.

4.2 Processus d'évaluation de la tâche de reconnaissance d'entités nommées

L'approche d'évaluation la plus utilisée en extraction d'entités nommées est l'approche appelée par « boîte noire » (*black box*), qui consiste à évaluer les systèmes de REN en se fondant uniquement sur leurs sorties sans s'intéresser en détail à leur architecture interne. L'utilisation de cette approche requiert le développement d'une plateforme d'évaluation qui permet :

- de vérifier que les solutions fournies répondent à la tâche fixée *a priori* ;
- de s'assurer que les comportements des systèmes sont conformes aux descriptions de la tâche fournie dans le guide d'annotation ;
- de comparer l'efficacité des solutions et des approches proposées sur une base commune.

La mise en place d'une telle plateforme comporte deux étapes importantes. La première consiste à rassembler des données de test en adéquation avec la tâche et typiques de celles que les systèmes auront à traiter dans des cadres applicatifs réels. La seconde requiert la sélection et l'implémentation de la métrique d'évaluation adéquate permettant de refléter au mieux les performances des systèmes.

4.2.1 Constitution des données de test

Le choix des données de test est une étape importante dans l'élaboration d'une procédure d'évaluation. Des débats méthodologiques sur la nature et la taille des données de test existent au sein de la communauté du TAL entre les adeptes des approches symboliques et ceux des approches statistiques. Certains chercheurs dont Chomsky [Chomsky, 2002] (pages 124-128) argumentent pour l'utilisation de jeux de test bien sélectionnés dans le but de se focaliser sur des phénomènes linguistiques rares mais informatifs plutôt que d'utiliser des grandes quantités de données contenant des redondances. Cette méthodologie permet de mieux étudier les phénomènes linguistiques pour mieux les traiter. Néanmoins, elle suscite des questionnements sur sa capacité à refléter les besoins des utilisateurs [Cunningham *et al.*, 1995]. Jones et Galliers [Jones et Galliers, 1996] pensent que l'utilisation de jeu de test doit être traité avec précaution car la langue naturelle est riche d'expressions « multifacettes » (incluant plusieurs phénomènes linguistiques à la fois), et séparer les phénomènes pour les isoler est un processus difficile à réaliser qui peut conduire à des exemples artificiels dont l'utilité est contestée [Jones et Galliers, 1996].

L'approche la plus utilisée pour la constitution de données de test dans les campagnes d'évaluation est l'approche à base de corpus. Elle consiste à collecter des données dans des conditions similaires aux conditions d'utilisation réelles. La base de données collectée doit être assez large pour permettre de dégager des perfor-

mances statistiquement valides. Cette approche vise à confronter les systèmes à la réalité des données pour évaluer la maturité de la technologie et déterminer les cas d'usage possibles.

L'évaluation de la tâche de REN dans les campagnes d'évaluation a toujours été fondée sur l'approche à base de corpus de test depuis sa première apparition dans MUC. Comme beaucoup de tâches en TAL, l'évaluation en REN consiste à comparer les performances des systèmes à celles des annotateurs humains. Le corpus test est ainsi annoté par des annotateurs experts selon les règles définies par les concepteurs de la tâche dans les guides d'annotation. Les annotations résultantes serviront de référence. Elles représentent une « vérité de terrain » (*ground truth*) à laquelle seront comparées les annotations hypothèses fournies par les systèmes de REN. Afin de limiter le risque d'erreur les annotations de référence sont annotées par plusieurs annotateurs. La version finale est obtenue en fusionnant les données multi-annotateurs pour obtenir un « *gold standard* ». La fusion permet de détecter et d'éliminer les erreurs provenant de la perte de concentration ou de la fatigue. Les scores des accords inter-annotateurs permettent d'évaluer la validité des données, et constituent une estimation de confiance accordée à la référence. Toutefois, il est à noter que les annotations de référence ne sont pas parfaites, ce qui n'est pas uniquement dû au facteur humain, mais peut également provenir de manque de règles d'annotations spécifiques à certains cas (guide d'annotation incomplet) ou à des cas d'ambiguïtés intrinsèques à la langue. Pour plus d'informations sur les pratiques à respecter lors de la création de données annotées manuellement, nous renvoyons le lecteur à la thèse de Karen Fort [Fort *et al.*, 2009] qui porte sur ce sujet.

4.2.2 Mesures de performance

L'évaluation fondée sur des corpus de référence utilise généralement des métriques qui permettent de mesurer la distance entre un ensemble de réponses correctes (la référence) et les hypothèses des systèmes automatiques. Cette pratique vise à déterminer si un système de traitement automatique des langues offre un comportement attendu, en observant uniquement ses sorties, et ce, dans le respect du principe de l'évaluation en boîte noire. La mesure de la distance qui sépare les réponses des systèmes des réponses attendues dépend de la nature de la tâche. Pour des tâches telles que la traduction automatique ou le dialogue homme-machine, la réponse attendue n'est pas forcément unique et, en pratique, il est impossible d'énumérer toutes les réponses possibles. C'est pourquoi, nous avons souvent recours à des juges humains qui utilisent leurs compétences linguistiques pour évaluer à quel point les hypothèses des systèmes appartiennent ou non à l'ensemble des réponses satisfaisantes. Pour des tâches telles que la REN, sachant que la réponse correcte (la réponse donnée par le *gold standard*) est unique, ou

avec une faible variation, la difficulté de l'évaluation va résider dans la mesure de la distance entre la référence et l'hypothèse, ce qui nécessite de choisir la métrique la plus adaptée pour la tâche à évaluer.
L'automatisation du processus d'évaluation est aussi une étape très importante car :

- elle permet de réduire le temps et le coût de l'évaluation ;
- elle garantit des conditions d'évaluation équitables à tous les participants ;
- elle facilite la reproduction de l'évaluation ce qui permet de mesurer le progrès de la technologie dans des conditions similaires à l'évaluation initiale ;
- elle permet aux développeurs d'optimiser leurs systèmes en faisant des mesures répétées pour choisir les paramètres qui optimisent les performances de leurs applications.

L'automatisation de l'évaluation de la tâche de REN implique une première étape qui est la phase de l'alignement. Cette phase consiste à projeter les annotations d'hypothèse sur les annotations de référence afin de déterminer l'ensemble de paires de segments (hypothèse et référence) à comparer. La seconde étape consiste à appliquer la métrique d'évaluation qui, en se fondant sur le résultat de l'alignement, permet de mesurer la distance entre l'hypothèse et la référence.
Le défi des évaluateurs consiste à définir ou à sélectionner l'algorithme d'alignement et la métrique d'évaluation qui peuvent refléter véritablement les performances des systèmes étant donné la tâche demandée.

4.3 Les mesures d'évaluation des systèmes de REN : points forts et points faibles

La précision (P) et le rappel (R) sont les mesures les plus utilisées en évaluation des systèmes d'extraction d'informations. Définis à l'origine pour l'évaluation de la recherche documentaire [Salton et Buckley, 1988], ils sont applicables à toute tâche visant à identifier des éléments pertinents parmi un ensemble d'éléments candidats :

$$P = \frac{C}{C + S + I} \tag{4.1}$$

$$R = \frac{C}{C + S + D} \tag{4.2}$$

Avec :

- C : le nombre de réponses correctes ;
- S : le nombre de substitutions ;
- D : le nombre d'omissions ;
- I : le nombre d'insertions (fausses alarmes).

4.4.3 Les mesures d'évaluation des systèmes de REN : points forts et points faibles

La précision est donnée par le ratio entre les réponses correctes et toutes les réponses données par un système. Il permet d'estimer la fiabilité des hypothèses fournies par un système donné. Alors que le rappel est donné par le ratio entre les réponses correctes et toutes les réponses attendues. Il permet d'estimer la capacité d'un système à couvrir l'ensemble des réponses se trouvant dans le test. Aucune des deux métriques ne peut être considérée comme une métrique complète pour mesurer la distance entre l'hypothèse et la référence, puisque la formule de la précision ne prend pas en compte les erreurs de suppression, et que la formule du rappel ne prend pas en compte les erreurs d'insertion. C'est leur moyenne harmonique, la F-mesure, qui est utilisée comme mesure unique afin de comparer les performances des systèmes entre eux. La formule générale de F-mesure est F_β définie comme suit :

$$F\text{-}mesure = (1 + \beta^2)\frac{P \times R}{(\beta^2 P) + R} \tag{4.3}$$

Classiquement c'est la F_1 ($F_\beta \; avec \; \beta = 1$) qui est utilisée. La F_1 permet d'accorder un poids égal à la précision et au rappel.

La F-mesure a été utilisée pour classer les participants durant de nombreuses campagnes d'évaluation. Toutefois, cette mesure présente des limites. D'abord, il a été démontré dans [Makhoul *et al.*, 1999] que la fusion de P et de R avec une moyenne harmonique diminue l'importance des erreurs d'omission et d'insertion par rapport aux substitutions. D'autre part, l'évolution de la complexité de la tâche de REN rend l'utilisation de la F-mesure inadaptée. En effet, les entités nommées utilisées aujourd'hui dans la plupart des campagnes d'évaluation ont une structure hiérarchique avec des types et des sous-types. Ceci a conduit à l'apparition de nombreux types d'erreurs de substitution : substitution de type, substitution de sous-type et substitution de frontières auxquels on voudra attribuer des poids différents. Voilà des exemples d'erreurs de substitution :

- REF : La <loc.fac> gare de Rungis </loc.fac>
- HYP1 : La <org.adm> gare de Rungis </org.adm> : substitution de type
- HYP2 : La <loc.fac> gare </loc.fac> de Rungis : substitution de frontière
- HYP3 : La <loc.adm.town> gare de Rungis </loc.adm.town> : substitution de sous-type
- HYP4 : La gare de <pers.ind> Rungis </pers.ind> : substitution de frontière + substitution de type

La précision, le rappel et, par conséquence, la F-mesure possèdent un fonctionnement binaire où chaque réponse ne peut être considérée que comme correcte ou fausse. Ainsi, il n'est pas possible d'affecter des poids différents aux erreurs selon leur gravité.

Inspiré par le taux d'erreur de mots utilisé classiquement pour évaluer les systèmes de reconnaissance automatique de la parole (RAP), des métriques fondées

sur le taux d'erreur ont été introduites comme alternatives à la F-mesure pour évaluer la reconnaissance d'entités nommées. Les métriques fondées sur le taux d'erreur visent à estimer le coût que représentent les erreurs du système de REN pour des utilisateurs ou des applications. L'amélioration des performances est alors proportionnelle à la réduction du taux d'erreur.

La première métrique fondée sur le taux d'erreur, ERR, a été introduite durant MUC-6 [Makhoul *et al.*, 1999]. Elle est définie comme suit :

$$ERR = \frac{S + D + I}{C + S + D + I}$$

Le fait de compter les insertions (I) dans le dénominateur permet d'avoir des taux d'erreur compris entre 0 et 100 %. Mais, comme le nombre d'insertions varie d'un système de REN à un autre, le dénominateur de ERR n'est pas constant. Par conséquent ERR ne peut pas être utilisée pour comparer les performances obtenues par des systèmes différents.

Le *Slot Error Rate* (SER) a été proposé par [Makhoul *et al.*, 1999] afin de remédier à cet inconvénient. Le SER est défini comme suit :

$$SER = \frac{S + D + I}{C + D + S} = \frac{S + D + I}{R}$$

avec R le nombre d'entités se trouvant dans la référence qui est une constante du corpus test. Cette définition a permis de résoudre les inconvénients de ERR. Le SER permet d'affecter des poids différents aux erreurs selon leur gravité comme le montre l'équation 4.4. Elle a été utilisée pour l'évaluation de la tâche de REN dans les premières campagnes d'évaluation de ACE et dans ESTER, ETAPE et QUAERO.

$$SER = \frac{\alpha_1 S_t + \alpha_2 S_f + \beta D + \gamma I}{R} \tag{4.4}$$

avec :

- S_t et S_f le nombre d'erreur de substitution de type et de frontières ;
- D et I le nombre d'erreur d'omission et d'insertion d'entité ;
- α_1, α_2, β et γ les poids affectés à chaque type d'erreur.

Le SER est fondé sur la mesure de distance entre les slots de l'hypothèse et les slots de la référence. Un slot est défini comme étant un segment de texte annoté, caractérisé par des frontières de début et de fin et par une étiquette. Ce principe suppose que tous les slots possèdent le même poids, c'est-à-dire ont la même importance pour la tâche à évaluer. Par conséquent, le SER ne peut être utilisé que pour l'évaluation de tâches simples dans lesquelles tous les slots jouent le même rôle.

L'apparition de sous-tâches, telles que la reconnaissance des occurrences et des relations dans les campagnes, a augmenté la complexité de la tâche de REN.

Ceci a rendu l'utilisation du SER inadaptée pour l'évaluation de la tâche. C'est ainsi qu'une nouvelle métrique EDT_{value} (*Entity Detection and Tracking value*) [Doddington *et al.*, 2004] a vu le jour durant ACE 2004 [NIST, 2004] dans le but de prendre en compte la reconnaissance d'occurrences lors de l'évaluation. Cette métrique consiste à calculer un score pour chaque entité et à fournir la somme de tous les scores comme résultat final.

$$EDT\_value_{sys} = \sum_i \text{value\_of\_sys\_entity}_i \tag{4.5}$$

La EDT_{value} a évolué pour donner le $LEDR_{value}$ (*Local Entity Detection and Recognition value*) durant ACE 2008 [NIST, 2008], qui consiste tout simplement à normaliser le score EDT_{value} par la somme des scores calculés à partir de la référence.

$$LEDR\_value_{sys} = \frac{\sum_i \text{value\_of\_sys\_token}_i}{\sum_i \text{value\_of\_ref\_token}_i} \tag{4.6}$$

EDT et LEDR ont permis de contourner le problème de la complexité de la tâche. Les scores offerts par ces métriques permettent de comparer et de classer les systèmes dans une campagne d'évaluation. Toutefois, l'interprétation des scores reste très difficile et à prendre avec beaucoup de précautions, puisque il ne s'agit ni d'un taux d'erreur ni d'une mesure de performance proprement dite.

4.4 Les métriques actuelles et l'évaluation des entités nommées structurées et compositionnelles

Le SER a été utilisé pour l'évaluation des systèmes de REN durant les campagnes d'évaluation QUAERO [Galibert *et al.*, 2010] et ETAPE [Galibert *et al.*, 2014]. Les organisateurs ont signalé que le SER ne permet pas de prendre en compte toutes les caractéristiques de la tâche de détection, classification et décomposition d'entités nommées [Grouin *et al.*, 2011]. En effet, l'utilisation du SER présuppose que tous les slots sont homogènes (ont la même importance). Ceci n'est pas vérifié dans le cas des entités nommées étendues et compositionnelles, puisque nous pouvons distinguer deux types de slots.

- les slots types nécessaires pour la classification des entités ;
- les slots composants nécessaires pour la décomposition des entités.

Par conséquent, chaque entité nommée n'est plus identifiable par un seul slot mais par un ensemble de slots : le slot type et l'ensemble des slots constituants qui composent l'entité. La structure des entités peut ainsi être assimilée à un arbre dans lequel le slot type est le noeud principal et les slots constituants sont les feuilles.

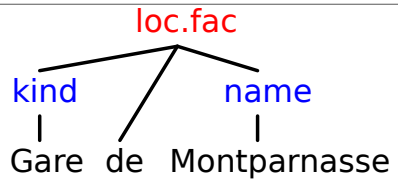

FIGURE 4.1 – Exemple d'une entité simple

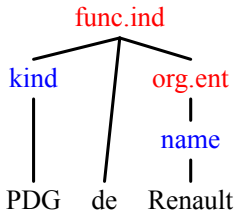

FIGURE 4.2 – Exemple d'entités imbriquées

L'interprétation d'une entité se fait obligatoirement en tenant compte de la hiérarchie qui relie les slots types et les slots composants. Sémantiquement, le type apporte une information sur la classe de l'entité alors que les composants fournissent une description détaillée de sa composition. Par exemple, dans le cas de *gare de Montparnasse*, figure 4.1, les deux composants *name* et *kind* ne peuvent pas être interprétés de façon isolée. Leur interprétation dépend du noeud type auquel ils sont reliés. Sachant que le noeud type est un *loc.fac*, nous pouvons déduire qu'il s'agit d'un nom (*name*) et d'un genre (*kind*) d'un bâtiment.

Non seulement, le SER ne permet pas de prendre en compte l'existence de deux catégories de slots différents (les types et les composants), mais aussi, il ne tient pas compte de la relation de dépendance qui relie les deux catégories de slots.

Pour illustrer cela, nous allons nous intéresser au cas d'imbrication. Nous parlons d'imbrication lorsqu'une large entité englobe une autre plus petite. Ainsi, un des constituants de l'entité la plus large est une autre entité. La figure 4.2 illustre un exemple d'entités imbriquées. Dans cet exemple le nom *Renault* est le nom de l'Organisation (*org.ent*) et non celui de la Profession (*func.ind*). Par conséquent, il est important de noter que l'interprétation des slots composants dépend nécessairement du noeud type auquel ils sont directement reliés. Cette relation de dépendance n'est pas prise en compte par le SER qui était à l'origine développé pour évaluer des entités à structure plate dans laquelle chaque slot était équivalent à une entité.

Un autre cas particulier qui découle des règles d'annotations des entités structurées et compositionnelles est le double étiquetage (utilisation de deux étiquettes

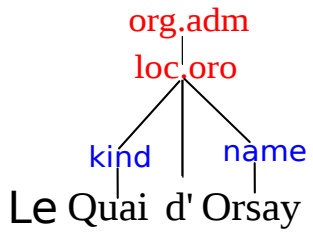

FIGURE 4.3 – Exemple de cas de métonymie

types) pour l'annotation des cas de métonymie. La première étiquette est intrinsèque (sens premier) et la seconde est liée au contexte. Un exemple de cas de métonymie est montré figure 4.3. Dans le cas de doubles étiquettes, l'application du SER revient à considérer qu'il s'agit de deux slots différents, donc de deux entités différentes, alors qu'il s'agit d'une unique entité ayant une étiquette particulière.
Toutes ces observations montrent que le SER n'est pas la métrique adaptée pour évaluer des tâches complexes telles que la détection, classification et décomposition d'entités nommées. Il est important de noter ici que précision, rappel et F-mesure sont également sensibles aux mêmes problèmes que nous venons de citer pour le SER.
Nous proposons dans la section suivante une nouvelle métrique alternative au SER, qui permet de prendre en compte toutes les caractéristiques de la tâche.

4.5 Entity Tree Error Rate : ETER

4.5.1 Objectifs

Nous venons de voir que les métriques (F-mesure et SER) utilisées actuellement ne sont pas adaptées pour l'évaluation des entités nommées structurées et compositionnelles car elles ne prennent pas en compte la structure et les caractéristiques de ces entités. Nous considérons qu'une meilleure métrique doit prendre en compte l'ensemble des caractéristiques suivantes :

- une entité ne peut pas être assimilée à un seul slot, mais à un ensemble de slots (types et composants) dépendants les uns des autres : structure arborescente ;
- l'organisation des catégories d'entités est hiérarchique avec des types et des sous-types ;
- des entités peuvent être imbriquées les unes dans les autres ;
- une double annotation en types est utilisée dans les cas de métonymies ;

- les entités nommées QUAERO comportent une annotation en types et une autre en composants, les informations apportées par chacune de ces annotations sont de natures différentes. Il serait donc intéressant de permettre aux utilisateurs de fixer les poids qu'ils souhaitent accorder à chacune de ces sous-tâches selon le cadre applicatif visé par l'évaluation.

La métrique que nous proposons peut être considérée comme une adaptation du SER à la tâche et à la structure des entités structurées et compositionnelles. En effet, comme nous l'avons montré dans la section précédente, l'utilisation du SER n'est pas pertinente dans le cas de la tâche de détection, classification et décomposition d'entités. Afin de combler ces limites nous proposons une nouvelle métrique ETER (*Entity Tree Error Rate*), qui calcule un taux d'erreur fondé sur les arbres (chaque arbre est constitué par l'ensemble des slots qui constituent l'entité) plutôt que sur les slots isolés.

Comme expliqué précédemment, la mesure de performance en REN consiste à estimer la distance entre les annotations de référence et celles de l'hypothèse. Ce processus nécessite deux étapes principales :

- l'alignement qui permet d'associer les éléments de la référence et de l'hypothèse à comparer ;
- la mesure de distance entre les éléments issus de la phase d'alignement.

4.5.2 Alignement

La première étape pour chaque métrique fondée sur une comparaison entre l'hypothèse et la référence consiste à aligner les éléments à comparer. Dans le cas du SER, l'alignement cherche à maximiser un score calculé sur des slots indépendants les uns des autres. Ceci donne lieu à des associations entre des slots appartenant à des entités (arbres) différentes comme illustré sur la figure 4.4. À première vue, cet alignement peut nous sembler correct, mais, en réalité, il casse le concept d'entité en associant des constituants appartenant à des entités différentes. Ainsi en considérant cet alignement, l'annotation de l'entité de type *func.ind* proposée dans l'hypothèse est parfaitement correcte, alors que ceci n'est pas vrai du fait que les deux composants *name* ne sont pas des noms de Fonction mais des noms de deux Lieux qui n'ont pas été détectés.

Afin, d'éviter ce type d'association, nous proposons d'effectuer un alignement fondé sur des arbres où chaque arbre est construit à partir de l'ensemble des slots appartenant à une même entité. Chaque arbre est identifiable grâce à son noeud père. Le noeud père est caractérisé par le slot type qui permet d'indiquer les frontières de début et de fin de l'entité ainsi que sa classe. À chaque noeud père nous associons la liste de tous les slots constituants qui se trouvent à un seul niveau en dessous du noeud père.

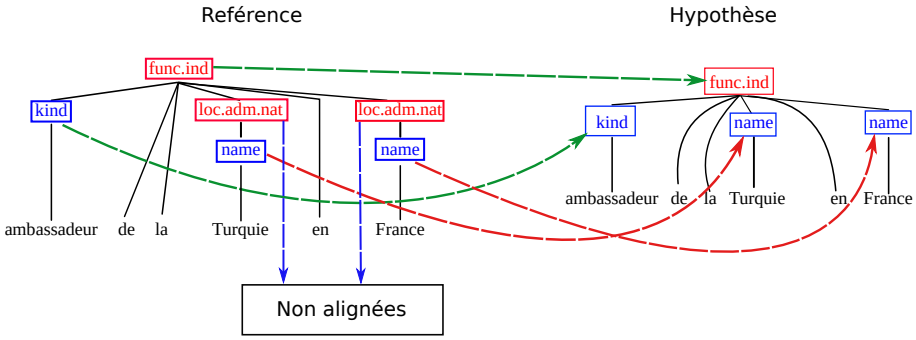

FIGURE 4.4 – Exemple d'alignement fondé sur les slots

Une fois que la construction des arbres à partir des annotations référence et hypothèse est terminée, nous appliquons un alignement à deux niveaux. Un premier niveau consiste à aligner les noeuds pères des arbres de la référence à ceux de l'hypothèse. L'algorithme vise à minimiser le taux d'erreur sur les slots types. Ensuite, le deuxième niveau consiste à aligner les constituants appartenant à chaque paire d'arbres issue de la première phase d'alignement en choisissant la solution qui permet de minimiser le taux d'erreur sur les constituants.

Un tel algorithme donne le résultat affiché figure 4.5, sur lequel, nous pouvons voir qu'il existe trois entités dans la référence et une seule dans l'hypothèse. Ainsi, l'arbre ayant comme noeud père *func.ind* est aligné avec l'arbre se trouvant dans l'hypothèse et ayant le même noeud père. Les deux autres arbres *loc.adm.nat* ne trouvent pas de correspondant côté hypothèse et sont considérés comme supprimés. Nous passons maintenant au résultat du deuxième niveau d'alignement pour la paire d'arbres *func.ind* où nous pouvons voir que dans ce cas de figure les slots *name* de l'hypothèse seront alignés avec les slots *loc.adm.nat* correspondant aux constituants de l'entité de la référence. Ceci permet de repérer l'erreur de décomposition commise par le système de REN. Cet alignement respecte mieux la relation de dépendance qui relie les slots, et ainsi, respecte mieux la définition de la tâche.

4.5.3 Mesure du taux d'erreur

La métrique proposée, ETER (*Entity Tree Error Rate*), utilise le résultat de la phase d'alignement, et permet de calculer la distance entre les paires d'entités (arbres) dans l'hypothèse et dans la référence. À la différence du SER qui offre un taux d'erreur fondé sur la comparaison de slots, ETER est aussi une mesure de taux d'erreur mais qui s'appuie sur la comparaison d'entités. La formule de ETER est donnée par l'équation 4.7 :

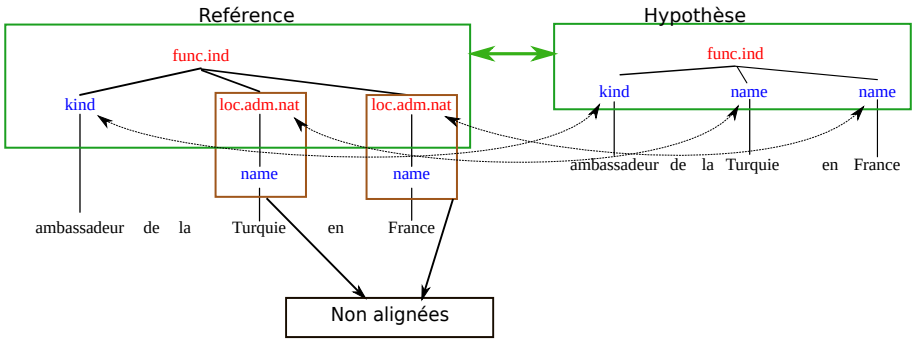

FIGURE 4.5 – Exemple d'un alignement à deux niveaux fondé sur les arbres construits à partir des entités

$$ETER = \frac{I + D + \sum_{(e_r, e_h)} E(e_r, e_h)}{N_E} \tag{4.7}$$

avec :

- I : le nombre d'insertion (fausses alarmes), les entités de l'hypothèse qui ne s'alignent avec aucune entité de la référence ;
- D : le nombre d'omission, les entités de la référence qui ne s'alignent avec aucune entité de l'hypothèse ;
- (e_r, e_h) : les paires d'entités, référence et hypothèse, associées à l'issue de l'alignement ;
- $E(r, h)$: l'erreur calculée pour chaque paire d'entités (e_r, e_h). L'erreur peut être égale à zéro ;
- N_E : le nombre d'entités dans la référence.

La somme des erreurs comporte trois composantes, les insertions, les omissions et les substitutions. Traditionnellement, les insertions et les omissions possèdent un poids de 1. La métrique repose essentiellement sur le calcul de l'erreur pour les paires d'entités $E(r, h)$. Cette erreur est répartie en deux membres, correspondant aux deux niveaux d'annotations. La première partie calcule l'erreur sur la détection et la classification de l'entité E_T. La seconde partie calcule l'erreur sur la décomposition E_C ;

$$E(r, h) = (1 - \alpha) E_T(e_r, e_h) + \alpha E_C(e_r, e_h), \alpha \in [0..1] \tag{4.8}$$

avec :

- $E_T(e_r, e_h)$: erreur de classification fondée sur la mesure de la distance entre les deux noeuds pères de la paire d'arbres (e_r, e_h) ;
- $E_c(e_r, e_h)$: erreur de décomposition fondée sur la mesure de la distance entre les différents constituants des deux arbres (e_r, e_h) ;
- α est un paramètre qui permet de fixer le poids de la décomposition par rapport à la classification.

Par défaut nous fixons $\alpha = 0,5$, nous considérons que cette valeur est la plus adaptée pour évaluer la tâche globale (détection, classification et décomposition d'entités nommées) car elle accorde un poids égal à la décomposition et à la classification. Toutefois, les utilisateurs peuvent modifier cette valeur et choisir celle qui s'adapte le mieux à leur cadre applicatif ou celle qui leur permet de faire des meilleures analyses des performances de leurs systèmes.
Le calcul de $E_T(e_r, e_h)$ nécessite une comparaison des labels types et sous-types du premier niveau d'annotation ainsi qu'une comparaison des frontières de début et de fin d'entité correspondant à la paire d'entités (e_r, e_h). L'estimation de l'erreur relative à chaque label type se fait en vérifiant la validité des informations suivantes :

- détection de la présence d'une entité lorsque il y en a une ;
- détection des frontières de début et de fin de l'entité ;
- affectation de la bonne classe (type) de l'entité ;
- affection de la bonne sous-classe (sous-type) de l'entité.

Puisque le taux d'erreur final est normalisé par rapport au nombre d'entités dans la référence (N_E), l'erreur totale la plus adaptée sur une entité est de 1. Ainsi, nous affectons un poids égal à 0,25 à chacune des quatre informations citées. En effet, qu'une mauvaise détection d'entité affectera l'ensemble des quatre informations et conduira ainsi à une erreur d'ommission ou d'insertion, ayant un poids de 1. L'affectation d'une classe incorrecte implique automatiquement une erreur pour la sous-classe et aura ainsi un poids total de 0,5.

Dans le cas de labels simples (pas de métonymie), nous procédons à une comparaison élémentaire telle que décrite dans l'équation 4.9 :

$$E_L(e_r, e_h) = E_{L1}(e_r, e_h) = \begin{cases} 0,5 & \text{Si le type est différent} \\ 0,25 & \text{Si le sous-type est difféent} \\ 0 & \text{Si les labels types sont entièrement identiques} \end{cases} \tag{4.9}$$

avec $E_L(e_r, e_h)$ est l'erreur élémentaire obtenue en comparant les labels types de la paire d'entités (e_r, e_h).
Pour une paire d'entité associées (l'entité détectée existe bien dans la référence) l'erreur maximale qui peut être attribuée est de 0,75, dans le cas où les frontières et le type sont tous les deux faux.

$$E_T(e_r, e_h) = E_L(e_r, e_h) + \begin{cases} 0,25 & \text{si les frontières ne sont pas correctes} \\ 0 & \text{si non} \end{cases} \tag{4.10}$$

Si aucune des deux entités, référence et hypothèse, ne présente un cas de métonymie nous utilisons directement une comparaison élémentaire. Si un cas de métonymie est observé sur seulement un des deux labels à comparer (référence ou hypothèse), nous calculons alors la moyenne entre une erreur de type (0,5) et la comparaison élémentaire qui donne le meilleur score parmi les deux choix

possibles. Si les deux entités (référence et hypothèse) présentent un cas de métonymie, nous calculons la moyenne des comparaisons élémentaires entre les deux paires de labels permettant la meilleure association possible (celle qui minimise l'erreur) entre les deux labels de la référence et de l'hypothèse.

$$E_L(\{e_{r1}\},\{e_{h1}\}) = E_{L1}(e_{r1},e_{h1}) \quad (4.11)$$

$$E_L(\{e_{r1},e_{r2}\},\{e_{h1}\}) = \min(\frac{0,5+E_{L1}(e_{r1},e_{h1})}{2}, \frac{0,5+E_{L1}(e_{r2},e_{h1})}{2}) \quad (4.12)$$

$$E_L(\{e_{r1}\},\{e_{h1},e_{h2}\}) = \min(\frac{0,5+E_{L1}(e_{r1},e_{h1})}{2}, \frac{0,5+E_{L1}(e_{r1},e_{h2})}{2}) \quad (4.13)$$

$$E_L(\{e_{r1},e_{r2}\},\{e_{h1},e_{h2}\}) = \min(\frac{E_{L1}(e_{r1},e_{h1})+E_{l1}(e_{r2},e_{h2})}{2}, \frac{E_{L1}(e_{r1},e_{h2})+E_{L1}(e_{r2},e_{h1})}{2}) \quad (4.14)$$

Le score ou l'erreur de décomposition $E_c(e_r,e_h)$ peut être assimilé à un SER local calculé pour l'ensemble des composants d'une seule entité. Il est fondé sur le résultat de l'alignement des constituants des paires d'entités (e_r,e_h).

$$E_C(e_r,e_h) = \frac{I_c(e_r,e_h)+D_c(e_r,e_h)+\sum_{(c_r,c_h)}E_{c1}(c_r,c_h)}{Nc(r)} \quad (4.15)$$

avec :

- $I_c(e_r,e_h)$: le nombre de constituants insérés ;
- $D_c(e_r,e_h)$: le nombre de constituants supprimés ;
- (c_r,c_h) : les paires de constituants, référence et hypothèse, associées à l'issue de l'alignement ;
- $E_{C1}(e_r,e_h)$: l'erreur calculée pour chaque paire de constituants associés. Cette erreur peut être égale à zéro ;
- $Nc(r)$: le nombre de constituants dans l'entité (arbre) référence.

Finalement, l'erreur $E_{C1}(e_r,e_h)$ entre les paires des composants est calculée ainsi :

$$E_{C1}(c_r,c_h) = \begin{cases} 0,5 & \text{Si les labels sont différents} \\ 0 & \text{Si non} \end{cases} \quad (4.16)$$
$$\begin{cases} 0,25 & \text{Si les frontières sont differentes} \\ 0 & \text{Si non} \end{cases}$$

Cette méthodologie nous permet de calculer un taux d'erreur pour la détection et la classification des entités $E_T(e_r,e_h)$, puis un taux d'erreur pour la décomposition $E_C(e_r,e_h)$. Une fusion linéaire des deux scores permet d'obtenir le score final

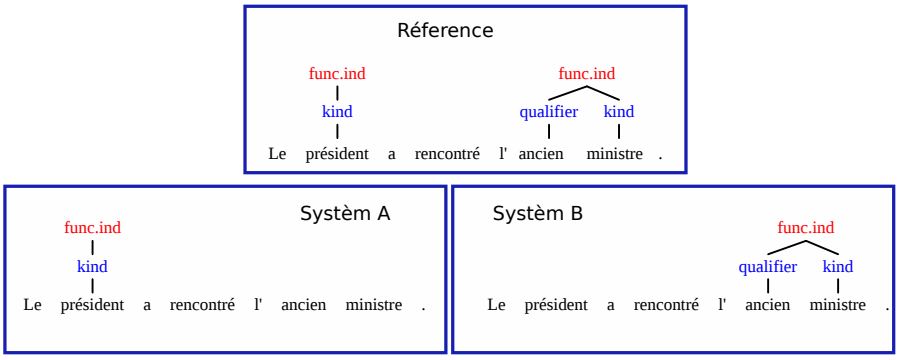

FIGURE 4.6 – Exemple de sorties de systèmes de REN

ETER.

$$ETER = \frac{I + D + \sum_{(e_r,e_h)}((1-\alpha)E_T(e_r,e_h) + \alpha E_C(e_r,e_h))}{N_E} \tag{4.17}$$

4.6 Analyses comparatives entre SER et ETER

Pour mieux comprendre la différence entre les mesures données par le SER et celles données par ETER, nous proposons de faire des analyses comparatives. D'abord, nous nous appuyons sur un jeu d'exemples choisis pour mettre l'accent sur les différences de comportement des deux métriques. Ensuite, nous utilisons des données réelles issues de campagnes d'évaluation, qui nous permettent d'interpréter l'impact de l'utilisation des deux métriques sur l'interprétation des performances des systèmes de REN.

4.6.1 Analyses fondées sur des exemples ciblés

Pour mieux comprendre les impacts de la structure complexe des entités nommées sur le comportement des métriques, nous allons nous appuyer sur quelques exemples pour comparer le comportement en termes de calcul de score entre SER et ETER et discuter des conséquences sur l'interprétation du résultat final.

Nous considérons d'abord la figure 4.6 dans laquelle nous pouvons voir que dans la référence nous avons deux entités, une première, *président*, qui contient un seul composant et une seconde entité, *ancien ministre*, qui contient deux composants.

Nous supposons maintenant que nous avons deux systèmes de REN différents A et B. Le système A annote correctement la première entité et oublie la deuxième, alors que le système B oublie la première mais annote correctement la deuxième.

Dans le cas du SER le calcul est effectué au niveau des slots et, par conséquent, le système A aura comme erreurs : D=3 (trois omissions), I=0, S=0 et C=2 (deux slots corrects). Le système B aura comme erreurs : D=2, I=0, S=0, et C=3. Ceci donne au système A un SER = $\frac{3}{5} = 60\ \%$ et au système B un SER =$\frac{2}{5} = 40\ \%$.

Ce résultat n'est pas cohérent puisque les deux systèmes de REN ont supprimé une seule entité, pourquoi donc considérer que le système B est meilleur que le système A ?
Dans le cas de ETER l'évaluation se fait au niveau des entités et non pas au niveau des slots. Par conséquent, les deux systèmes obtiennent D=1, I=0 et $E_T = 0$ (erreur sur les entités détectées) et par suite un ETER= 50%, ce qui représente le score le plus adapté dans cette situation.

Afin de mettre en évidence la prise en compte de la structure arborescente lors du calcul du score final dans ETER, nous considérons maintenant l'exemple des figures 4.4 et 4.5 comportant des entités imbriquées.

Nous pouvons voir dans les annotations référence que l'entité *ambassadeur de la Turquie en France* a comme classe *func.ind* et comporte trois composants, un *kind* (*ambassadeur*) et deux lieux (*loc.adm.nat* : *Turquie* et *France*). *Turquie* et *France* sont aussi deux entités qui contiennent chacune un composant *name*. Nous avons donc au total dans la référence trois entités, alors que le système de REN n'a détecté qu'une seule *func.ind*, *Turquie* et *France* sont alors considérées comme nom de Fonction.

Les alignements obtenus dans le cas de SER et ETER sont représentés respectivement dans les figures 4.4 et 4.5. En s'appuyant sur ces alignements, nous obtenons dans le cas du SER, quatre slots corrects et deux slots supprimés $SER = \frac{2}{6} = 33,33\ \%$. Alors que dans le cas de ETER, nous obtenons deux entités supprimées et une entité correctement détectée et classifiée mais avec des erreurs de décomposition $E_T = 0,16$, $ETER = \frac{2,16}{3} = 72\ \%$. Nous considérons que le score donné par ETER est pertinent dans ce cas de figure puisque nous avons deux entités supprimées et une mal décomposée.
L'interprétation de ces deux exemples montre que la métrique que nous proposons prend en compte la structure hiérarchique des entités nommées et les relations de dépendance qui relient les différents slots formant une entité.

4.7 Analyses comparatives fondées sur des données réelles

Afin d'illustrer la différence entre les mesures offertes par ETER et celles offertes par SER sur des données réelles, nous proposons ici une analyse comparative fondée sur l'interprétation des performances des systèmes de REN ayant participés à la campagne d'évaluation ETAPE [Galibert *et al.*, 2014]. La tâche de

reconnaissance d'entités nommées adoptée durant ETAPE est conforme aux règles définies durant le projet QUAERO. Pour la description détaillée du corpus ETAPE voir le chapitre 3.4. Le tableau 4.1 affiche les taux d'erreur mesurés par les deux métriques SER et ETER pour les différents participants de la campagne ETAPE, ainsi que les classements obtenus à partir de chacune des deux mesures.

Systèmes REN	SER		ETER	
	Taux d'erreur	Classement	Taux d'erreur	Classement
REN-1	35,4	1	34,4	1
REN-2	38,0	4	35,5	2
REN-3	51,4	8	50,0	7
REN-4	36,4	2	40,4	5
REN-5	37,3	3	38,7	4
REN-6	39,2	5	37,9	3
REN-7	50,0	7	52,6	9
REN-8	56,4	9	50,1	8
REN-9	44,6	6	42,8	6
REN-10	85,6	10	81,4	10

TABLEAU 4.1 – Comparaison des mesures de taux d'erreur offertes par SER et ETER (avec $\alpha = 0{,}5$) sur le test ETAPE

En comparant les mesures offertes par les deux métriques affichées dans le tableau 4.1, nous pouvons remarquer que pour tous les systèmes de REN, les taux d'erreur mesurés par ETER sont différents de ceux mesurés par SER. Nous pouvons noter également que pour certains systèmes de REN (REN-4, REN-5 et REN-7) le taux d'erreur mesuré par ETER est plus élevé que celui mesuré en utilisant le SER. Alors que d'autres (REN-1, REN-2, REN-3, REN-10...) affichent le comportement inverse. Par conséquent, les classements obtenus selon les mesures des deux métriques sont différents. Ces observations confirment que ETER et SER possèdent des comportements différents.

Pour mieux en comprendre les raisons, nous allons nous focaliser en détail sur l'analyse des performances de trois systèmes de REN (REN-4, REN-5 et REN-8) montrant des comportements différents vis-à-vis des deux métriques et qui abordent la problématique de reconnaissance d'entité nommées avec des approches différentes :

- REN-4 : [Raymond, 2013], ce système ignore la structure des entités et considère que tous les slots types et composants sont complètement indépendants ;
- REN-5 : [Dinarelli et Rosset, 2012], ce système construit les arbres (entités) en deux étapes. Tout d'abord, un modèle CRF détecte les composants, et

puis un modèle PCFG applique des méthodes d'analyse afin de construire l'arborescence des entités ;

- REN-8 : [Hatmi *et al.*, 2013], ce système ne prend pas la décomposition en compte et considère ainsi que tous les composants sont des **name**.

	D	I	S	SER	ETER
REN-4 types	1 925	155	541	39,5 %	
REN-4 composants	2 304	310	587	33,7 %	
REN-4 tâche globale	4 229	465	1128	36,4 %	40,0 %
REN-5 types	1 476	543	888	41,0 %	
REN-5 composants	2 092	525	589	33,7 %	
REN-5 tâche globale	3 568	1 068	1 477	37,3 %	38,6 %
REN-8 types	1 511	290	818	37 %	
REN-8 composants	4 295	266	2 908	69,7 %	
REN-8 tâche globale	5 806	556	3 726	56,4 %	50 %

TABLEAU 4.2 – Statistiques sur les types d'erreurs commises par les trois systèmes REN-4, REN-5 et REN-8, ainsi que les mesures de performance données par SER et ETER

Le tableau 4.2 affiche les résultats obtenus par ces trois systèmes. Nous pouvons distinguer quatre mesures différentes :

- un SER calculé sur les slots types uniquement ;
- un SER calculé sur les slots composants uniquement ;
- un SER global calculé sur tous les slots types et composants ;
- un ETER global calculé avec une valeur de $\alpha = 0,5$.

Pour essayer de comprendre les causes des comportements différents des deux métriques, nous allons nous intéresser aux performances des systèmes pour chacune des sous-tâches séparément. Une approche pour faire cela consiste à calculer le SER pour les slots types et les slots composants séparément. Nous pouvons voir dans le tableau 4.2 que REN-8 affiche un SER moins élevé sur les slots types (37,0 %) que sur les slots composants (69,7 %). Sur la tâche globale, ce même système obtient un ETER (50,0 %) moins élevé que le taux d'erreur donné par SER (56,4 %).

Les deux autres systèmes REN-4 et REN-5 affichent des SER moins élevés sur les slots composants (33,7 % et 33,7 %) que ceux calculés sur les slots types (39,5 % et 41,0 %). Contrairement au comportement de REN-8, ces deux systèmes obtiennent des ETER (40,0 % et 38,6%) plus élevés que les taux d'erreur donnés par SER (36,4 % et 37,3 %) sur la tâche globale.

Ces comportements sont dus à l'existence d'un biais dans les taux d'erreur mesurés par SER qui favorise les systèmes ayant des meilleures performances en

	Test-ETAPE	Pourcentage
Nombre de slots types	5 954	40,8 %
Nombre de slots composants	8 627	59,2 %
Total	14 581	100%

TABLEAU 4.3 – Description du corpus test ETAPE en termes de nombre de slots types et composants

décomposition (traitement des composants) par rapport à la classification (traitement des types). Ainsi, les systèmes ayant des bonnes performances en décomposition comme dans le cas de REN-4 et REN-5 auront un SER qui surestime leurs performances sur la tâche globale. Alors que les systèmes ayant des mauvaises performances en décomposition, comme dans le cas de REN-8, verront leurs performances abaissées sur la tâche globale.
Pour mieux comprendre la source de ce biais nous proposons d'étudier en détail le calcul fait par SER :

$$SER = \frac{\text{Nombre de slots erronés}}{\text{Nombre de slots dans la référence}} \tag{4.18}$$

$$= \frac{\overbrace{\sum_{i=0}^{NbT}(\text{erreur de type})}^{\text{somme des erreurs de classification}} + \overbrace{\sum_{j=0}^{NbC}(\text{erreurs de composant})}^{\text{somme des erreurs de décomposition}}}{\text{Nombre de slots dans la référence}} \tag{4.19}$$

avec :
NbT : nombre de slots types dans l'hypothèse ; NbC : nombre de slots composants dans l'hypothèse ; NbC $\geq$ NbT.

Comme nous pouvons le voir dans le tableau 4.3, dans le cas des données réelles NbC est toujours supérieur à NbT (NbC est 1,5 fois plus grand que NbT). Par conséquent, l'utilisation d'une métrique qui comptabilise les erreurs par rapport aux slots comme le fait le SER, conduit à une favorisation de la décomposition par rapport à la classification, et ce, à cause de la supériorité numérique des slots composants, introduisant ainsi un biais important lors de l'évaluation des performances des systèmes pour la tâche globale. Dans le cas de ETER les erreurs sont comptabilisées au niveau des entités et non pas au niveau des slots. L'utilisation de ETER avec un $\alpha = 0,5$ garantit un poids égal à la classification et à la décomposition indépendamment du nombre de slots concernés par chaque sous-tâche.

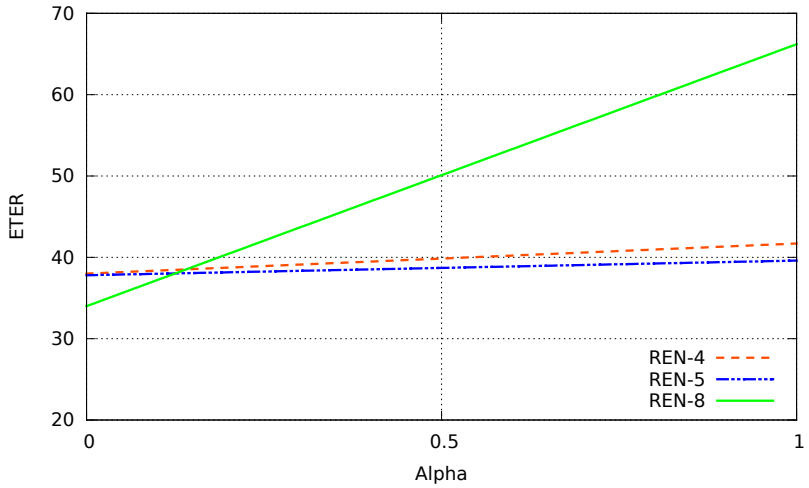

FIGURE 4.7 – Évolution des performances des trois systèmes de REN en fonction de la variation d'alpha

4.8 Impact du changement du paramètre alpha sur l'interprétation du taux d'erreur

Par défaut, la valeur d'alpha est fixée à 0,5 qui est la valeur la plus équilibrée pour l'évaluation de la tâche globale. Toutefois, l'utilisateur peut faire varier la valeur d'alpha dans le but de prendre plusieurs mesures qui lui permettront de comprendre le comportement de son approche et de détecter les points forts et les points faibles de son approche.

La figure 4.7 illustre l'évolution des performances des trois systèmes de REN (REN-4, REN-5 et REN-8) en fonction de la variation d'alpha.

- $\alpha = 0$: évaluation des performances en détection et classification d'entités nommées, la décomposition n'est pas prise en compte ;
- $\alpha = 0,5$: valeur la plus adaptée pour l'évaluation de la tâche globale, elle offre un poids égal à la classification et à la décomposition ;
- $\alpha = 1$: évaluation des performances en détection et décomposition d'entités nommées, la classification n'est pas prise en compte.

Comme nous pouvons le remarquer sur la figure 4.7, quand $\alpha = 0$ REN-8 affiche des meilleurs performances que les autres participants. Ceci signifie que REN-8 est le meilleur système en détection et classification d'entités nommées. En revanche, son taux d'erreur grimpe très rapidement dès que la valeur de α commence à monter pour atteindre un ETER très élevé quand $\alpha = 1$. Ce comportement laisse voir que ce système n'est pas bon en décomposition.

Sur la même figure, nous pouvons voir que REN-4 et REN-5 obtiennent le même ETER quand $\alpha = 0$. Ceci veut dire que les deux systèmes possèdent des

performances similaires en détection et classification d'entités nommées. En revanche, avec une valeur de $\alpha = 1$ REN-5 affiche un taux d'erreur inférieur à celui de REN-4, ce qui signifie que REN-5 est plus performant sur la détection et la décomposition des entités nommées que REN-4, et par la suite sur la tâche globale également quand $\alpha = 0,5$.

Si nous comparons maintenant les résultats affichés dans le tableau 4.2 avec ceux de la figure 4.7, nous pouvons remarquer que REN-4 et REN-5 possèdent tous les deux le même SER pour les composants (33,7 %), alors que les résultats affichés sur la figure 4.7 suggèrent que REN-5 est meilleur sur la décomposition puisque il obtient un meilleur ETER quand $\alpha = 1$. Pour comprendre l'interprétation de cette observation nous allons regarder en détail comment le taux d'erreur pour la décomposition est calculé dans le cas de chaque métrique. Pour le SER, nous avons :

$$SER_{composant} = \frac{I_{composant} + D_{composant} + S_{composant}}{N_c}$$

Avec :

- $I_{composant}$: le nombre de composants insérés ;
- $D_{composant}$: le nombre de composants supprimés ;
- $S_{composant}$: le nombre de composants substitués ;
- N_c : le nombre de constituants dans l'entité référence.

Calculé ainsi, le $SER_{composant}$ correspond à une évaluation d'une tâche de détection et de classification de composants qui, dans ce cas, est supposée être indépendante de la tâche globale demandée qui est la détection, classification et décomposition d'entités nommées. Ceci cause de nombreux problèmes. D'abord, le $SER_{composant}$ suppose que la détection des composants est indépendante de la détection des entités, alors qu'en pratique ce n'est pas du tout le cas. En effet, la majorité des erreurs de $I_{composant}$ et $D_{composant}$ découlent d'erreurs d'insertion et d'omission d'entités qui impliquent automatiquement l'insertion ou l'omission de l'ensemble des composants contenus dans ces entités. Un tel biais met en cause la validité des mesures offertes par $SER_{composant}$. Plus important encore, si nous nous référons à la définition de la tâche globale, ce qui est demandé est la décomposition des entités détectées et non pas une détection de composants. Pour ETER, nous avons :

$$ETER_{\alpha=1} = \frac{I + D + E_{composant}}{N}$$

Avec :

- I : le nombre d'entités insérées ;
- D : le nombre d'entités supprimées ;
- $E_{composant}$: la somme des erreurs de décomposition calculée pour les entités correctement détectées ;
- N : le nombre d'entités dans la référence.

	Substitutions de type	Substitutions de sous-types	Substitutions de frontières
REN-4	167	56	260
REN-5	307	119	485
REN-8	371	137	341

TABLEAU 4.4 – Distribution des différents types d'erreur de substitution

Le $ETER_{\alpha=1}$ obtenu correspond à un taux d'erreur de détection et de décomposition d'entités nommées, ceci le rend plus pertinent avec la définition de la tâche à évaluer.

Nous pouvons aussi remarquer dans le tableau 4.2 que REN-4 et REN-5 ont obtenu des SER différents sur les slots types, mais qu'ils obtiennent les mêmes performances en détection et classification d'entités nommées sur la figure 4.7 quand $\alpha = 0$. Afin de pouvoir interpréter ces observations, nous allons regarder en détail les erreurs commises par les systèmes. Nous pouvons voir que REN-5 fait moins d'erreurs de suppression que REN-4 mais plus d'erreurs d'insertion et de substitution. Les erreurs de substitution sont traitées différemment dans les deux métriques. Dans le cas de SER, toutes les substitutions possèdent le même poids (toutes les substitutions coûtent 0,5). Alors que dans le cas de ETER nous distinguons trois types différents de substitution :

- substitutions de types : quand la classe principale n'est pas correcte (coût 0,5) ;
- substitutions de sous-types : quand la classe principale est correcte mais la sous classe est incorrecte (coût 0,25) ;
- substitutions de frontières : quand les frontières de début ou/et de fin d'entité sont incorrectes (coût 0,25).

REN-5 a commis plus d'erreurs de substitution que REN-4 et la plus grande partie de ces erreurs de substitution sont des substitutions de sous-types (voir tableau 4.4). Puisque ces derniers sont moins pénalisés dans ETER, REN-4 obtient de meilleures performances en détection et classification d'entités quand ETER est utilisé.

4.9 Discussion

L'évaluation est une étape importante dans le cycle de développement des technologies issues de la recherche. Le choix des procédures et des protocoles d'évaluation adaptés au contexte de l'évaluation et aux cadres applicatifs visés conditionne le succès de cette étape. C'est ainsi que pour l'évaluation de la technologie de REN, nous avons observé l'emploi de procédures d'évaluation différentes durant

les campagnes d'évaluation, notamment, celles qui se sont étalées sur de longues durées et qui comprenaient des modifications touchant la définition de la tâche, comme c'était le cas des campagnes MUC et ACE. L'introduction de la tâche de détection, classification et décomposition d'entités dans le cadre du projet QUAERO a conduit à l'apparition des entités nommées structurées et compositionnelles. Cette nouvelle tâche qui requiert deux niveaux d'annotations a donné lieu à des structures d'annotations complexes. Évaluer cette tâche en utilisant des métriques telles que le SER et la F-mesure développées initialement pour l'évaluation de tâches simples biaise les mesures de performance, principalement à cause de la non-prise en compte de la structure des annotations et des relations qui relient les composants aux types. Étant donné que la tâche de REN, que nous étudions dans le cadre de cette thèse, est celle de la détection, classification et décomposition d'entités, nous avons développé une procédure d'évaluation qui permet de remédier aux limites observées pour les autres métriques et de prendre en compte les caractéristiques de la tâche. Cette procédure inclut une nouvelle approche d'alignement ainsi qu'une nouvelle métrique d'évaluation.

Alignement Pour prendre en compte la relation qui relie les constituants aux types, nous avons proposé un alignement à deux niveaux. Le premier niveau permet d'aligner les entités de l'hypothèse avec celles de la référence en se fondant sur les informations des slots types, tout en choisissant la solution qui optimise le taux d'erreur relatif à ces slots types. Le second niveau permet d'aligner les constituants appartenant aux paires d'entités résultantes du premier niveau d'alignement tout en choisissant la solution qui optimise le taux d'erreur de décomposition. Cette solution a permis surtout de prendre en compte la structure *multi-slots* des entités nommées structurées et compositionnelles, pour lesquelles l'entité est constituée d'un ensemble de slots, le slot type et l'ensemble des slots constituants (au moins un) lesquels se trouvent à un seul niveau en dessous du slot type. L'approche proposée permet ainsi d'éviter d'aligner les slots d'une même entité avec des slots appartenant à des entités différentes. Toutefois, la solution de double optimisation (une première durant le premier niveau d'alignement et une autre durant le second niveau) ne garantit pas forcément une optimisation globale du taux d'erreur sur l'ensemble de l'entité. En effet, le résultat du premier niveau d'alignement conditionne le deuxième, mais le résultat du deuxième niveau d'alignement n'a aucune influence sur le résultat du premier niveau. Une optimisation globale consisterait à établir une liste de tous les alignements possibles entre les slots types de la référence et de l'hypothèse, puis une liste de tous les alignements possibles de leurs constituants respectifs, et enfin de choisir la solution qui optimise le score global. L'optimisation globale est beaucoup plus complexe algorithmiquement que la double optimisation avec un résultat similaire, puisque les seuls cas où il y aura des résultats différents sont les cas d'imbrications multiples qui sont des cas très rares.

Métrique d'évaluation La métrique proposée ETER est une mesure de taux d'erreur. Elle suit le même principe que le SER et permet de combler certaines de ses limites. Elle se distingue par sa capacité à évaluer des tâches complexes et offre la possibilité de gérer l'importance (le poids) des sous-tâches. Ces critères permettent à ETER d'être la métrique la plus adaptée pour évaluer la tâche de détection, classification et décomposition d'entités.

Elle peut être utilisée pour faire de l'évaluation comparative ou également pour faire de l'évaluation de diagnostic durant la phase d'optimisation des systèmes.

Dans le cas de l'évaluation comparative, grâce à sa capacité à équilibrer les poids des sous-tâches (classification et décomposition), elle permet d'éviter des biais pouvant favoriser certains systèmes par rapport à d'autres comme nous l'avons montré dans l'analyse des résultats.

L'évaluation de diagnostic, nécessite l'utilisation de plusieurs mesures et de différentes métriques pour pouvoir caractériser les problèmes des systèmes et identifier les solutions qui permettent de les résoudre. Pour ce type d'évaluation, le choix de la métrique d'évaluation dépend de l'aspect qui intéresse l'évaluateur. Dans le cas de la tâche de détection, classification et décomposition d'entités, il est important d'analyser les performances des systèmes sur chacune des sous-tâches, pour mieux comprendre le comportement du système.

ETER permet aux développeurs de séparer facilement les performances relatives à chacune des sous-tâches (classification et décomposition) et les aide ainsi dans leurs tâches d'optimisation.

Le choix de la métrique d'évaluation dépend également d'une autre dimension qui est le contexte applicatif visé. Ainsi, suivant le cas dans lequel sera utilisé le système de détection d'entités nommées, l'importance de chacune des sous-tâches peut être différente. C'est pour cette raison que nous avons choisi d'accorder à l'évaluateur la possibilité de choisir le poids relatif aux sous-tâches.

4.10 Conclusion

Dans ce chapitre, nous avons détaillé le processus d'évaluation de la tâche de REN, et nous avons discuté les points faibles et les points forts des métriques d'évaluation utilisées dans la littérature. Après avoir démontré que les métriques existantes ne sont pas adaptées pour évaluer la tâche de détection, classification et décomposition d'entités, nous avons introduit une nouvelle métrique ETER (*Entity Tree Error Rate*) qui prend en compte la structure arborescente des entités structurées et compositionnelles. ETER permet à l'utilisateur de fixer le poids qu'il souhaite accorder à chacune des deux sous-tâches (la classification et la décomposition des entités) suivant le but expérimental ou/et le cadre applicatif visés.

4.4.10 Conclusion

Afin de valider notre proposition, nous avons comparé le comportement de ETER à celui de SER (métrique précédemment utilisée pour l'évaluation de la tâche) en utilisant une sélection d'exemples et en s'appuyant sur des données réelles issues de la campagne d'évaluation ETAPE.

Étant donné que la tâche de REN étudiée dans le cadre de cette thèse est celle de détection, classification et décomposition d'entités introduite dans le projet QUAERO, nous adoptons ETER comme mesure de performance pour l'évaluation de cette tâche dans la suite les expériences réalisées.

Troisième partie

Évaluation en contexte applicatif

Chapitre 5

Estimation de la qualité de la transcription automatique pour l'extraction d'entités nommées

5.1 Introduction

La reconnaissance automatique de la parole (RAP) constitue une brique technologique clé pour de nombreuses applications car elle donne accès au contenu d'une grande quantité de données se trouvant sous format audio et audiovisuel. Grâce aux efforts de la communauté du traitement automatique de la parole, les systèmes de RAP arrivent aujourd'hui à reconnaître de la parole continue à grand vocabulaire, et ont acquis de la robustesse contre les sources de bruits et de variations se trouvant dans le signal de la parole. Toutefois, la technologie n'est pas encore parfaite et des erreurs subsistent dans les sorties des systèmes de RAP. Ces erreurs résiduelles sont plus ou moins gênantes suivant l'application voulue. Nous avons vu dans le chapitre 2 que de nombreuses campagnes d'évaluation s'intéressent et encouragent le développement des systèmes de traitement automatique de la parole. Nous avons également vu que des cas d'incohérence entre les mesures données par le WER (*Word Error Rate*, la métrique classiquement utilisée pour l'évaluation des systèmes de RAP) et les performances globales de l'application ont été signalés et que quelques mesures alternatives au WER ont été proposées. Mais, malheureusement, la plupart de ces mesures n'ont pas été testées sur des données réelles et n'ont pas été très utilisées.

Dans ce chapitre, nous commençons par présenter brièvement les principales causes des erreurs faites par les systèmes de RAP. Nous utilisons ensuite les données des campagnes d'évaluation ETAPE et QUAERO pour examiner l'évolution des performances des systèmes de reconnaissance d'entités nommées (REN) en fonction du WER et analyser l'impact des erreurs de transcription sur les systèmes

de REN. Ensuite, nous proposons une nouvelle mesure, ATENE (*Automatic Transcription Evaluation for Named Entities*), permettant d'évaluer la qualité des transcriptions automatiques pour la REN. Enfin, nous évaluons ATENE en la comparant aux mesures proposées dans la littérature.

5.2 L'origine des erreurs des systèmes de RAP

Les performances des systèmes de RAP varient selon le type et la qualité des signaux à transcrire. Même si le taux d'erreur est faible sur des données de bonne qualité, les performances peuvent chuter très vite dès qu'il s'agit de données bruitées ou de parole non native [Galibert *et al.*, 2014]. De nombreuses études se sont intéressées à la caractérisation des sources d'erreurs des systèmes de RAP dans le but de mieux les maîtriser et, par la suite, les réduire. Nous présentons dans cette section les principales causes d'erreurs de RAP afin de comprendre pourquoi ces erreurs persistent et pourquoi leur élimination reste difficile et très coûteuse.

5.2.1 La robustesse au bruit

Des parasites peuvent provenir des bruits se trouvant dans l'environnement. Les bruits peuvent être de natures différentes : musique de fond, bruit de voiture, souffle du vent, claquement de porte... Plusieurs méthodes ont été proposées dans la littérature pour la reconnaissance de la parole en milieu bruité. Nous pouvons ainsi trouver des méthodes fondées sur la séparation des sources [Makino *et al.*, 2007], la déréverbération de la parole [Naylor et Gaubitch, 2010] ou encore sur l'utilisation de canaux multiples pour l'amélioration de la qualité du signal (*multi-channel*) [Virtanen *et al.*, 2012]. D'autres méthodes plus spécifiques existent également pour le traitement des bruits non stationnaires [Li *et al.*, 2014] et pour la séparation de la musique [Demir *et al.*, 2013]. Malgré toutes les avancés atteintes, le problème de la reconnaissance de la parole en milieu bruité est loin d'être résolu.

5.2.2 La variabilité dans le signal de parole

Les variations peuvent être liées aux conditions d'enregistrement, comme l'acoustique de la chambre et les caractéristiques du canal (microphone, téléphone...) qui induisent une grande variabilité dans le signal de parole. À cela, s'ajoutent les variabilités liées aux locuteurs qui peuvent être des variations intralocuteurs ou interlocuteurs. Les variations intralocuteurs sont les variations qui peuvent être détectées dans la parole d'un même locuteur. Elles peuvent être liées à son émotivité, son débit de paroles, son style de vocalisation... Les variabilités interlocuteurs sont les variations qui peuvent être détectées dans la parole de locuteurs différents.

Elles peuvent être liées à une différence de sexe, d'âge, d'accent, de langue maternelle...
L'adaptation des systèmes de RAP à ces types de variations nécessite l'utilisation de données d'entraînement provenant de conditions d'enregistrement différentes et de données multi-locuteurs, afin d'éviter la spécialisation des systèmes de RAP à un locuteur ou à une condition.

5.2.3 Les difficultés inhérentes à la langue parlée

Si les variabilités et la qualité du signal constituent des défis auxquels la communauté du traitement automatique de la parole a plus au moins réussi à faire face, les ambiguïtés intrinsèques à la langue parlée constituent, quant à elles, un véritable casse-tête. Dans le cadre du projet EPAC (Exploration de masse de documents audio pour l'extraction et le traitement de la parole conversationnelle), une analyse des caractéristiques de la parole en français a permis de caractériser les principales causes d'ambiguïté pour les systèmes de RAP. Nous présentons ici brièvement les principales causes d'erreurs observées lors de cette étude, ainsi que des exemples tirés de [Bazillon *et al.*, 2008].

Les homophones Les homophones sont des mots qui possèdent une même forme orale (même suite de phonèmes) mais qui s'écrivent différemment. Le français est une langue particulièrement riche en homophones du fait que les monosyllabes homophones sont beaucoup plus nombreuses que dans les autres langues [Bazillon *et al.*, 2008]. Exemples :

- *foi/fois/foie/Foix* ;
- *lait/les/lais/laie*.

Le problème des homonymes ne se limite pas aux monosyllabes homophones. En effet, dans plusieurs cas, les systèmes de RAP arrivent à trouver la bonne suite phonémique, mais lui associent une mauvaise transcription orthographique. Voilà quelques exemples :

- *là je viens d'ouvrir* $\Rightarrow$ *l'âge viens d'ouvrir* ;
- *proches, hein !* $\Rightarrow$ *prochain*.

La confusion «e» ouvert et «e» fermé Cette ambiguïté rend parfois très difficile la distinction entre l'imparfait, le participe passé ou l'infinitif comme dans l'exemple : *l'enfant aimait sauter dans l'eau et l'enfant aimé sautait dans l'eau*.
Elle peut également créer une confusion entre des suites de mots monosyllabiques

et des verbes. En voilà quelques exemples :

- *je l'ai* ⇒ *geler* ;
- *les faits* ⇒ *l'effet* ;
- *des faits* ⇒ *défais*.

Assimilation « C'est la variation phonétique, entraînant la modification de la prononciation d'une consonne sourde au contact d'une consonne voisine sonore (ou l'inverse) » [Bazillon *et al.*, 2008]. Ce phénomène est plus souvent perçu dans le cas de la parole spontanée et il est à l'origine de confusions pour les systèmes de RAP. Voilà quelques exemples :

- *envie d(e) passer* ⇒ *vite passé* ;
- *pas d(e)sanitaires* ⇒ *pattes sanitaires* ;
- *coup d(e) fil* ⇒ *coûte fils*.

Les disfluences Ce sont l'ensemble des hésitations, faux départs, lapsus, répétitions... Ces phénomènes sont assez fréquents en parole spontanée et posent régulièrement des problèmes pour les systèmes de RAP. Voilà quelques exemples :

- *(beaucoup de) beaucoup de* ;
- *(peut-être) alors peut-être* ;
- *(pour le) pour l'événement*.

Les mots hors vocabulaire Les systèmes de reconnaissance de la parole sont fondés sur une approche bayésienne. Dans cette approche le signal de parole est assimilé à une suite de vecteurs de paramètres X. Reconnaître la suite des mots W qui a été prononcée dans signal revient à trouver celle ayant la meilleure probabilité $P(W|X)$ c'est-à-dire la suite de mots la plus probable étant donné le signal.

$$P(W|X) = \text{argmax}P(X|W).P(W) \tag{5.1}$$

$P(X|W)$ est donnée par un modèle acoustique. Elle représente la probabilité d'avoir la suite de vecteurs de paramètres X lorsque la suite de mots W est prononcée.
$P(W)$ représente la probabilité du mot W dans la langue cible. Elle est donnée par un modèle de langage. Le modèle de langage est appris sur de grandes quantités de données afin de couvrir le plus large vocabulaire possible de la langue à transcrire. Toutefois, si un mot n'est pas inclus dans le vocabulaire du modèle de langage, sa probabilité P(W) (donnée par le modèle de langage) est à égale

zéro et, par la suite, la probabilité de son apparition dans une séquence de mots P(W|X) est aussi égale à zéro (ou très petite si des probabilités très petites ont été prévues pour éviter les probabilités nulles)[Ketabdar *et al.*, 2007].

Les mots qui ne figurent pas dans le vocabulaire des systèmes de RAP sont appelés des mots hors vocabulaire (OOV, *Out Of Vocabulary*). Les OOV concernent principalement les mots rares et peu fréquents. C'est notamment le cas des mots issus des procédés d'agglutination (comme en allemand ou en turc), et des noms propres (noms de personnes, noms de lieux, noms d'organisations...). En effet, recenser tous les noms propres est très délicat du fait de leur grand nombre, de leur diversité et de leur généricité (de nouveaux noms propres apparaissent tous les jours). La détection des nouveaux noms propres et la mise à jour du vocabulaire sont très coûteuses et difficiles à entretenir [Karanasou et Lamel, 2010]. Les OOV demeurent un grand défi, difficile à relever pour la technologie RAP [San José *et al.*, 2010], d'autant que les noms propres sont particulièrement importants dans plusieurs cadres applicatifs dont la reconnaissance d'entités nommées.

5.2.4 Les enjeux en analyse d'erreurs de transcription automatique de la parole

La quasi-totalité des études qui se sont consacrées à l'analyse des erreurs des systèmes de RAP s'est intéressée aux causes des erreurs plutôt qu'à leur impact. Ces études ont dévoilé les ambiguïtés auxquelles les développeurs des systèmes de RAP doivent faire face pour réduire les taux d'erreur et ont aidé ainsi à rendre la technologie de RAP plus efficace. Toutefois, les erreurs de transcription persistent, l'élimination de ces erreurs résiduelles devient de plus en plus difficile et de plus en plus coûteuse en temps et en ressources [Liberman, 2010] et [Adda, 2011].
Malgré ces erreurs, les performances des systèmes de RAP restent satisfaisantes et permettent leur utilisation dans différents cadres applicatifs. Il est donc important de mesurer l'impact des erreurs de RAP sur les briques technologiques qui exploitent ces transcriptions bruitées.

Malheureusement, l'étude de l'impact des erreurs des systèmes de RAP sur des modules appliqués en aval est problématique. Ceci est principalement dû à l'absence d'une taxonomie fondée sur l'impact des erreurs de transcription sur des modules ou des applications. En effet, prédire l'impact des erreurs n'est pas évident, car il existe une infinité d'erreurs différentes. Ces erreurs peuvent toucher des mots jouant des rôles différents dans le texte (verbe, nom, déterminant, conjonction...). De plus, les erreurs apparaissent dans des contextes (syntaxique ou sémantique) différents que ce soit en isolation ou en présence d'autres erreurs qui peuvent aussi

être du même type ou de types différents. Un autre problème, lié à la nature des erreurs de transcription, vient du fait qu'une erreur de transcription peut déclencher une suite d'erreurs en cascade et créer un empan d'erreurs dans lequel il est difficile de déterminer l'impact des erreurs élémentaires. En effet, les erreurs d'un même empan sont dépendantes les unes des autres et il est difficile de déterminer leur impact si elles apparaissent isolées. Une étude, qui s'est intéressée à creuser la question de la gravité des erreurs de RAP [Luzzati *et al.*, 2014], montre qu'il est difficile de définir la notion de gravité surtout si le cadre applicatif dans lequel les transcriptions automatiques seront utilisées n'est pas connu à l'avance. Cette étude montre qu'il est difficile pour des annotateurs humains de se mettre d'accord (accord inter-annotateurs faible) sur le degré de gravité des erreurs hors contexte applicatif (le contexte n'était pas précisée aux annotateurs).

Étant donné l'absence de taxonomie fondée sur la gravité des erreurs de transcription qui pourrait permettre d'évaluer la qualité des sorties de RAP autrement qu'en utilisant les mesures quantitatives existantes (voir chapitre 2), les développeurs des systèmes de RAP continuent d'utiliser ces dernières malgré les désavantages que cela pourrait représenter. Ainsi, les questions que nous nous posons sont les suivantes :

- À quel point les mesures utilisées aujourd'hui (notamment le WER) permettent d'évaluer la qualité des transcriptions automatiques en cadre applicatif ?
- Quel est l'impact des erreurs des systèmes de RAP sur les systèmes de REN appliqués en aval ? Comment peut-on évaluer cet impact ?
- Comment peut-on évaluer la qualité des transcriptions automatiques de la parole en cadre applicatif (en particulier pour la REN) ?

5.3 Le WER et l'évaluation de la qualité des transcriptions automatiques pour la REN

Nous avons discuté dans le chapitre 2 des exemples d'études traitant la problématique de l'évaluation de la qualité des transcriptions automatiques selon le cadre applicatif. Ces études montrent que dans certains cas le WER reste adapté, alors que dans d'autres les mesures données par le WER sont incohérentes. Qu'en est-il pour la tâche de reconnaissance d'entités nommées (REN) à partir de la parole ?

5.5.3 Le WER et l'évaluation de la qualité des transcriptions automatiques pour la REN

Nous avons à notre disposition les données des campagnes d'évaluation ETAPE et QUAERO qui incluaient toutes les deux la tâche de reconnaissance d'entités nommées à partir de la parole. Ces deux campagnes françaises ont adopté la même tâche de REN que celle mise en place dans le cadre du projet QUAERO. Elles impliquent les entités nommées hiérarchiques et compositionnelles pour lesquelles nous avons décrit la définition et les méthodes d'évaluation dans la partie II (chapitre 3 et chapitre 4). La description des données (corpus ETAPE et QUAERO) est donnée dans la section 3.4.

5.3.1 Interprétation des résultats de la campagne ETAPE

Le corpus de test ETAPE a été transcrits par cinq systèmes de RAP différents auxquels s'ajoutent des transcriptions issues d'un *rover* fondé sur les sorties des cinq systèmes. Sept systèmes de REN ont été appliqués sur l'ensemble des transcriptions automatiques et sur les transcriptions manuelles de référence.

La figure 5.1 montre les performances des systèmes de REN mesurées en ETER (*Entity Tree Error Rate*) en fonction du WER. Ces résultats sont obtenus en utilisant les données de la campagne ETAPE.

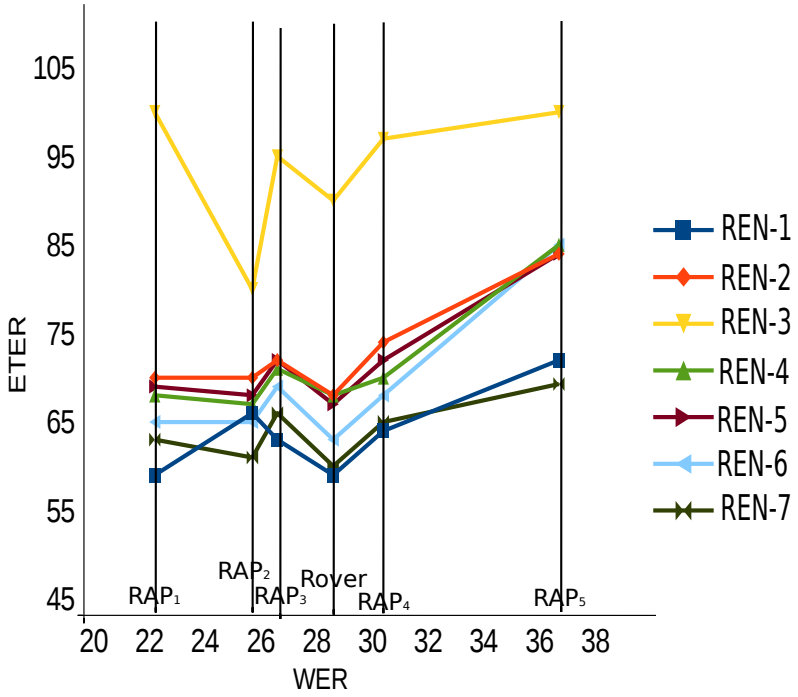

FIGURE 5.1 – Résultats de la campagne ETAPE avec les performances des systèmes de REN mesurées en ETER en fonction du WER.

Sur cette figure, nous pouvons voir que les meilleurs résultats en REN sont obtenus sur les sorties du rover malgré le fait que les transcriptions de ce dernier affichent un taux d'erreur de 28,7 % et sont ainsi classées quatrièmes en termes de WER.
Nous pouvons également remarquer que, pour la plupart des systèmes de REN (REN-3, REN-4, REN-5, REN-6 et REN-7), les sorties de RAP-2, classé deuxième en termes de WER (25,7 %), permettent d'obtenir de meilleures performances en REN que les sorties de RAP-1 classé premier avec 22,2 %.
Les transcriptions du système RAP-4 (30,4 %) permettent aux systèmes de REN d'obtenir des performances comparables (voire meilleures pour REN-3, REN-4, REN-6 et REN-7) à celles obtenues sur les sorties du système RAP-3 (26,6 %) malgré la différence assez importante entre leurs deux sorties en termes de WER. Le système RAP-5 classé dernier en termes de WER (36,7 %) permet, quant à lui, d'obtenir les plus mauvais résultats pour l'ensemble des systèmes de REN. L'écart important qui le sépare des autres systèmes donne ainsi une tendance croissante à la courbe.
À partir des résultats de la campagne ETAPE affichés sur la figure 5.1, nous pouvons déduire qu'une augmentation du WER n'implique pas forcément une diminution des performances en REN. Toutefois, un écart important en termes de WER semble se traduire par des baisses de performances pour les systèmes de REN. Dans ce cas, quand pouvons-nous considérer que la différence en termes de WER sera traduite en baisse de performances pour les systèmes de REN ? Les résultats suggèrent qu'il n'y a pas de tendance claire. En effet, un écart assez important en termes de WER (3,5 %) entre RAP-1 et RAP-2 n'engendre pas une baisse de performances pour la REN. Un comportement similaire peut être observé également en comparant les performances des systèmes de REN obtenues sur les sorties des systèmes RAP-3 et RAP-4 qui affichent une différence de 3,8% de WER. Alors qu'un écart réduit entre RAP-2 et RAP-3 (0,9 % de WER) se traduit par une baisse des performances en REN facilement observable.

5.3.2 Interprétation des résultats de la campagne QUAERO

Malheureusement, la campagne QUAERO a enregistré moins de participants que la campagne ETAPE. Trois participants ont soumis les sorties de leurs systèmes de RAP (transcriptions du corpus test). Trois systèmes de REN différents ont été appliqués sur ces trois sorties et sur les transcriptions manuelles de référence.

La figure 5.2 montre les performances observées des systèmes de REN en fonction du WER, sur les résultats de la campagne QUAERO.

Sur cette figure nous pouvons voir que les transcriptions du système RAP-2, classé deuxième en termes de WER, permettent d'obtenir les meilleures performances en REN pour tous les systèmes de REN. Nous pouvons aussi remarquer une différence assez importante entre les performances de RAP-3 (29 %) et celle

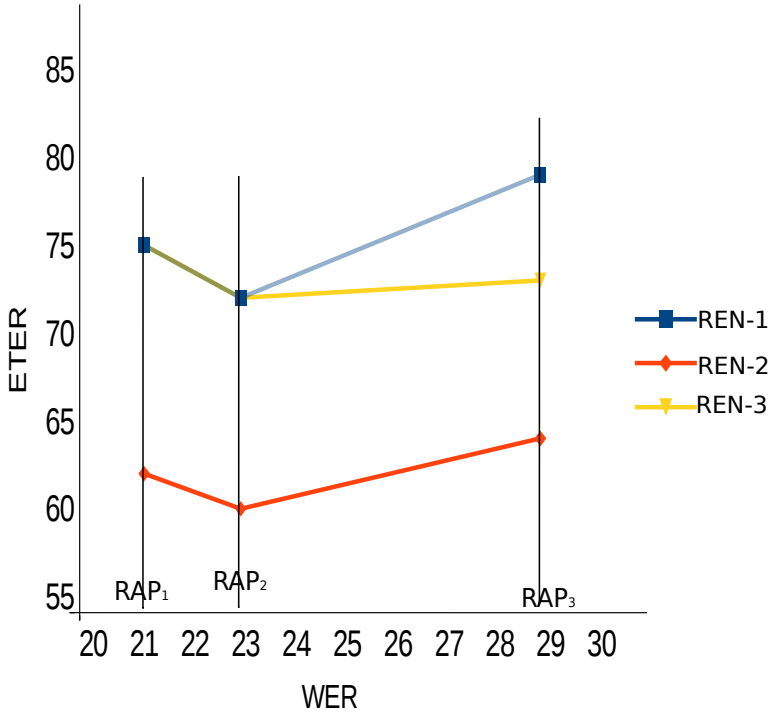

FIGURE 5.2 – Résultats de la campagne QUAERO. Les performances des systèmes de REN sont mesurées en ETER en fonction du WER. Les systèmes REN-1 et REN-3 obtiennent les mêmes résultats sur les sorties de RAP-1 et RAP-2.

de RAP-1 (21 %) et RAP-2 (23 %) en termes de WER qui ne se traduit pas par une baisse importante des performances en REN.

5.3.3 Discussion

Les résultats des campagnes d'évaluation ETAPE et QUAERO montrent que le WER n'est pas un bon indicateur de la qualité des transcriptions automatiques pour la REN. Toutefois, il semble que, dans les cas où la différence en termes de taux d'erreur est très importante (supérieure à 7 %), les prédictions du WER restent assez fiables, même si la baisse des performances des systèmes de REN n'est pas linéaire. Ce résultat n'est pas surprenant puisque les erreurs de transcription n'ont pas toutes la même importance pour la REN, alors que le WER reste une mesure qui affecte un poids égal à toutes les erreurs.

5.3.4 Impact des erreurs de RAP sur les systèmes de REN

L'interprétation des résultats des deux campagnes d'évaluation ETAPE et QUAERO montre qu'une augmentation des taux d'erreur de transcription n'implique pas forcément une baisse de performances pour les systèmes de REN appliqués en aval. Alors, quel est donc l'impact des erreurs de transcription automatique sur les systèmes de REN ? Est-ce que certaines erreurs sont plus gênantes que d'autres pour la REN ? Est-ce que des types d'erreurs de RAP différents engendrent un comportement également différent des systèmes de REN appliqués en aval ? Quels sont ces comportements ?

Les systèmes de REN font principalement trois types d'erreurs différents :

- les omissions ou délétions (D) : les entités de la référence qui ne sont pas détectées par le système et dont l'augmentation cause une baisse du rappel du système ;
- les insertions (I) (ou fausses alarmes) : les entités qui ne figurent pas dans la référence mais existent dans l'hypothèse et dont l'augmentation cause une baisse de la précision du système ;
- les substitutions (S) : les entités de la référence et de l'hypothèse qui sont alignées, mais possèdent des frontières de début ou de fin différentes, ou étiquettes (classes ou sous-classes) différentes.

Nous nous intéressons à l'étude de l'impact de la qualité des transcriptions automatiques sur les différents types d'erreurs de REN. Les données des campagnes d'évaluation ETAPE et QUAERO incluent les sorties des systèmes de REN sur les transcriptions de référence (manuelles) et sur les transcriptions automatiques des différents systèmes de RAP ayant participé aux campagnes. Afin de déterminer si certaines transcriptions automatiques favorisent un type d'erreur particulier pour les systèmes de REN appliqués en aval, nous proposons d'étudier la distribution des différents types d'erreurs de REN sur les sorties des systèmes de RAP.

Les systèmes de REN ne sont pas non plus parfaits et font des erreurs même sur des transcriptions manuelles ne contenant pas d'erreurs de transcription. Afin de séparer les erreurs de REN dues à celles des systèmes de RAP et celles causées par les défaillances des systèmes de REN, et plutôt que de regarder l'ensemble des erreurs de REN faites sur les transcriptions automatiques, nous nous intéressons uniquement aux erreurs de REN qui n'existent pas sur les transcriptions de référence. Pour ce faire, nous calculons pour chaque système de REN le pourcentage d'augmentation des erreurs (PAE) en faisant la différence entre les erreurs faites sur les transcriptions manuelles et celles faites sur les transcriptions automatiques. Pour chaque type d'erreur de REN (D, I et S) le PAE est calculé comme suit :

$$PAE(e) = 100 \times \frac{NB_A(e) - NB_M(e)}{NB_M(e)} \tag{5.2}$$

Avec :

- e est une erreur de REN de type omission, insertion ou substitution ;
- NB_A est le nombre des erreurs de type e sur les transcriptions automatiques ;
- NB_M est le nombre des erreurs de type e sur les transcriptions manuelles.

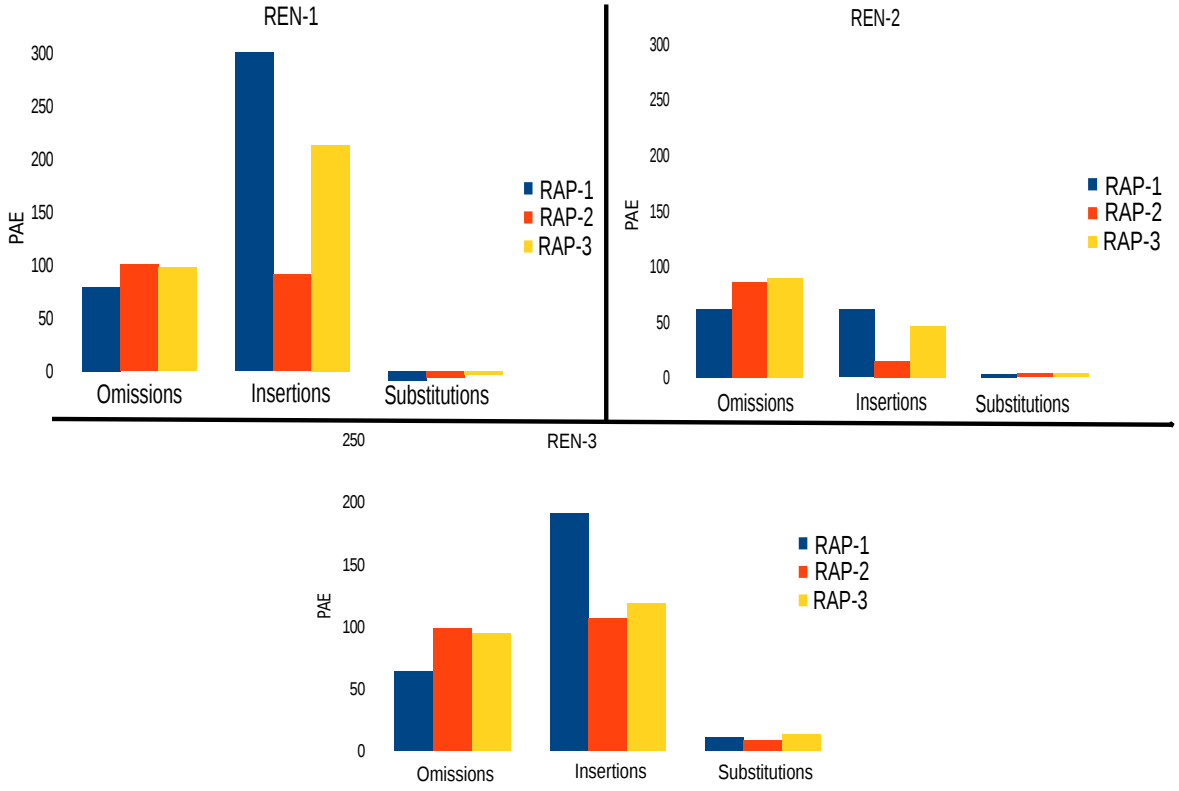

FIGURE 5.3 – Pourcentages d'augmentation des erreurs de REN sur les transcriptions automatiques par rapport aux transcriptions manuelles, résultats de la campagne QUAERO.

La figure 5.3 affiche les pourcentages d'augmentation des erreurs pour les trois systèmes de REN ayant participé à l'évaluation QUAERO. En regardant d'abord globalement les résultats affichés sur cette figure, nous pouvons voir que les erreurs (tous les types d'erreurs) augmentent d'une façon différente d'un système de REN à un autre. Ceci montre que les systèmes de REN possèdent des fonctionnements différents reflétés notamment par leur tolérance à certain types d'erreurs. Ainsi, nous pouvons voir que d'une façon globale le système REN-2 tolère moins les erreurs d'insertion que les deux autres systèmes.
En revanche, l'effet des erreurs de transcription automatique faites par les trois systèmes de RAP est le même sur l'ensemble des systèmes de REN. Ainsi, nous pouvons voir que les sorties de RAP-1 (en bleu) causent moins d'erreurs d'omission que les deux autres systèmes de RAP pour l'ensemble des systèmes de REN. En même temps, ces mêmes transcriptions (de RAP-1) sont celles qui causent le

plus d'erreurs d'insertion pour les systèmes de REN. Inversement, nous pouvons voir que les sorties de RAP-2 (en rouge) sont celles qui causent le plus d'erreurs d'omission mais le moins d'erreurs d'insertion pour les systèmes de REN. Le fait que ces comportements sont observables sur tous les systèmes de REN quelle que soit leur tolérance pour les erreurs d'insertion ou d'omission d'entités confirme que les transcriptions automatiques peuvent favoriser un type d'erreur particulier pour les systèmes de REN appliqués en aval.

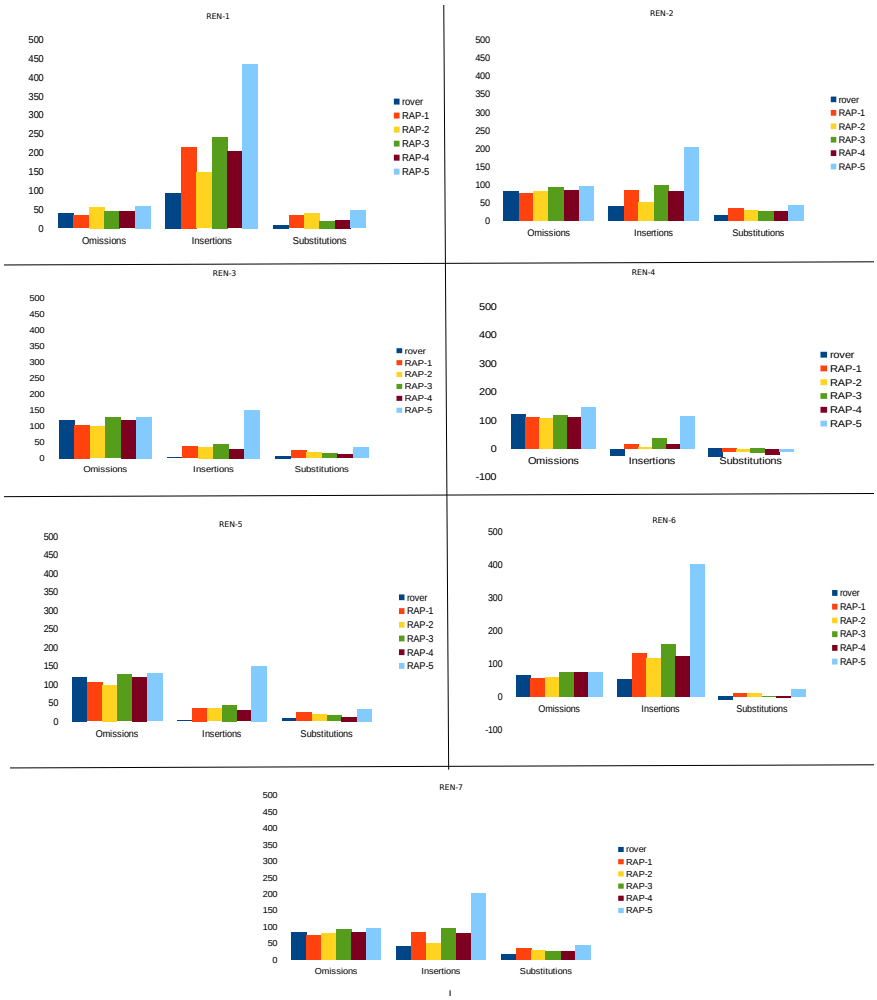

FIGURE 5.4 – Pourcentages d'augmentation des erreurs de REN sur les transcriptions automatiques par rapport aux transcriptions manuelles, résultats de la campagne ETAPE.

Afin de vérifier que ces tendances sont généralisables, la même expérience est réalisée sur les données de la campagne ETAPE. La figure 5.4 affiche le pourcentage d'augmentation des erreurs pour les systèmes de REN ayant participé à l'évaluation ETAPE. Sur cette figure, nous pouvons voir que les transcriptions du

rover (en bleu foncé) causent moins d'erreurs d'insertions que les transcriptions des autres systèmes de RAP, tandis que ces mêmes transcriptions causent plus d'erreurs d'omission que la majorité des autres transcriptions. Un comportement similaire, mais moins marqué, peut être observé sur les sorties de RAP-2 (en jaune).

Les observations faites sur les résultats des deux campagnes d'évaluation ETAPE et QUAERO montrent que les erreurs de transcription automatique ne conduisent pas simplement à une dégradation plus au moins forte des performances des systèmes de REN appliqués en aval, mais induisent des performances différentes. En effet, certaines erreurs favorisent l'omission des entités, d'autres favorisent leur insertion, ce qui implique une baisse du rappel ou de la précision des systèmes de REN. Ces tendances sont observées pour des systèmes de REN paramétrés différemment vis-à-vis de leur tolérance à chaque type d'erreur. Ceci veut dire que même un système de REN programmé pour réduire les erreurs d'insertion, par exemple, va voir ce type d'erreur augmenté si elles sont favorisées par les erreurs du système de RAP.

5.3.5 Insertion et suppression de mots *versus* insertion et suppression d'entités

Pourquoi certaines transcriptions automatiques favorisent-elles les omissions d'entités alors que d'autres favorisent leur insertion ? Une hypothèse intuitive consiste à dire que plus les systèmes de RAP insèrent de mots, plus il y aura d'entités insérées par les systèmes de REN, et plus ils suppriment de mots, plus les systèmes de REN suppriment des entités. Afin de vérifier cette hypothèse, nous proposons de tracer les courbes de PAE d'erreurs d'insertion et d'omission d'entités en fonction du pourcentage d'erreurs d'insertion et de suppression de mots réalisé par les systèmes de RAP.

5.3.5.1 Suppression de mots *versus* suppression d'entités

Les figures 5.5 et 5.6 affichent le comportement du PAE d'omission d'entités (fait par les systèmes de REN) en fonction du pourcentage des erreurs de suppression des mots par les systèmes de RAP respectivement sur les données des campagnes d'évaluation ETAPE et QUAERO. Les résultats de la campagne QUAERO, figure 5.6, suggèrent qu'il y a une corrélation entre le pourcentage des mots supprimés par les systèmes de RAP et l'augmentation des erreurs d'omission d'entités faite par les systèmes de REN. Cette observation reste à vérifier, d'autant que les participants (systèmes de RAP et systèmes de REN) à la campagne QUAERO ne sont pas nombreux et que ces observations sont en contradiction avec les résultats obtenus sur les données de la campagne ETAPE, figure 5.5, pour laquelle nous

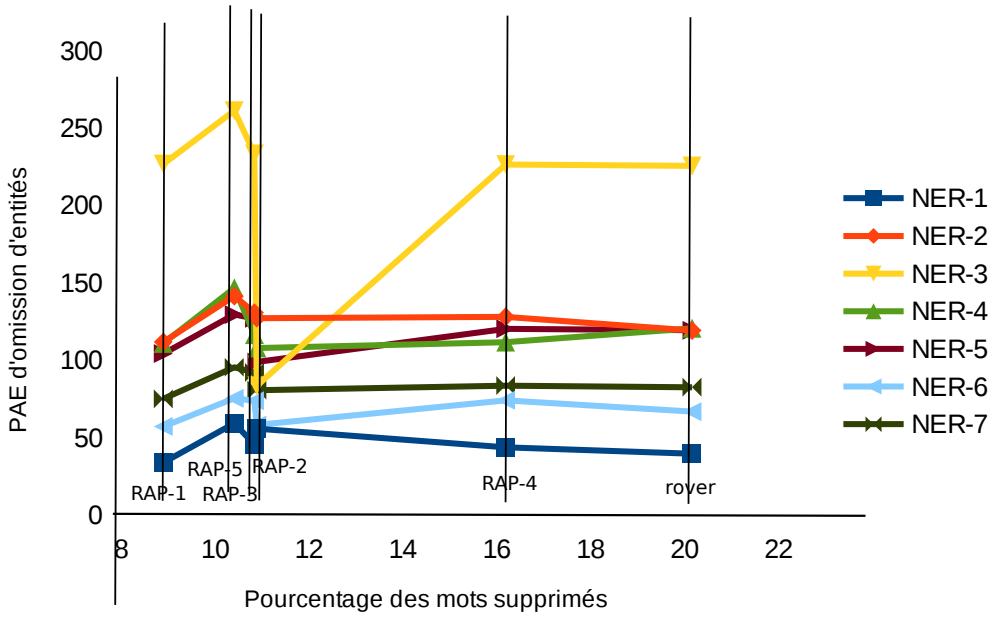

FIGURE 5.5 – Pourcentage d'augmentation des erreurs (PAE) d'omission d'entités en fonction du pourcentage des mots supprimés par la RAP, résultats de la campagne ETAPE.

disposons de plus de données. Ces résultats de la campagne ETAPE montrent qu'il n'y a aucune corrélation entre le pourcentage de suppression des mots dans les transcriptions automatiques et le PAE d'omission d'entités fait par les systèmes de REN.

5.3.5.2 Insertion de mots *versus* insertion d'entités

Les figures 5.7 et 5.8 affichent le comportement du PAE des insertions d'entités en fonction du pourcentage d'erreurs d'insertion fait par les systèmes de RAP et obtenu à partir des données des campagnes d'évaluation ETAPE et QUAERO. La figure 5.8, obtenue à partir des résultats de la campagne QUAERO, suggère que il n'y pas forcément de relation entre le pourcentage des mots insérés par les systèmes de RAP et l'augmentation des erreurs d'insertion d'entités, puisque les sorties du systèmes RAP-1 qui causent le plus d'erreurs d'insertion pour les systèmes de REN ne sont pas celles qui contiennent le pourcentage de mots insérés le plus élevé parmi les trois systèmes de RAP ayant participé à l'évaluation. Alors que les résultats de la campagne ETAPE, figure 5.7, pour laquelle nous disposons de plus de données, montrent que les systèmes de REN ont tendance à insérer plus

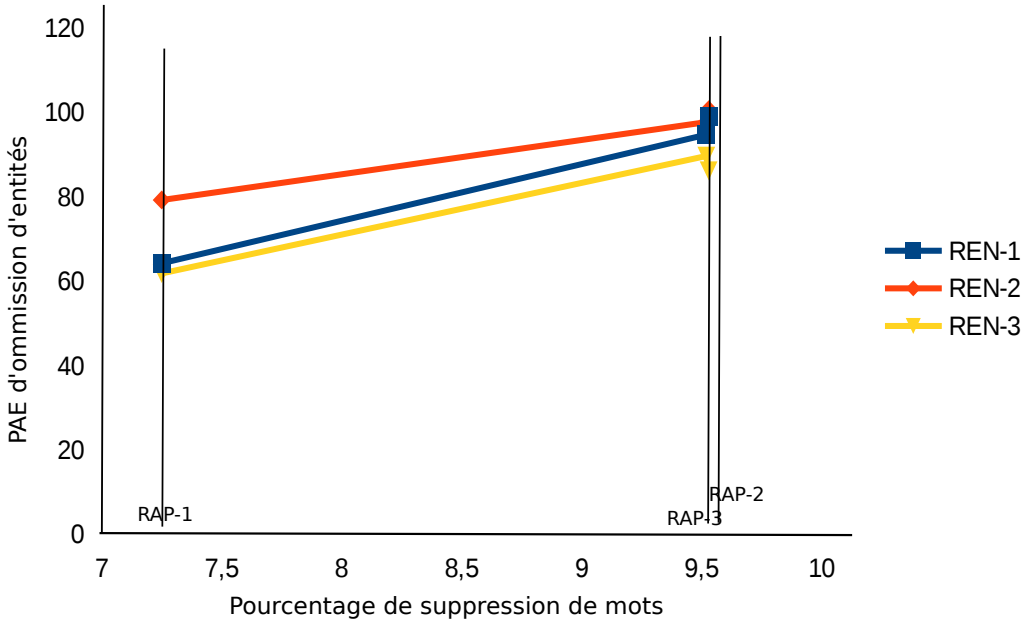

FIGURE 5.6 – Pourcentage d'augmentation des erreurs (PAE) d'omission d'entités en fonction du pourcentage des mots supprimés par la RAP, résultats de la campagne QUAERO.

d'entités quand les systèmes de RAP insèrent plus de mots. Cette tendance semble confirmée quand l'écart, en termes de pourcentage d'erreurs d'insertion de mots, est important entre deux sorties de RAP, comme celui que nous pouvons observer entre les sorties du rover (0,53 % de mots insérés) et celles de RAP-5 (5 % de mots insérés). Toutefois, lorsque les sorties de RAP affichent des écarts réduits en termes de pourcentage d'insertion de mots, il est difficile d'anticiper quelle sortie causera plus d'insertions d'entités pour les systèmes de REN.

5.3.5.3 Discussion

L'ensemble de ces observations montrent que les taux d'erreur d'insertion et de suppression des mots commis par les systèmes de RAP ne sont pas des bons indicateurs pour prédire les types d'erreurs de REN qui seront favorisés. Ceci n'est pas surprenant puisque à peu près 50 % des erreurs de transcription sont des erreurs de substitution de mots et que leur impact n'est pas pris en compte lorsque nous nous intéressons uniquement aux erreurs de suppression et d'insertion de mots. Nous avons noté qu'un taux élevé d'insertions de mots peut conduire à une

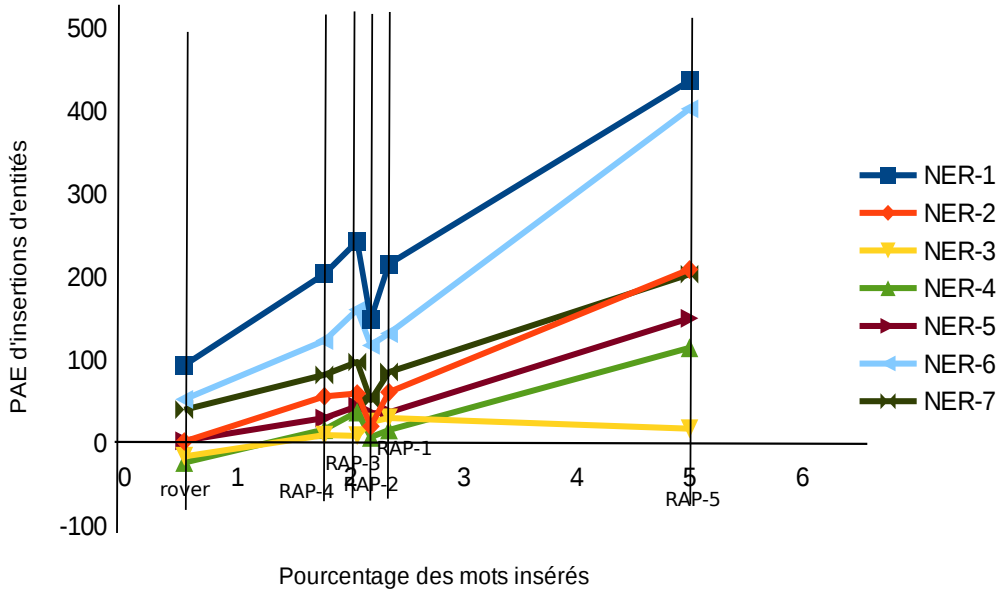

FIGURE 5.7 – Pourcentage d'augmentation des erreurs (PAE) d'insertion d'entités en fonction du pourcentage des mots insérés par la RAP, résultats de la campagne ETAPE.

augmentation des erreurs d'insertion d'entités. Toutefois, comme nous pouvons le voir sur les figures 5.7 et 5.8, il y a des nombreuses exceptions à cette règle. Ceci suggère qu'il existe des facteurs, au-delà d'un simple dénombrement d'erreurs, à prendre en compte pour l'évaluation de la qualité des transcriptions automatiques ainsi que leur impact sur les systèmes de REN. Dans la section suivante, nous nous intéressons aux approches et aux paramètres utilisés pour le développement des systèmes de REN afin de comprendre la nature des erreurs qui peuvent nuire à leur fonctionnement.

5.4 Les approches utilisées pour la reconnaissance d'entités nommées

Étudier les approches utilisées dans le développement des systèmes de REN nous permet d'identifier les paramètres sur lesquels sont fondées les décisions de ces systèmes. Ces paramètres sont importants puisque leur modification par des erreurs de RAP peut conduire les systèmes de REN à prendre des mauvaises décisions.

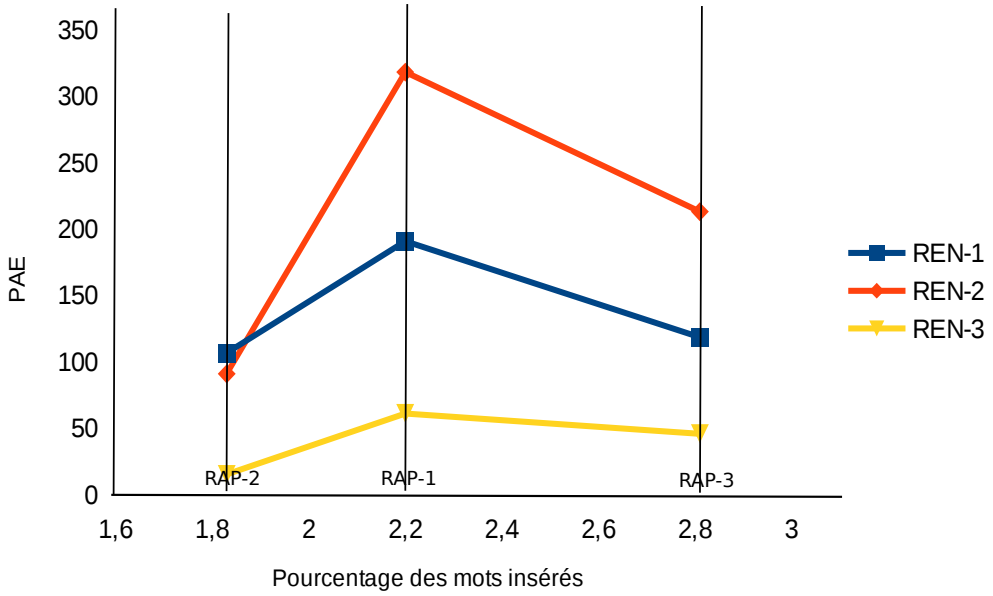

FIGURE 5.8 – Pourcentage d'augmentation des erreurs (PAE) d'insertion d'entités en fonction du pourcentage des mots insérés par la RAP, résultats de la campagne QUAERO.

Principalement, il existe deux familles d'approches permettant d'aborder la problématique de la REN, les approches orientées connaissances et les approches orientées données. Afin de profiter des avantages offerts par les approches, certains systèmes dits « hybrides » sont fondés sur des concepts tirés des deux approches. Nous n'allons pas traiter en détail toutes les étapes et les prétraitements nécessaires pour le développement d'un système de REN, nous allons nous restreindre à la description des approches et des paramètres utilisés. Pour plus de détails et d'exemples, nous renvoyons le lecteur aux travaux suivants : [Poibeau, 2001], [Friburger, 2002], [Fourour, 2002], [Nouvel, 2012] et [Nadeau et Sekine, 2007].

5.4.1 Les approches orientées connaissances

Les systèmes de REN orientés connaissances sont fondés sur des lexiques (des dictionnaires de noms propres de personnes, de pays, de villes, d'organisations...) et sur de règles produites manuellement par des experts. Les lexiques peuvent être formalisés à l'aide des règles pour créer des automates ou des grammaires génératives permettant de modéliser les contextes d'apparition des entités. Les approches orientées connaissances sont donc très dépendantes des lexiques. Mal-

heureusement, les lexiques ne sont pas exhaustifs, puisqu'il est impossible d'établir une liste de toutes les entités nommées à cause de la caractéristique dynamique des langues qui fait que des nouveaux noms propres apparaissent tous les jours. La mise à jour et l'entretien des dictionnaires deviennent alors laborieux et peu efficaces. De plus, si la tâche de la REN demandée va au-delà d'une détection de noms propres, cela devient encore plus délicat de gérer tous les types de syntagmes avec des listes. Pour contourner cet obstacle, les systèmes à base de règles (orientés connaissances) se sont tournés vers l'utilisation d'indices morphologiques et morphosyntaxiques [McDonald, 1996]. Voici quelques règles fondées sur les traits (indices) morphologiques et morphosyntaxiques :

- les noms propres commencent par une majuscule (*Juppé*) ;
- les dates contiennent des chiffres arabes et des noms de mois (*14 juillet 1945*) ;
- les noms d'organisations peuvent être des acronymes (*CNRS*).

Pour des descriptions de systèmes de REN orientés connaissances, nous renvoyons le lecteur aux travaux suivant : [Wacholder *et al.*, 1997], [McDonald, 1996] et [Ji-Hwan Kim, 2000] pour l'anglais et [Brun *et al.*, 2010], [Stern et Sagot, 2010] et [Maurel *et al.*, 2011] pour le français.

5.4.2 Les approches orientées données

Les systèmes de REN orientés données se fondent sur les mêmes observations (indices) que les systèmes orientés connaissances pour détecter les entités nommées, mais, à la différence de ces derniers, ils apprennent à extraire automatiquement les règles leur permettant d'utiliser les observations. Cette approche requiert de grandes quantités de données d'entraînement annotées manuellement. En effet, disposer de nombreux exemples sous forme brute (sans annotation) et annotés permet d'apprendre à des systèmes à passer d'un format à l'autre en se fondant sur un ensemble de traits (*features*). Différents types de traits peuvent être trouvés dans la littérature, nombre d'entre eux sont communs avec les approches orientées connaissances.

Voilà quelques exemples de traits :

- n-grammes de mots (bigrammes, trigrammes...) ;
- morphologiques (les préfixes, suffixes, des n-grammes de suffixes et de préfixes...) ;
- syntaxiques (comme les parties du discours, arbres syntaxiques...) ;
- sémantiques (sorties de systèmes d'annotations sémantiques) ;
- dictionnaires de noms propres ;
- ...

L'apprentissage des systèmes se fait alors automatiquement grâce à des procédures itératives permettant d'ajuster les configurations du système. Les algorithmes d'apprentissage les plus utilisés sont les machines à vecteurs supports

(SVM) [Isozaki et Kazawa, 2002], les modèles de Markov à états cachés (HMM) [Bikel *et al.*, 1997], les modèles de champs conditionnels aléatoires (CRF) (voir [Béchet et Charton, 2010] et [Dinarelli et Rosset, 2012]) et les arbres de décision [Isozaki, 2001]. Ces algorithmes sont classiquement très utilisés dans des tâches de classification, alors que l'annotation en entités nommées implique également une tâche de segmentation. La méthode la plus utilisée pour prendre en compte la segmentation est fondée sur le format BIO (*Begin, Inside, Outside*). Ce format consiste à affecter le label « B-type » (par exemple B-pers pour une entité de type Personne) pour les mots se trouvant en début d'entité. Si l'entité est composée de plusieurs mots, les mots qui ne sont pas en début d'entité reçoivent le label « I-type ». Enfin, les mots qui ne font partie d'aucune entité reçoivent le label « O » (*outside*). Ce format permet ainsi de déterminer à chaque fois la suite de mots composant une même entité.

5.5 Mesure proposée

Les observations faites dans la section 5.3 montrent que le WER n'est pas un bon indicateur de qualité des transcriptions automatiques pour la REN. Elles montrent que les sorties de RAP engendrent des types d'erreurs différents en reconnaissance d'entités nommées favorisant une baisse du rappel ou de la précision des systèmes de REN. Il y a donc un besoin de mesures alternatives qui permettent de mieux refléter l'impact des erreurs faites par les systèmes de RAP sur le fonctionnement des systèmes de REN. En effet, les sorties des systèmes de RAP constituent une source importante de données pour de nombreuses applications. Les intégrateurs voulant inclure un module de RAP dans une application demandent des évaluations centrées sur le cadre applicatif visé, permettant de prévoir l'impact des erreurs du module de RAP sur le fonctionnement de l'application global. Une évaluation selon le cadre applicatif permet également aux développeurs des systèmes de RAP de configurer leurs systèmes selon l'application visée et ainsi de proposer des solutions plus efficaces.

Idéalement l'évaluation des systèmes de RAP, selon le besoin applicatif, peut se faire en intégrant directement les systèmes de RAP dans l'application finale pour mesurer l'impact sur la performance globale. Pratiquement cette solution est difficile, puisque les différents modules de l'application sont souvent développés par des laboratoires ou des entreprises différents. De plus, effectuer une évaluation systématique, par paire (système de REN, système de RAP), pose des difficultés d'accès aux systèmes et des coûts d'intégration bien plus large que ceux engendrés par l'intégration des deux systèmes finalement choisis. Pouvoir ainsi découpler l'évaluation des systèmes tout en conservant le contexte de l'application finale a un intérêt scientifique et économique évident.

5.5.1 Contraintes et propositions

Les erreurs de transcription des systèmes de RAP déforment les observations en entrée des systèmes de REN en modifiant des mots et/ou leurs contextes d'apparition et influent ainsi sur les décisions des systèmes de REN. Évaluer l'impact de cette influence sur les performances des systèmes de REN se heurte aux contraintes suivantes :

- la notion de gravité des erreurs est difficile à définir et nous ne disposons pas de taxonomie d'erreurs de transcription qui peut nous informer sur l'impact des erreurs de transcription pour l'extraction d'entités nommées. De plus il est très difficile de mettre en place de telle typologie (voir section 5.2.4) ;
- la quantité d'erreurs se trouvant dans les transcriptions automatiques ne permet pas de prédire leur impact sur les systèmes de REN, puisque les prédictions offertes par le WER ne sont pas en cohérence avec les performances obtenues par les systèmes de NER (voir section 5.3) ;
- les erreurs de RAP peuvent favoriser une baisse du rappel ou de la précision des systèmes de REN, un impact difficile à prévoir à partir d'une simple mesure du WER ou des erreurs d'insertion et de suppression de mots (voir section 5.3.4).

Nous pouvons identifier, à partir de la littérature, principalement deux types d'approches pour résoudre la problématique de l'évaluation des systèmes de RAP en contexte applicatif (discutée en détail dans la section 2.6). Il s'agit des approches fondées sur la mesure de la perte d'information générale telles que RIL (*Relative Information Loss*), WIL (*Word Information Lost*) et des approches fondées sur la mesure de taux d'erreur ciblé sur des passages importants pour l'application visée telle que NE-WER. Comme nous l'avons discuté dans la section 2.6 chacune de ces approches possède ses forces et ses faiblesses. Nous pensons qu'une meilleure solution consiste à améliorer et à combiner les principes proposés dans ces deux approches. Pour ce faire, nous proposons de mesurer la perte d'information utile causée par les erreurs de RAP, en ciblant les passages de textes importants pour l'application (dans notre cas la REN à partir de la parole). Pour mettre en évidence notre proposition, il faut nécessairement répondre aux questions suivantes :

- Quelles sont les informations utiles pour le fonctionnement des systèmes de REN ?
- Comment évaluer la perte d'information causée par les erreurs de RAP ?
- Quels sont les passages de textes importants pour la REN et comment les identifier ?

5.5.2 Méthodologie

5.5.2.1 Modélisation de l'information utile pour la REN

Les systèmes de REN, comme nous l'avons discuté dans la section 5.4, se fondent sur un ensemble de traits (observations) pour détecter et classifier les entités nommées. Même si la plupart des traits sont communs à l'ensemble des approches de REN, la manière de les utiliser varie d'un système à un autre. De plus, les développeurs des systèmes choisissent des sélections de traits différentes suivant leur méthodologie, le type des données à traiter ou les outils (analyseurs morphologique, syntaxique, sémantique ou autres) à leur disposition. L'importance d'un trait par rapport à un autre est difficile à cerner puisqu'elle varie suivant le contexte. Prenons, par exemple, la suite des mots « sur le plateau » présente dans les trois exemples ci-dessous :

- *sur le <loc.phy> plateau de Saclay </loc.phy>* ;
- *sur le plateau de <org.ent> BFM </org.ent>* ;
- *sur le plateau d'argent* ;

Cette suite de mots semble contenir des indices permettant de révéler la présence d'une entité nommée. Mais comment déterminer la gravité d'une erreur ou d'un ensemble d'erreurs qui modifient les traits à extraire à partir de cette suite de mots ? Comment savoir, étant donné le contexte, si la quantité d'informations transmise par les systèmes de RAP reste suffisante pour la REN ou non ?

Répondre à ces questions simplement en comparant les suites des mots de la référence (transcription manuelle) et celles de l'hypothèse nous semble insuffisant étant donné que les métriques de l'état de l'art utilisant ce principe restent peu efficaces. Nous proposons donc de comparer plutôt la probabilité de présence d'entités nommées étant donné les transcriptions de référence (sans erreur) et les transcriptions de l'hypothèse (contenant des erreurs). Plus les erreurs de transcription font baisser les probabilités des réponses recherchées (pour la REN) par rapport aux probabilités obtenues sur les transcriptions de référence, plus elles sont graves pour la REN.

Pour mettre en place cette méthodologie nous avons besoin d'un modèle probabiliste permettant de modéliser la présence d'entités nommées dans des données propres (ne contenant pas d'erreurs). Ce modèle doit aussi être simple, facilement reproductible et à faible coût. Mais, également, le modèle ne doit pas favoriser une approche de REN particulière.

Nous proposons d'utiliser un simple classifieur fondé sur des traits basiques, et utilisés dans les différentes approches de REN, notamment les sacs de mots (n-grammes de mots), les préfixes, suffixes et la présence de majuscules (si fournies par la RAP). Pour entraîner notre modèle statistique, nous avons choisi d'utiliser la méthode du maximum d'entropie (MaxEnt) qui a démontré son efficacité dans de nombreuses tâches de classification. Ces modèles sont particulièrement

adaptés à la prise en compte de multiples traits discriminants, qui peuvent être interdépendants. En cas de disponibilité de données, le développement des modèles MaxEnt est peu coûteux en temps et en développement. En plus, l'algorithme d'apprentissage est implémenté dans l'outil Wapiti [Lavergne *et al.*, 2010] développé au LIMSI (Laboratoire d'Informatique pour la Mécanique et les Sciences de l'Ingénieur). L'outil Wapiti permet, entre autres, l'accès à tous les paramètres du modèle nécessaire pour le développement de nos outils d'évaluation.

Pour le maximum d'entropie, la probabilité pour un mot *m* d'appartenir à une classe *c* selon un ensemble de fonctions caractéristiques (traits) $f_1 ... f_k$ et leurs poids associés $\lambda_1 ... \lambda_k$ (avec $Z(m)$ un facteur de normalisation sur le mot) est estimée selon la formule [Berger *et al.*, 1996].

$$P(c|m) = \frac{\exp(\sum_{i=1}^{n} \lambda_i \times f_i(t,c))}{Z(m)} \tag{5.3}$$

Les divers traits seront donc exploités par le modèle afin de déterminer les poids $\lambda_1 ... \lambda_k$ selon la formule 5.4. La théorie de l'apprentissage automatique fournit une procédure qui ajuste itérativement les poids à l'aide d'un algorithme de descente de gradient.

$$\lambda = \underset{\lambda}{\text{argmax}} \sum_{m,c} p(m,c) \times \log\left(p(c|m)\right) \tag{5.4}$$

Ce modèle tire avantageusement parti des multiples traits, potentiellement interdépendants, qui peuvent être relevés pour un mot donné. Dans les modèles MaxEnt l'estimation des probabilités des classes pour un mot donné ne dépend pas des estimations calculées pour les mots le précédant et lui succédant. Ceci les rend peu efficaces pour des tâches d'annotation de séquences telles que la REN puisque la non-prise en compte des relations de dépendance peut donner lieu à un étiquetage différent pour des mots appartenant à une même entité et rend difficile la détection des frontières de début et de fin d'entité. Toutefois, cette caractéristique les rend particulièrement intéressants dans notre cas puisque elle assure que les probabilités déterminées pour chaque mot ne dépendent que des observations faites sur l'ensemble de traits et empêchent ainsi la propagation de possibles erreurs ou imprécisions dues au modèle.

5.5.2.2 Les segments importants pour la REN

Identifier les segments importants pour la REN est nécessaire pour mettre en place une méthodologie d'évaluation ciblée. Afin de pouvoir identifier le type d'erreur de REN favorisé par les erreurs de la transcription automatique, nous avons décidé de distinguer deux types de segments dans les documents textuels :

- les segments contenant des entités nommées, sur lesquels les erreurs de RAP peuvent engendrer des erreurs d'omission ou de substitution d'entités pour les systèmes de REN ;

- les segments hors entités sur lesquels les erreurs de RAP peuvent engendrer des erreurs d'insertion d'entités pour les systèmes de REN.

La figure 5.9 montre un exemple de sortie de système de RAP contenant des erreurs dans les deux types de segments et signale les fautes de REN pouvant être engendrées par ces erreurs.

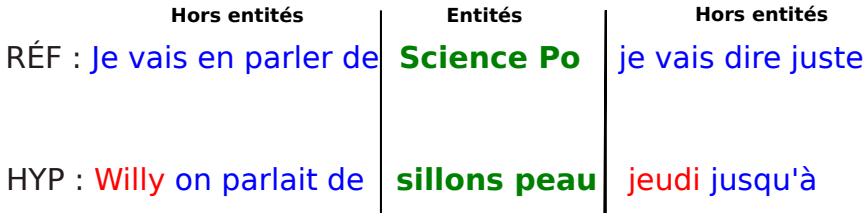

FIGURE 5.9 – Exemple de segmentation de texte en segments entités et segments hors entités, en haut le texte référence et en bas l'hypothèse RAP. Les segments entités sont en vert et les segments hors entités sont en bleu. Les mots écrits en rouge sont des mots hors entités qui ont été modifiés par des erreurs de RAP et qui sont susceptibles de causer des erreurs de REN.

Les erreurs de RAP qui modifient les entités (segments entités en vert) ou les mots se trouvant dans leurs contextes proches peuvent conduire à une omission des entités ou à leurs substitutions (erreurs sur les frontières ou la classe de l'entité) par les systèmes de REN. Nous pensons que la comparaison entre les probabilités de présence d'entités données sur la référence et l'hypothèse permet de nous renseigner sur la perte d'information causée par les erreurs de RAP. Ainsi plus la probabilité de présence d'entités sur l'hypothèse de RAP baisse par rapport à celle calculée sur les transcriptions de référence plus les erreurs de RAP sont graves pour la REN.

Les erreurs de RAP qui modifient les segments hors entités peuvent conduire les systèmes de REN à insérer des entités qui n'existent pas dans la référence. Typiquement les erreurs introduisant des noms propres, des dates ou des quantités (tels que les mots en rouge dans la figure 5.9) augmentent le risque d'insertion d'entités par les systèmes de REN.

La mesure que nous proposons est alors composée de deux mesures élémentaires. Une première mesure permet d'évaluer le risque d'erreurs d'omission et de substitution d'entités calculée sur les segments entités nommées et vise à prévenir contre la baisse de rappel des systèmes de REN. Une deuxième mesure permet d'évaluer le risque d'erreur d'insertion d'entités calculée sur les segments hors entités et vise à prévenir contre la baisse de précision des systèmes de REN. Fina-

lement, une mesure unique est tirée à partir des deux mesures élémentaires pour évaluer la qualité des sorties de RAP pour la REN.

5.5.2.3 Projection et segmentation

Notre méthodologie nécessite une segmentation des transcriptions manuelles et automatiques en segments entités et hors entités. Il est donc nécessaire que les transcriptions de référence (manuelles) soient annotées manuellement en entités nommées. Ainsi les frontières des entités constituent aussi les frontières des segments. Pour obtenir la même segmentation sur les transcriptions automatiques, nous projetons les annotations de la référence sur les sorties de RAP en utilisant une approche proche de celle décrite dans [Galibert *et al.*, 2011]. La projection se fait en plusieurs étapes. D'abord nous appliquons un alignement forcé du signal sur les transcriptions de référence, ceci nous permet de disposer des positions temporelles possibles des entités. Nous utilisons ensuite ces positions temporelles plus une marge pour déterminer les frontières possibles des entités. Nous gardons enfin la solution la plus proche phonétiquement de la référence.

Grâce à cette projection nous pouvons disposer d'une segmentation des sorties de RAP en segments entités et hors entités qui correspond à la même segmentation présente dans les transcriptions de référence.

5.5.3 La mesure ATENE

Dans le but de mesurer la perte d'information utile pour la REN causée par les erreurs de RAP, nous proposons une mesure fondée sur une comparaison des probabilités relatives à la présence d'entités sur les transcriptions de référence et sur les hypothèses de RAP plutôt qu'une comparaison classique de graphèmes. L'avantage de cette méthode est qu'elle permet de prendre en compte l'ensemble des informations contextuelles. Ainsi, le score calculé à chaque point va dépendre de la quantité d'information globale (sur l'ensemble des observations ou traits) perdue.

La tâche de REN, qui consiste à détecter et à classifier (déterminer la classe d'entités parmi les *n* classes possibles) les entités, est surtout un problème de prise de décision dans lequel les probabilités absolues sont moins importantes que l'écart entre les probabilités des différentes solutions possibles. Nous avons donc décidé de comparer la discriminance de la réponse correcte plutôt que des probabilités absolues. Étant donné une tâche de REN avec *n* classes d'entités différentes, l'ensemble des réponses possibles Y est de taille *n+1* (les *n* classes d'entités plus l'étiquette *Outside*), $Y = y_1, y_2..y_{n+1}$. Nous appelons discriminance de la réponse correcte l'écart de probabilités entre la probabilité de la réponse se trouvant

dans la référence $P(\hat{y})$ et celle de la plus probable des autres réponses possibles $\max_{y \neq \hat{y}} P(y)$. La discriminance de la réponse correcte est notée $Marge$ et elle est donnée par l'équation 5.5 :

$$Marge(x) = P(\hat{y}|x) - \max_{y \neq \hat{y}}(P(y|x)) \tag{5.5}$$

avec x un vecteur de traits (mots, préfixes, POS...), extrait à une position donnée du texte.

Une partie des erreurs faites par les systèmes de REN sur les transcriptions automatiques est due à des défaillances dans les systèmes de REN eux-mêmes. La différence entre la $Marge$ calculée sur les sorties RAP et la $Marge$ calculée sur les transcriptions de référence renseigne uniquement sur l'impact des erreurs de RAP. Nous notons la différence des marges (la $Marge$ sur sortie RAP et la $Marge$ sur référence) Δ_M, elle est donnée par l'équation 5.6 :

$$\Delta_M(x_H, x_R) = Marge(x_H) - Marge(x_R) \tag{5.6}$$

avec, x_H et x_R des vecteurs de traits extraits à partir de l'hypothèse et de la référence aux mêmes positions (les positions sont déterminées par la projection). $\Delta_M(x_H, x_R)$ peut ainsi être positive, négative ou nulle. Elle est positive si les erreurs de transcription favorisent la réponse recherchée $\hat{y}$, elle est négative si les erreurs de transcription augmentent le risque d'erreurs pour les systèmes de REN, et elle nulle s'il n'y a pas d'erreur de transcription.

5.5.3.1 La mesure élémentaire relative aux risques d'omission et de substitution d'entités : $ATENE_{DS}$

Les erreurs d'omission et de substitution peuvent se produire dans les segments d'entités nommées. Ces segments sont limités par les frontières de début et de fin d'entité. Dans le but d'évaluer le risque d'erreurs d'omission et de substitution introduit par l'hypothèse du système de RAP, nous calculons $\Delta_M(x_H, x_R)$ pour tous les mots se trouvant en début et fin d'entité. La mesure globale $ATENE_{DS}$ pour l'ensemble des transcriptions est la moyenne arithmétique de toutes les $\Delta_M(x_H, x_R)$, elle est donnée par l'équation 5.7 :

$$ATENE_{DS} = \frac{\sum_{i=1}^{N} \Delta_M(\text{début}_i) + \Delta_M(\text{fin}_i)}{2N} \tag{5.7}$$

avec N est le nombre d'entités dans la référence.
Plus le score donné par $ATENE_{DS}$ est négatif plus le risque d'erreurs d'omission et de substitution d'entités est élevé. Par la suite une baisse du rappel des systèmes de REN est attendue. Nous utilisons deux classifieurs différents pour calculer $ATENE_{DS}$. Un premier modèle de type « BO » (B : *Begin-type, O : Outside*)

pour le calcul de Δ_M pour les mots en début d'entité et un deuxième modèle de type « EO » (E : *End-type, O : Outside*) pour le calcul de Δ_M pour les mots en fin d'entité. Nous avons choisi d'utiliser deux modèles (« BO » et « EO ») puisque notre objectif n'est pas de développer un système de REN, mais uniquement de se concentrer sur les endroits où les entités sont présentes. Les classifieurs s'appuient sur des traits extraits à partir d'une fenêtre de décision de taille deux autour de la position du mot (les traits utilisés seront décrits plus tard dans la section 5.6). Le processus de calcul de $ATENE_{DS}$ est décrit dans la figure 5.10.

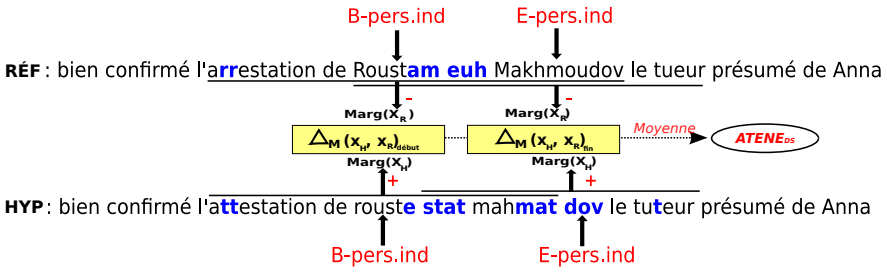

FIGURE 5.10 – Processus de calcul de $ATENE_{DS}$.

5.5.3.2 La mesure élémentaire relative au risque d'insertion d'entités : $ATENE_I$

Les erreurs de RAP dans les segments hors entités peuvent engendrer des erreurs d'insertion d'entités produites par les systèmes de REN. L'évaluation du risque d'insertion est sujette à des contraintes supplémentaires par rapport à l'évaluation du risque d'erreurs d'omission et de substitution. Tout d'abord, rappelons que déterminer a priori les mots susceptibles de causer des erreurs d'insertion est dépendant du système de REN donc difficile. Ensuite, il faut savoir que notre méthode de projection détermine la position des entités dans l'hypothèse de RAP, mais ne cherche pas à apparier un à un les mots de l'hypothèse et de la référence. Ceci nous empêche d'envisager une application de comparaison mot par mot des Marges sur les segments hors entités. Nous avons donc cherché à visualiser les distributions des erreurs d'insertion d'entités dans les segments hors entités. La figure 5.11 montre la distribution du nombre d'erreurs d'insertion d'entité par segment, calculée à partir des résultats de la campagne d'évaluation ETAPE. Nous pouvons voir que le plus grand nombre de segments contient zéro erreur d'insertion, quelques segments contiennent une seule entité insérée et très peu de segments contiennent plus qu'une seule insertion.

Pour plus de précisions, nous présentons sur la figure 5.12 les distributions des erreurs d'insertion pour différents systèmes de REN en nous focalisant sur les segments contenant au moins une erreur d'insertion. Comme nous pouvons

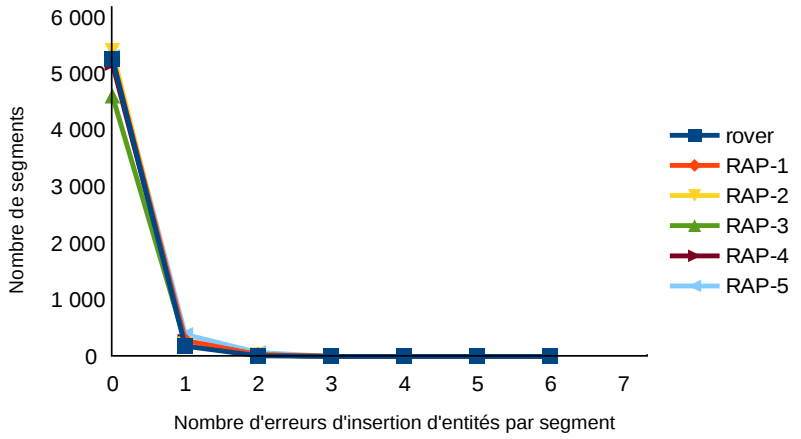

FIGURE 5.11 – Distribution moyenne des erreurs d'insertion faites par les systèmes de REN dans les segments hors entités, résultats de l'évaluation ETAPE.

le voir sur cette figure, le nombre de segments contenant une seule insertion est largement supérieur (au moins cinq fois plus) au nombre de segments contenant plus qu'une insertion. Nous pouvons également voir que les sorties de RAP sur lesquelles il y a le plus de segments contenant une seule erreur sont aussi celles qui contiennent le plus grand nombre de segments contenant plus qu'une seule insertion.

En se fondant sur ces observations, nous considérons qu'il suffit pour chaque segment hors entités d'estimer le risque d'avoir au moins une insertion.
Dans un segment hors entités tous les mots doivent être classés « O » (*Outside*), il y a donc un risque d'insertion si la probabilité de la réponse recherchée $P(O)$ ($\hat{y} = O$ *O : Outside*) est inférieure à la probabilité d'une autre réponse $P(y)\ avec\ y \neq O$. Dans ce cas la marge sera inférieure à zéro ($Marge < 0$). Pour comparer le risque d'erreur d'insertion dans les transcriptions de référence et dans l'hypothèse, nous comparons donc pour chaque paire de segments (hypothèse, référence) hors entités, les $Marges$ les plus négatives, afin de voir si le risque d'avoir au moins une insertion a augmenté à cause des erreurs de RAP. Si tous les mots du segment affichent des $Marges$ positives nous considérons qu'il n'y a pas de risque d'insertion dans le segment et nous fixons la $Marge$ du segment à 1. Ainsi, pour les segments hors entités la marge ($Marge'_O(S)$) est donnée par l'équation 5.8 :

$$Marge'_O(S) = \begin{cases} Marge_O(S) & si\ Marge_O(S) < 0 \\ 1 & si\ non \end{cases} \tag{5.8}$$

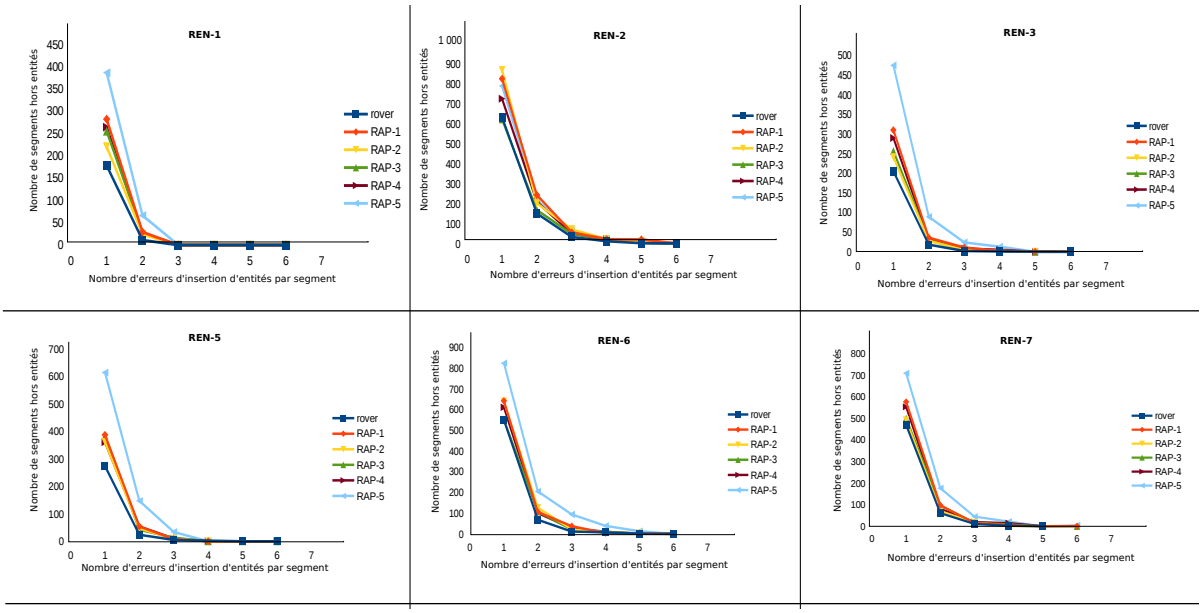

FIGURE 5.12 – Distribution des erreurs d'insertion dans les segments hors entités contenant au moins une insertion par différents systèmes de REN, résultats de l'évaluation ETAPE.

avec, S est un segment hors entités et $Marge_O(S)$ est la $Marge$ la plus négative calculée sur le segment S, elle donnée par l'équation 5.9 :

$$Marge_O(S) = \min_{x \in S} \left(P(O|x) - \max_{y \neq O} P(y|x) \right) \tag{5.9}$$

Pour évaluer si le risque d'avoir une insertion a augmenté sur un segment de l'hypothèse (sortie RAP), nous calculons $\Delta_{MO}(S)$ qui est la différence entre les marges $Marge'_O(S_H)$ et $Marge'_O(S_R)$ obtenues sur le segment de hypothèse et le segment de la référence. Un $\Delta_{MO}(S)$ négatif indique une augmentation du risque d'insertion d'entités dans l'hypothèse, alors qu'une valeur positive indique l'effet inverse. Finalement, un $\Delta_{MO}(S)$ nul indique que le risque d'insertion d'entités sur le segment est le même pour les transcriptions de l'hypothèse et de la référence. De fait, $\Delta_{MO}(S)$ est donné par l'équation 5.10 :

$$\Delta_{MO}(S) = Marge'_O(S_H) - Marge'_O(S_R) \tag{5.10}$$

La mesure globale $ATENE_I$ pour l'ensemble des transcriptions est la moyenne arithmétique de tous les $\Delta_{MO}(S)$, elle est donnée par l'équation 5.11 :

$$ATENE_I = \frac{\sum_{i=1}^{N_S} \Delta_{MO}(S_i)}{N_S} \tag{5.11}$$

avec N_S est le nombre de segments hors entités.

Plus le score donné par $ATENE_I$ est négatif plus le risque d'erreur d'insertion d'entités est élevé, avec par la suite, une baisse de la précision des systèmes de REN attendue.

Nous utilisons un classifieur MaxEnt de type « IO » (*I : Inside-type et O : Outside*) pour le calcul de $ATENE_i$.

5.5.3.3 Mesure de la qualité des transcriptions automatiques pour la REN : ATENE

Il est important d'avoir une mesure unique de la qualité des transcriptions pour des raisons pratiques, que ce soit pour faire une simple comparaison entre systèmes ou pour mettre en place une procédure d'optimisation. Nous définissons la mesure unique ATENE comme étant la moyenne des deux mesures élémentaires $ATENE_I$ et $ATENE_{DS}$. Les deux mesures élémentaires sont obtenues à partir d'une moyenne de scores compris entre -1 et 1 et elles sont calculées sur un nombre de segments presque égal. Ce comportement similaire laisse suggérer qu'une simple moyenne peut suffire. Ainsi, ATENE est donnée par l'équation 5.12 :

$$ATENE = -100\frac{ATENE_{DS} + ATENE_I}{2} \tag{5.12}$$

Nous multiplions le score final (négatif) par -100 pour améliorer la lisibilité des résultats.

5.6 Méthodologie de validation

L'expérience que nous décrivons ici a pour objectif de valider les hypothèses avancées et d'évaluer si les objectifs visés ont été atteints. Le choix des procédures à adopter est donc très important.
La nouvelle mesure proposée (ATENE) vise à évaluer la qualité des transcriptions automatiques pour l'application en aval d'un système de REN, et a pour objectif de permettre :

- la sélection du meilleur paramétrage [1] du système de RAP dans le cadre applicatif de reconnaissance d'entités nommées à partir de la parole ;
- la prédiction de l'impact des erreurs des systèmes de RAP sur le fonctionnement des systèmes de REN, notamment la baisse du rappel (si les omissions d'entités sont favorisées) et la baisse de précision (si les erreurs d'insertion sont favorisées).

1. Certains paramètres des systèmes de RAP sont fixés pendant la phase d'optimisation, comme le *fudge factor* (poids du modèle acoustique par rapport au modèle de langage) et le poids d'ajout d'un mot [Watanabe et Le Roux, 2014]

Dans le but d'atteindre ces objectifs, ATENE mesure la perte d'information (pour la REN) causée par les systèmes de RAP en ciblant les segments de texte importants pour la REN. Elle est fondée sur une approche statistique permettant de modéliser les informations nécessaires pour la REN. L'utilisation d'une telle approche n'est pas habituelle dans l'évaluation des systèmes en TAL et pose donc un certain nombre de questions. Particulièrement les questions suivantes :

- Est-ce que cette approche sera compatible avec des systèmes symboliques (orientés connaissances) ? Autrement dit, est-ce que cette approche ne va pas biaiser les résultats en faveur des systèmes à base de modèles statistiques ?
- Quels sont les traits à utiliser pour l'entraînement des modèles statistiques nécessaires à l'évaluation ? Quel est l'impact des changements de ces traits sur les mesures ?

Nous proposons ATENE comme une mesure alternative à ce qui existe dans la littérature qui permette d'évaluer la qualité des sorties RAP pour la REN avec plus de précisions et en offrant plus d'informations. Il est donc nécessaire de mettre en place une procédure offrant la possibilité de la comparer avec les différentes mesures existantes (WER, NE-WER, WIL...), en utilisant des données réelles issues de campagnes d'évaluation.
Mises à part les études qui se sont intéressées à mesurer la compatibilité du WER pour l'évaluation en contexte applicatif, il est difficile de trouver des tests utilisant les autres métriques proposées sur des données réelles. Ceci est probablement dû à la non-disponibilité des données. Il est encore plus difficile de trouver des études comparatives fondées sur des mesures faites sur des données issues de campagnes d'évaluation, probablement parce qu'il n'existe pas de plateforme ni d'outils permettant de mettre en place facilement une telle étude. Nous proposons donc de comparer ATENE aux autres mesures en utilisant les données des campagnes d'évaluation ETAPE et QUAERO à notre disposition. Cette comparaison sera à la fois utile pour évaluer notre proposition (ATENE) et aussi pour identifier les forces et les faiblesses de chaque mesure.

5.6.1 Description des données des campagnes d'évaluation ETAPE et QUAERO

Nous décrivons ici les données générés par les systèmes de RAP et de REN durant les campagnes d'évaluation ETAPE et QUAERO. Nous donnons également quelques précisions sur les différents systèmes. Les corpus d'apprentissage et de test sont décrits dans la section 3.4.

5.6.1.1 Description des données de la campagne ETAPE

La tâche de RAP de la campagne ETAPE consiste à transcrire automatiquement le corpus test composé quinze émissions de radio ou de télévision différentes. Cinq participants ont soumis les sorties de leurs systèmes pour l'ensemble des émissions. Un *rover* a également été généré à partir des hypothèses des cinq systèmes de RAP. Les performances des différents systèmes en termes de WER sont affichées dans le tableau 5.1.

	RAP-1	RAP-2	RAP-3	RAP-4	RAP-5	Rover
WER	22,3	25,7	26,6	30,4	36,7	28

TABLEAU 5.1 – Performances des systèmes de RAP en termes de WER pour la campagne évaluation ETAPE.

Les sorties du système RAP-2 sont capitalisées (contiennent des majuscules) alors que toutes les sorties des autres systèmes sont entièrement en minuscules.

La tâche de REN de la campagne ETAPE consiste à annoter automatiquement le corpus de test en entités nommées structurées et compositionnelles selon les règles du guide d'annotation QUAERO. La tâche inclut deux conditions, l'annotation du corpus test de référence (transcriptions manuelles) et l'annotation du corpus test transcrit automatiquement par les cinq systèmes de RAP. Sept participants ont soumis les sorties de leurs systèmes de REN pour les deux conditions.

La figure 5.1 affiche les performances des systèmes de REN en termes de ETER sur les sorties des différents systèmes de RAP. Les sept systèmes de REN utilisent des approches différentes pour aborder le problème. Seul le système REN-6 est fondé sur une approche purement symbolique (orientée connaissances), les autres systèmes sont plutôt fondés sur des approches statistiques (orientées données) ou hybrides.

5.6.1.2 Description des données de la campagne QUAERO

La tâche de RAP de la campagne QUAERO a consisté à transcrire automatiquement le corpus test composé dix-huit émissions de radio et de télévision différentes. Trois participants ont soumis les sorties de leurs systèmes pour l'ensemble des émissions. Les sorties des trois systèmes étaient capitalisées. Les performances des différents systèmes en termes de WER sont affichées dans le tableau 5.2.

La même tâche de la REN de la campagne ETAPE a été également adoptée dans la campagne QUAERO, avec également deux conditions que sont l'annotation du corpus test de référence (transcriptions manuelles) et l'annotation du corpus test transcrit automatiquement par les trois systèmes de RAP. Trois systèmes de REN ont soumis les sorties de leurs systèmes pour les deux conditions. La figure 5.2

	RAP-1	RAP-2	RAP-3
WER	21,0	22,9	28,8

TABLEAU 5.2 – Performances des systèmes de RAP en termes de WER à l'évaluation QUAERO.

affiche les performances des systèmes de REN en termes de ETER sur les sorties des différents systèmes de RAP. Le système de REN-3 était un système fondé sur une approche symbolique, alors que les deux autres étaient des systèmes principalement statistiques.

5.6.2 Les modèles statistiques et la sélection des traits

Notre méthode d'évaluation est fondée sur l'utilisation de probabilités fournies par des modèles statistiques. Les modèles statistiques sont entraînés selon l'approche MaxEnt en se fondant sur un ensemble de traits. Afin d'étudier l'impact du choix des traits et l'existence potentielle d'un biais, nous allons tester différentes sélections de traits. Comme nous l'avons expliqué précédemment, les traits doivent être simples (faciles à extraire) pour ne pas augmenter le coût de développement et doivent aussi être communs aux différentes approches utilisées en REN.

Pour l'extraction des POS (*Part Of Speech* ou parties du discours), nous avons utilisé l'analyseur syntaxique du LIMSI [Allauzen et Bonneau-Maynard, 2008]. Cet analyseur syntaxique offre des informations à plusieurs niveaux : syntaxique (fonctions syntaxiques des mots) et morphologique (genre, nombre, temps...). Nous n'avons utilisé que les informations syntaxiques dans les modèles que nous avons testés.

Nous avons donc choisi d'expérimenter l'ensemble des traits suivants :
- les uni-grammes et bigrammes de mots se trouvant dans une fenêtre de [-2,+2] autour du mot cible (position de décision) ;
- les préfixes et les suffixes des mots se trouvant dans une fenêtre de [-2,+2] autour du mot cible ;
- les uni-grammes et bigrammes de POS se trouvant dans une fenêtre de [-1,+1] autour du mot cible.

En se fondant sur ces traits nous avons développé quatre types de modèles différents :
- le premier (baseline) est fondé uniquement sur les uni-grammes et les bigrammes de mots ;
- le deuxième utilise les préfixes et les suffixes en plus de traits de la baseline (baseline+préf.+suf.) ;

- le troisième utilise les POS en plus de traits de la baseline (baseline+POS) ;
- le quatrième utilise tous les traits (baseline+préf.+suf.+POS).

La présence de majuscules est aussi un trait largement utilisé en REN. Toutefois, les systèmes de RAP ne fournissent pas tous les majuscules dans leurs sorties. Nous avons donc décidé de développer une variante pour chacun des quatre modèles proposés qui considère la présence de majuscules en plus des traits de base. Les différents modèles ont été entraînés sur les corpus d'apprentissage ETAPE et QUAERO à l'aide de l'outil Wapiti [Lavergne *et al.*, 2010].

5.6.3 Méthodologie de comparaison des mesures

Le choix de la méthodologie adaptée pour comparer la capacité des différentes mesures qui évaluent la qualité des transcriptions automatiques pour la REN est très important, à la fois pour la validation de ATENE et pour une bonne interprétation des résultats. Mais, avant de définir les différentes étapes de cette méthodologie, et avant de choisir les mesures nécessaires pour bien la mener, il est nécessaire de préciser ce que nous cherchons à mesurer exactement.

Nous proposons ATENE comme une mesure alternative aux autres mesures se trouvant dans la littérature pour évaluer la qualité des transcriptions pour la REN. Il est important de noter ici que pour un même système de RAP, nous nous attendons naturellement à ce qu'une augmentation du nombre d'erreurs de RAP conduise à une augmentation du nombre d'erreurs de REN. Nous souhaitons ici comparer des systèmes de RAP différents (approches différentes ou paramétrages différents). Or, comme nous l'avons montré dans la section 5.3 le taux d'erreur n'est pas la mesure la plus utile puisque les systèmes de RAP ont des comportements différents et font donc des erreurs de types différents. Il semble plus important, étant donné notre objectif, de pouvoir classer les systèmes de RAP suivant leurs impacts sur la baisse des performances des systèmes de REN.

Par conséquent, comparer la capacité des mesures d'évaluation à jauger de l'impact des sorties des systèmes de RAP sur les performances des systèmes de REN revient à comparer leur capacité à classer les systèmes de RAP selon les performances que ces derniers induisent en REN. Nous nous situons dans un cas de comparaison des classements entre ceux des systèmes de RAP selon les performances de REN et ceux des systèmes de RAP selon les scores fournis par les mesures d'évaluation. Or, la solution la plus adaptée pour comparer des classements est la mesure de corrélation de rangs. En effet, la corrélation de rangs, comme son nom l'indique, permet de mesurer la corrélation entre deux classements en se fondant sur les rangs établis par rapport à deux variables ou à deux observations différentes. Les méthodes les plus populaires pour la mesure de corrélation de rangs sont la corrélation de Kendall notée « τ » et celle de Spearman notée «

ρ » [Kendall, 1948]. La corrélation de Spearman reflète le degré de concordance et de discordance entre les rangs de deux classements, alors que la corrélation de Kendall reflète le nombre des concordances et des discordances entre les rangs de deux classements [Chok, 2010]. Les deux mesures de corrélation « τ » et « ρ » donnent des valeurs comprises entre -1 et +1 dont la valeur absolue indique la puissance de corrélation entre les deux variable testées. La corrélation de Kendall peut atteindre 1 pour un plus large nombre de scénario que la corrélation de Spearman [Chok, 2010]. Afin de nous assurer que nos résultats sont stables nous avons décidé de calculer les deux mesures τ et ρ pour mesurer la corrélation entre les classements des systèmes de RAP selon les performances obtenues par les systèmes de REN et les classements des systèmes de RAP selon différentes mesures d'évaluation de la qualité des transcriptions.

Le classement des systèmes de RAP, selon les performances des différents systèmes de REN, n'est pas unique et il est possible d'obtenir des classements différents selon le système de REN considéré. Ceci est probablement dû à des approches différentes utilisées pour le développement des systèmes différents eux aussi. Nous allons considérer des classements différents (classement des systèmes de RAP selon les performances en REN), un pour chaque système de REN, et considérer la moyenne des corrélations calculée pour tous les systèmes comme mesure de corrélation globale. L'avantage de cette méthode est qu'elle nous permet également d'observer si certaines mesures d'évaluation sont plus en corrélation avec des systèmes de REN particuliers ou non.
Plus le nombre d'exemples est important plus les mesures de corrélation sont robustes. Comme nous ne disposons que de six systèmes de RAP, sept systèmes de REN, dans le cas des données ETAPE, et seulement trois systèmes de RAP et de REN, dans le cas des données QUAERO, nous avons décidé de mesurer les corrélations pour les transcriptions relatives à chaque émission de radio et de télévision (quinze émissions par sortie RAP pour ETAPE et dix-huit pour QUAERO). Ceci nous permet de réduire l'incertitude et d'augmenter la puissance de notre test. La mesure de corrélation entre le classement des sorties de RAP selon les performances d'un système de REN donné et le classement fondé sur une des mesures d'évaluation sera donc la moyenne des corrélations calculée pour les transcriptions de l'ensemble des émissions.

5.7 Comparaison des mesures d'évaluation de la qualité des transcriptions automatiques pour la REN

Dans cette section nous présentons le résultat de la comparaison que nous avons mise en place en respectant la méthodologie décrite dans la section 5.6.3. Les mesures d'évaluation que nous comparons sont le WER (*Word Error Rate*), le

NE-WER (*Named Entities Word Error Rate*), le WIL (*Word Information Lost*) et la F-mesure (moyenne harmonique de précision et de rappel fondée sur les erreurs de RAP) et bien entendu ATENE (*Automatic Transcriptions Evaluation for Named Entities*). Nous rappelons que les différentes mesures tirées de la littérature sont décrites dans la section 2.6. Nous commençons d'abord par comparer les différentes mesures d'évaluation sur la base de leurs capacités à évaluer l'impact de la qualité des sorties de RAP sur les performances globales des systèmes de REN. Nous faisons ensuite la comparaison par rapport à leurs capacités à prédire les types d'erreurs de REN induites par les sorties de RAP.

5.7.1 Évaluation de l'impact des sorties de RAP sur les performances globales des systèmes de REN

		REN-1	REN-2	REN-3	REN-4	REN-5	REN-6	REN-7	moyenne
	WER	0,57	0,52	0,29	0,55	0,51	0,48	0,53	0,49
	NE-WER	**0,79**	0,68	0,31	0,64	0,60	0,67	0,59	0,61
	WIL	0,47	0,42	0,23	0,47	0,46	0,38	0,45	0,41
	F-mesure	0,47	0,42	0,23	0,47	0,46	0,38	0,45	0,41
ATENE	*Baseline*	0,70	0,79	0,47	**0,80**	0,64	0,73	0,71	0,69
	Base+préf.+suf.	0,74	**0,82**	0,45	0,77	0,66	**0,78**	0,73	**0,71**
	Base+pos	0,71	0,78	0,42	0,78	0,66	0,74	0,73	0,69
	Base+préf.+suf.+pos	0,74	0,80	0,43	**0,80**	**0,67**	**0,78**	**0,74**	**0,71**
	Baseline+maj.	0,39	0,64	0,59	0,69	0,53	0,68	0,66	0,60
	Base+préf.+suf.+maj.	0,43	0,75	**0,66**	0,77	0,56	0,71	**0,74**	0,66
	Base+pos+maj.	0,37	0,69	0,57	0,69	0,49	0,65	0,67	0,59
	Base+préf.+suf.+pos+maj.	0,47	0,74	0,56	0,77	0,61	0,69	0,73	0,65

TABLEAU 5.3 – Corrélations de Spearman moyennes entre les classements fondés sur les performances des systèmes de REN mesurées en ETER sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, pour les données de la campagne d'évaluation ETAPE.

Le tableau 5.3 affiche les corrélations de Spearman moyennes entre les classements des systèmes de RAP selon les performances globales des systèmes de REN mesurées en ETER et les classements des systèmes de RAP fondés sur les différentes mesures d'évaluation. Les mesures de corrélation de Kendall permettent les mêmes interprétations et seront donc affichées en annexe A.
Comme nous pouvons le voir sur le tableau 5.3, ce sont les mesures données par ATENE qui correspondent le mieux avec les performances des systèmes de REN. En effet, les mesures obtenues par ATENE utilisant des modèles sans majuscule

sont les meilleures et montrent une meilleure corrélation pour l'ensemble des systèmes de REN, sauf pour REN-1 pour lequel les mesures données par NE-WER semblent être un peu plus en accord. Ceci se traduit par un écart important en termes de corrélation moyenne entre ATENE (fondée sur les modèles sans majuscule) et les autres mesures qui montrent que notre méthode d'évaluation est plus efficace pour l'évaluation de la qualité des systèmes de RAP pour la REN. Nous pouvons remarquer également que les corrélations obtenues pour les mesures de ATENE fondées sur les quatre premiers modèles (sans majuscule) affichent très peu de variations bien que des combinaisons de traits différents soient utilisées. Ceci renforce notre méthodologie et montre que ATENE peut être utilisée même si les traits employés pour le développement des systèmes de REN cibles (pour l'application finale) ne sont pas connus. Nous observons une légère baisse de corrélation pour les mesures données par ATENE lorsque nous utilisons les modèles incluant les majuscules dans les traits. Ce comportement est causé par la présence de majuscules uniquement dans les sorties de RAP-2 ce qui conduit à biaiser les mesures de ATENE en faveur de ce système (pour les modèles qui incluent le trait des majuscules).

Nous pouvons également remarquer que les mesures de ATENE correspondent mieux avec les performances de REN-3 lorsque les majuscules sont prises en considération, alors qu'un comportement inverse est observé pour REN-1. Ceci suggère que les majuscules sont des traits importants pour REN-3, mais pas pour REN-1. Pour les autres systèmes de REN nous observons très peu de changements. Ceci est peut être dû à l'utilisation de systèmes de REN différents pour gérer les deux conditions (avec et sans les majuscules). Nous pouvons également voir que les mesures de corrélation obtenues entre ATENE et les performances du système REN-6 fondé sur des approches symboliques sont similaires aux mesures de corrélation obtenues pour les autres systèmes de REN qui s'appuient plutôt sur des approches statistiques.

ATENE a été définie suite à une étude du corpus ETAPE, il importe donc de s'assurer qu'il n'y a pas de biais lié à l'analyse de ce corpus spécifique. Afin de vérifier cela, nous appliquons la même méthodologie de comparaison et de validation sur les données de la campagne d'évaluation QUAERO. Les résultats de la corrélation de Spearman sont affichés dans le tableau 5.4. Ces résultats confirment que ce sont les mesures données par ATENE qui correspondent le mieux avec les performances des systèmes de REN. En revanche, nous pouvons remarquer que, contrairement aux données ETAPE, se sont les mesures de ATENE obtenues en utilisant les modèles qui prennent en considération les majuscules qui offrent les meilleurs résultats. Ce résultat était attendu, puisque, comme nous l'avons mentionné précédemment, tous les systèmes de RAP ayant participé à l'évaluation QUAERO ont soumis des transcriptions automatiques capitalisées (contenant des majuscules). Par conséquent, nous nous attendions à ce que ce trait important soit conservé par

5.5.7 Comparaison des mesures d'évaluation de la qualité des transcriptions automatiques pour la REN

		REN-1	REN-2	REN-3	moyenne
WER		0,39	0,36	0,11	0,28
NE-WER		0,19	0,17	-0,03	0,11
WIL		0,39	0,36	0,11	0,28
F-mesure		0,39	0,36	0,11	0,28
ATENE	*Baseline*	0,22	0,33	0,33	0,30
	Base+préf.+suf.	0,30	0,44	0,36	0,37
	Base+pos	0,33	0,13	0,30	0,25
	Base+préf.+suf.+pos	0,27	0,30	0,27	0,28
	Baseline+maj.	0,19	0,36	0,47	0,34
	Base+préf.+suf.+maj.	**0,44**	**0,50**	**0,55**	**0,50**
	Base+pos+maj.	0,19	0,16	0,5	0,28
	Base+préf.+suf.+pos+maj.	0,38	0,52	0,52	0,48

TABLEAU 5.4 – Corrélations de Spearman moyennes entre les classements fondés sur les performances des systèmes de REN mesurées en ETER sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, pour les données de la campagne d'évaluation QUAERO.

les développeurs des systèmes de REN pour le traitement des sorties de RAP. Nous pouvons également remarquer que les mesures de ATENE fondées sur des modèles qui utilisent les POS donnent des prédictions particulièrement mauvaises. Nous pensons que ce comportement est causé par le fonctionnement de l'analyseur morphosyntaxique qui a tendance à identifier tous les mots contenant des majuscules comme des noms propres et à accentuer, par la suite, l'importance des majuscules par rapport aux autres traits. Ce trait optionnel est donc à utiliser avec précaution surtout que nos résultats montrent qu'il ne s'agit pas d'un trait indispensable.

Les valeurs de corrélation assez importantes obtenues pour le système REN-3 qui est fondé sur des approches symboliques confirment nos observations faites sur les données ETAPE et montrent que ATENE reste une mesure valable quel que soit le type d'approche utilisé pour le développement des systèmes de REN.

5.7.2 Évaluation de l'impact des sorties de RAP sur les erreurs commises par les systèmes de REN

Nous avons montré dans la section 5.3.4 qu'en plus de la dégradation des performances globales des systèmes de REN, les erreurs de RAP peuvent aussi favoriser soit une baisse du rappel (en augmentant le risque d'omission d'entités), soit

une baisse de la précision (en augmentant le risque d'insertion d'entités). C'est pour cette raison que nous souhaitons tester également la capacité des différentes mesures à prédire ces types d'impacts en plus de la prédiction de la dégradation des performances globales.

5.7.2.1 Évaluation de l'impact des sorties de RAP sur les erreurs d'omission et de substitution d'entités

C'est la mesure élémentaire $ATENE_{DS}$ qui permet d'évaluer le risque d'erreurs d'omission et de substitution dans notre méthode d'évaluation. En appliquant la même méthodologie de comparaison que celle décrite dans la section 5.6.3, nous mesurons cette fois la corrélation des rangs entres les classements des systèmes de RAP fondés sur leur taux d'erreur d'omission et de substitution et les classements fondés sur les mesures d'évaluations $ATENE_{DS}$, WER, NE-WER, WIL et rappel (rappel fondé sur les erreurs de RAP). Nous avons choisi d'utiliser le rappel plutôt que la F-mesure et la précision puisque il permet de montrer la relation entre les erreurs d'omission de mots faites par les systèmes de RAP et les erreurs d'omission d'entités faites par les systèmes de REN.

Nous mesurons le taux d'erreur d'omission et de substitution d'entités, et calculons ETER sans prendre en compte les erreurs d'insertion d'entités. Nous notons cette mesure $ETER_{DS}$ qui est donnée par l'équation 5.13 (voir chapitre 4 pour plus de détail).

$$ETER_{DS} = \frac{D + \sum_{(e_r,e_h)} E(e_r, e_h)}{N_E} \tag{5.13}$$

Nous rappelons que :

- D est le nombre d'entités supprimées ou omises ;
- $\sum_{(e_r,e_h)} E(e_r, e_h)$ est la somme des erreurs de substitution et de décomposition calculée pour les entités détectées par les systèmes de REN ;
- N_E est le nombre d'entités dans la référence.

Le tableau 5.5 affiche les corrélations de Spearman moyennes entre les classements des systèmes de RAP selon les taux d'erreur d'omission et de substitution des systèmes de REN mesurés en $ETER_{DS}$ et les classements des systèmes de RAP fondés sur les différentes mesures d'évaluation. Pour les mesures de $ATENE_{DS}$, nous affichons uniquement les résultats pour les modèles qui n'utilisent pas les majuscules, les autres modèles affichent des taux de corrélation légèrement moins bons comme le laissent penser les résultats présentés pour les performances globales pour les raisons que nous avons discutées précédemment. Nous pouvons voire dans le tableau 5.5 que les mesures de $ATENE_{DS}$ obtiennent des corrélations légèrement meilleures que celles des autres mesures. Ce résultat renforce notre méthodologie d'évaluation et confirme que notre mesure élémentaire, $ATENE_{DS}$, permet effectivement d'évaluer le risque d'erreurs d'omission et de substitution

5.5.7 Comparaison des mesures d'évaluation de la qualité des transcriptions automatiques pour la REN

		REN-1	REN-2	REN-3	REN-4	REN-5	REN-6	REN-7	moyenne
	WER	0,47	0,66	**0,70**	0,70	0,59	0,76	0,62	0,64
	NE-WER	**0,69**	0,69	0,43	0,51	0,47	0,67	0,67	0,59
	WIL	0,39	0,61	0,69	0,67	**0,60**	0,77	0,58	0,62
	Rappel	0,36	0,56	0,65	0,65	**0,60**	0,74	0,56	0,59
$ATENE_{DS}$	*Baseline*	0,61	0,75	0,55	0,72	0,58	0,73	0,65	0,66
	Base+préf.+suf.	**0,69**	0,75	0,50	0,66	0,57	0,75	**0,70**	0,66
	Base+pos	0,60	**0,77**	0,61	**0,76**	**0,60**	**0,82**	0,69	**0,69**
	Base+préf.+suf.+pos	0,68	0,75	0,48	0,66	0,56	0,74	0,69	0,65

TABLEAU 5.5 – Corrélations de Spearman moyennes entre les classements fondés sur les taux d'erreur d'omission et de substitution des systèmes de REN mesurés en $ETER_{DS}$ sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, résultats de la campagne ETAPE.

d'entités causé par les erreurs de RAP.

		REN-1	REN-2	REN-3	moyenne
	WER	0,72	0,52	**0,77**	0,67
	NE-WER	**0,80**	0,58	**0,77**	**0,72**
	WIL	0,72	0,52	**0,77**	0,67
	Rappel	0,72	0,52	**0,77**	0,67
$ATENE_{DS}$	*Baseline+maj.*	0,36	0,47	0,47	0,43
	Base+préf.+suf.+maj.	0,66	0,55	0,63	0,62
	Base+pos+maj.	0,66	0,61	0,69	0,65
	Base+préf.+suf.+pos+maj.	0,69	**0,63**	0,69	0,67

TABLEAU 5.6 – Corrélations de Spearman moyennes entre les classements fondés sur les taux d'erreur d'omission et de substitution des systèmes de REN mesurés en $ETER_{DS}$ sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, résultats de la campagne QUAERO.

La même expérience appliquée sur les données de la campagne QUAERO donne les résultats affichés dans le tableau 5.6. Nous pouvons voir dans ce tableau que les mesures de $ATENE_{DS}$ obtiennent des taux de corrélation similaires à ceux obtenus sur les données de la campagne ETAPE. Ces résultats confirment la stabilité de notre mesure élémentaire $ATENE_{DS}$. Toutefois, nous pouvons remarquer que, cette fois, les autres mesures d'évaluation et notamment NE-WER sont

celles qui offrent les meilleures prédictions pour le risque d'erreur d'omission et de substitution. Ces résultats suggèrent que NE-WER, qui est plus facile à calculer que $ATENE_{DS}$, reste un choix possible si l'évaluation vise à prévoir uniquement contre la baisse de rappel des systèmes de REN.

5.7.2.2 Évaluation de l'impact des sorties de RAP sur les erreurs d'insertion d'entités

C'est la mesure élémentaire $ATENE_I$ qui permet d'évaluer le risque d'erreur d'insertion d'entités dans notre méthode d'évaluation. En appliquant la même méthodologie de comparaison décrite dans la section 5.6.3 nous mesurons cette fois la corrélation des rangs entre les classements des systèmes de RAP fondés sur leur taux d'erreur d'insertion et les classements fondés sur les mesures d'évaluation $ATENE_I$, WER, NE-WER, WIL et P (précision fondée sur les erreurs de RAP).

Nous avons choisi d'utiliser la précision plutôt que la F-mesure et le rappel puisqu'il permet de montrer la relation entre les erreurs d'insertion de mots faites par les systèmes de RAP et les erreurs d'insertion d'entités faites par les systèmes de REN.

Nous mesurons le taux d'erreur d'insertion d'entités selon l'équation 5.14 que nous notons $ETER_I$ (voir chapitre 4 pour plus de détails) :

$$ETER_I = \frac{I}{N_E} \tag{5.14}$$

Avec I le nombre d'entités insérées (fausse alarmes) par le système de REN.

		REN-1	REN-2	REN-3	REN-4	REN-5	REN-6	REN-7	moyenne
WER		0,34	0,38	-0,16	0,34	0,27	0,21	0,35	0,25
NE-WER		0,47	0,48	0,22	0,53	0,56	0,46	0,43	0,45
WIL		0,28	0,28	-0,24	0,25	0,21	0,12	0,26	0,16
Précision		0,30	0,32	-0,20	0,29	0,22	0,17	0,31	0,20
$ATENE_I$	*Baseline*	0,71	**0,76**	0,32	0,83	0,75	0,69	0,72	0,68
	Base+préf.+suf.	**0,75**	**0,76**	0,28	**0,85**	0,75	**0,76**	**0,74**	**0,70**
	Base+pos	0,71	0,70	0,29	0,81	**0,77**	0,71	0,69	0,67
	Base+préf.+suf.+pos	0,72	0,74	**0,35**	0,82	0,74	**0,76**	0,70	0,69

TABLEAU 5.7 – Corrélations de Spearman moyennes entre les classements fondés sur les taux d'erreur d'insertion des systèmes de REN mesurés en $ETER_I$ sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, résultats de la campagne ETAPE.

5.5.7 Comparaison des mesures d'évaluation de la qualité des transcriptions automatiques pour la REN

Le tableau 5.7 affiche les corrélations de Spearman moyennes entre les classements des systèmes de RAP selon les taux d'erreur d'insertion des systèmes de REN mesurés en $ETER_I$ et les classements des systèmes de RAP fondés sur les différentes mesures d'évaluation. Nous pouvons voir dans ce tableau que les mesures de $ATENE_I$ obtiennent des taux de corrélation élevés. Ceci valide la méthodologie mise en place pour calculer cette mesure élémentaire. Les corrélations assez faibles (par rapport à $ATENE_I$) obtenues par les autres mesures montrent que ces mesures ne permettent pas d'évaluer un risque de baisse de précision des systèmes de REN engendrée par les erreurs de RAP, notamment la mesure P qui met en évidence les erreurs d'insertion de mots par les systèmes de RAP. La même expérience appliquée sur les données de la campagne QUAERO donne les résultats affichés dans le tableau 5.8. Ces résultats confirment les observations que nous avons faites sur les données de l'évaluation ETAPE. Toutefois, nous pouvons remarquer que les corrélations sont moins élevées que celles obtenues sur les données de la campagne ETAPE pour toutes les mesures d'évaluation. Ceci est peut-être dû au fait que nous disposons de moins de données pour la campagne QUAERO, seulement trois systèmes de RAP et trois systèmes de REN, alors que nous disposons du double en termes de nombre de systèmes pour la campagne ETAPE. Mais, malgré cette baisse de corrélation globale, $ATENE_I$ conserve un écart important en termes de corrélation avec les autres mesures.

		REN-1	REN-2	REN-3	moyenne
WER		0,08	-0,13	-0,36	-0,13
NE-WER		-0,19	-0,41	-0,52	-0,38
WIL		0,08	-0,13	-0,36	-0,13
Précision		0,08	-0,13	-0,36	-0,13
$ATENE_I$	*Baseline+maj.*	0,38	0,44	0,27	0,37
	Base+préf.+suf.+maj.	**0,50**	**0,55**	0,22	0,42
	Base+pos+maj.	0,38	0,41	**0,50**	**0,43**
	Base+préf.+suf.+pos+maj.	**0,50**	**0,55**	0,22	0,42

TABLEAU 5.8 – Corrélations de Spearman moyennes entre les classements fondés sur les taux d'erreur d'insertion des systèmes de REN mesurés en $ETER_I$ sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, résultats de la campagne QUAERO.

Ceci suggère que c'est cette mesure élémentaire qui donne un avantage considérable à notre mesure globale ATENE par rapport aux autres mesures, pour l'évaluation de la qualité des systèmes de RAP pour la REN. Ceci fait de ATENE une mesure complète qui permet de prendre en compte tous les types d'erreurs de REN pouvant être engendrés par les systèmes de RAP contrairement aux autres

mesures qui, comme le montrent les résultats, ne permettent de mesurer avec précision que le risque d'omission et de substitution d'entités.

5.8 Discussion

L'étude de l'état de l'art de l'évaluation des systèmes de RAP montre qu'il y a un besoin en mesures d'évaluation permettant d'évaluer les systèmes de RAP en prenant en compte le cadre applicatif dans lequel ils seront utilisés. Ce besoin a été renforcé suite à l'observation de nombreux cas d'incohérences entre les mesures données par le WER et les performances obtenues au niveau des applications globales. Ainsi, des mesures alternatives au WER ont été proposées dans la littérature. Nous distinguons particulièrement les mesures RIL et WIL qui sont fondées sur la mesure de perte d'information causée par les erreurs de RAP, et le NE-WER qui vise à calculer un taux d'erreur ciblé sur les passages de textes importants pour l'application visée. Toutefois, ces métriques n'ont pas été très utilisées et n'ont pas été testées sur des données réelles pour prouver leur efficacité.

Afin d'aborder cette problématique, nous nous sommes intéressé au cadre applicatif de la REN à partir de la parole. Nous avons donc commencé par vérifier si le WER permettait effectivement d'évaluer la qualité des transcriptions automatiques pour la REN en utilisant les résultats des campagnes d'évaluation ETAPE et QUAERO qui incluent toutes les deux la même tâche de REN à partir de sorties de RAP. Nous avons ainsi montré que les mesures données par le WER ne permettent pas une bonne estimation de la qualité des systèmes de RAP pour la REN, puisque le classement des systèmes de RAP suivant le WER ne reflète pas celui obtenu suivant les performances des systèmes de REN sur les résultats des deux campagnes d'évaluation. Nous avons également étudié l'impact des erreurs de RAP sur la REN et nous avons montré que certaines sorties de RAP favorisent les erreurs d'omission d'entités alors que d'autres favorisent les erreurs d'insertion.

En nous appuyant sur les résultats de nos analyses et sur notre étude de l'état de l'art, nous avons défini une nouvelle mesure ATENE qui permet d'évaluer la qualité des systèmes de RAP et l'impact de leurs erreurs pour des systèmes de REN appliqués en aval. Notre mesure vise à mesurer la perte d'information causée par les erreurs de RAP, en ciblant les passages de textes importants pour les systèmes de RAP. Elle est fondée sur l'utilisation de modèles statistiques permettant de prendre en compte les relations de dépendance qui relient les différentes informations contextuelles nécessaires pour la REN.

ATENE offre la possibilité de comparer les probabilités de la présence d'entités sur les transcriptions de référence et d'hypothèses plutôt que de faire une comparaison directe des graphèmes. Elle est composée de deux mesures élémentaires :

- $ATENE_{DS}$ qui permet l'évaluation des risques d'erreurs d'omission et de substitution d'entités ;
- $ATENE_I$ qui permet d'évaluer le risque d'erreur d'insertion d'entités causé par les erreurs de RAP.

Pour valider notre mesure, nous avons mis en place une expérience qui permet de la comparer avec les autres mesures proposées dans la littérature sur la base de données réelles tirées de campagnes d'évaluation. Il s'agit de comparer la corrélation des classements des systèmes de RAP suivant les mesures d'évaluation et les classements fondés sur les performances des systèmes de REN. Les résultats montrent que ce sont les mesures de ATENE qui permettent le mieux d'évaluer la qualité des systèmes de RAP pour la REN sur les données des deux campagnes d'évaluation. En revanche, les résultats observés pour les mesures tirées de la littérature (que nous avons testées) ne sont pas stables. Alors que les mesures de NE-WER obtiennent des meilleures corrélations que WER, WIL et F-mesure sur les données ETAPE, ce sont ces dernières (WER, WIL et F-mesure) qui obtiennent les meilleures corrélations sur les données QUAERO. Le résultat montre également que le WIL et la F-mesure ne sont pas de meilleures mesures que le WER dans le contexte applicatif étudié.

Notre méthodologie de validation inclut également des mesures de corrélation entre les classements des systèmes de RAP donnés par les mesures élémentaires $ATENE_{DS}$ et $ATENE_I$ et les classements donnés respectivement par $ETER_{DS}$ (taux d'erreur d'omission et de substitution d'entités) et $ETER_I$ (taux d'erreur d'insertion d'entités). Les corrélations assez élevées obtenues pour les deux mesures élémentaires confirment qu'elles ont les comportements voulus et valident ainsi notre approche.

L'utilisation d'une approche statistique dans notre méthodologie peut laisser craindre une plus mauvaise corrélation avec les systèmes de REN symboliques. L'obtention de valeurs de corrélation élevées avec tous les systèmes de REN ayant participé aux évaluations et, en particulier, les systèmes fondés sur des approches symboliques (REN-6 pour ETAPE et REN-3 pour QUAERO), montre que notre méthodologie reste valable quelle que soit l'approche utilisée pour le développement des systèmes de REN.

L'utilisation d'une approche statistique fait du choix des traits un paramètre important de notre mesure. Nous avons choisi d'utiliser des traits très basiques, à fois pour simplifier le développement et aussi parce que ces traits simples sont communs à la plupart des approches en REN. Nous avons testé différentes combinaisons de traits pour l'entraînement de nos modèles dans le but d'évaluer l'impact de ce paramètre sur le fonctionnement de ATENE. Les résultats montrent qu'il y a très peu de variations causées par le changement des traits parmi les com-

binaisons que nous avons choisies. Même le modèle de base fondé uniquement sur les unigrammes et les bigrammes de mots permet de meilleures estimations que les mesures de l'état de l'art. Les traits morphologiques (préfixes et suffixes) semblent aider à améliorer nos mesures, contrairement aux parties du discours qui ne semblent pas apporter beaucoup d'amélioration. Toutefois, ce trait (POS) reste dépendant du fonctionnement de l'analyseur morphosyntaxique utilisé, et il est difficile d'en tirer des conclusions fortes. Finalement, la capitalisation reste un trait important, mais à utiliser uniquement si les majuscules sont fournies dans la plupart des sorties de celles de RAP à évaluer.

Les résultats des comparaisons que nous avons effectuées pour les mesures élémentaires $ATENE_{DS}$ et $ATENE_I$ montrent que le point fort de notre mesure réside dont sa capacité à évaluer le risque d'erreur d'insertion d'entités qui n'est pas bien assuré dans le cas des autres mesures. Ceci fait de ATENE une mesure complète qui permet de prendre en compte tous les types d'erreurs de REN pouvant être engendrés par les systèmes de RAP. En revanche, les corrélations obtenues pour les mesures de $ATENE_{DS}$ ne sont pas toujours meilleures que celles obtenues pour les autres mesures (WER, NE-WER, WIL et R). Par conséquent, il est peut être possible d'envisager de combiner $ATENE_I$ avec d'autres mesures. Nous pensons notamment à NE-WER et rappel qui dans leur logique sont proches de $ATENE_{DS}$ et plus simples à calculer.

Toutefois, l'utilisation de ATENE reste conditionnée par la disposition de données de test à la fois transcrites manuellement et annotées en entités nommées. Nous discutons des perspectives de solutions possibles à explorer pour réduire ou se séparer de cette dépendance dans la section 5.9.

5.9 Conclusion

Dans ce chapitre, nous avons utilisé les données des campagnes d'évaluation ETAPE et QUAERO pour observer les performances des systèmes de REN en fonction du WER et nous avons montré que les mesures données par le WER ne permettaient pas une bonne estimation de la qualité des systèmes de RAP pour la REN puisque le classement des systèmes de RAP suivant le WER ne reflète pas celui obtenu suivant les performances des systèmes de REN sur les résultats des deux campagnes d'évaluation. Nous avons également étudié l'impact des erreurs de RAP sur la REN et nous avons montré que certaines sorties de RAP favorisaient les erreurs d'omission d'entités alors que d'autres favorisaient les erreurs d'insertion.

En se fondant sur les résultats de nos analyses nous avons défini une nouvelle mesure ATENE pour l'évaluation de la qualité des systèmes de RAP et leur impact sur des systèmes de REN appliqués en aval. ATENE permet de comparer les pro-

babilités de présence d'entités sur les transcriptions de référence et d'hypothèses plutôt que de faire une comparaison directe des graphèmes. Elle est composée de deux mesures élémentaires, $ATENE_{DS}$ qui permet l'évaluation des risques d'erreur d'omission et de substitution d'entités et $ATENE_I$ qui permet d'évaluer le risque d'erreur d'insertion d'entités causé par les erreurs de RAP.

Afin de valider notre approche, nous avons mis en place une procédure de comparaison avec les autres mesures de la littérature (WER, NE-WER, WIL, F-mesure). La procédure de validation consiste à mesurer la corrélation entre les classements des systèmes de RAP fondés sur les différentes mesures d'évaluation et ceux fondés sur les performances des systèmes de REN. Les résultats obtenus montrent que ce sont les mesures de ATENE (et ses mesures élémentaires) qui obtiennent les valeurs de corrélation les plus élevées. Ces résultats permettent de valider notre approche.

Conclusion et perspectives

Les travaux présentés dans ce mémoire portent sur l'évaluation des technologies issues de la recherche, plus précisément du TAL (traitement automatique des langues). Notre travail s'est concentré sur deux technologies et sur leur enchaînement : la reconnaissance automatique de la parole (RAP) et la reconnaissance d'entités nommées (REN). La REN est la brique élémentaire qui permet d'extraire des informations d'un flux textuel. La RAP est une brique technologique importante qui permet, quant à elle, d'accéder à une grande quantité de données se trouvant sous format audio et audiovisuel. Cette brique est intégrée dans de nombreuses applications de traitement automatique de la parole visant à accomplir des tâches complexes, telles que l'extraction d'informations à partir de la parole, la compréhension de la parole et la traduction de la parole. Malgré les progrès atteints, les systèmes de RAP ne sont pas parfaits et font encore des erreurs de transcription, notamment à cause des variations qui peuvent toucher le signal de la parole et à cause des ambiguïtés intrinsèques à la langue. Une étape d'évaluation est donc nécessaire pour s'assurer de leur efficacité avant la phase d'intégration.
Nous présentons dans la section suivante un bilan de nos travaux qui ont porté sur la proposition de deux métriques. La première métrique, ETER, est conçue pour l'évaluation de la tâche de détection, classification et décomposition d'entités, et la seconde, ATENE, est conçue pour l'évaluation des systèmes de RAP pour la reconnaissance d'entités nommées. Enfin, nous décrivons dans la dernière section quelques perspectives de recherche ouvertes à l'issue de ce travail.

Conclusion

Dans cette thèse nous nous sommes intéressés au cadre applicatif de la reconnaissance d'entités nommées à partir de la parole, notre choix était motivé par la popularité de la tâche et par la diversité des cas d'utilisation possibles. Nous avons commencé par présenter la tâche de détection, classification et décomposition d'entités que nous avons considérée. Cette tâche qui requiert deux niveaux d'annotations a donné lieu à des structures d'annotations complexes. Évaluer cette tâche en utilisant des métriques telles que le SER (*Slot Error Rate*) et la F-mesure développées initialement pour l'évaluation de tâches simples, biaise les mesures

de performance, principalement à cause de la non-prise en compte de la structure des annotations et des relations qui relient les composants aux types. Étant donné que la tâche de REN, que nous étudions dans le cadre de cette thèse, est celle de la détection, classification et décomposition d'entités, nous avons développé une procédure d'évaluation qui permet de remédier aux limites observées pour les autres métriques et de prendre en compte les caractéristiques de la tâche. Cette procédure inclut une nouvelle approche d'alignement ainsi qu'une nouvelle métrique d'évaluation ETER. ETER permet à l'utilisateur de fixer le poids qu'il souhaite accorder à chacune des deux sous-tâches (la classification et la décomposition des entités) suivant le but expérimental ou/et le cadre applicatif visés. Afin de valider notre proposition, nous avons comparé le comportement de ETER à celui de SER en utilisant une sélection d'exemples et en nous appuyant sur des données réelles issues de la campagne d'évaluation ETAPE. Nos expériences ont montré que ETER était la métrique la plus adaptée pour l'évaluation de la tâche de détection, classification et décomposition d'entités étudiée. Nous l'avons donc adoptée pour mesurer les performances des systèmes de REN durant notre étude.

Pour aborder la problématique de l'évaluation des systèmes de RAP, nous avons commencé par dresser l'historique des défis proposés dans les campagnes d'évaluation. Ceci nous a permis de nous familiariser avec les procédures d'évaluation et de mettre en évidence l'évolution des buts applicatifs visés selon le degré de maturité de la technologie.
Le WER est la métrique classiquement utilisée pour évaluer la qualité des sorties de RAP. Cette métrique a prouvé son efficacité quand il s'agit d'évaluer les systèmes de RAP en isolation, elle a permis l'évaluation et l'optimisation des systèmes de RAP pendant de longues années. Toutefois, avec l'intégration des systèmes de RAP dans des chaînes de traitement où d'autres modules utilisent leur sortie comme entrée, de nombreuses études ont remarqué des incohérences entre les mesures données par le WER et les performances obtenues au niveau de l'application globale. Ainsi, des alternatives au WER ont été proposées dans la littérature. Nous distinguons particulièrement RIL (*Relative Information Loss*) et WIL (*Word Information Lost*) qui sont fondées sur la mesure de la perte d'informations causée par les erreurs de RAP et le NE-WER qui vise à calculer un taux d'erreurs ciblé sur les passages de textes importants pour l'application visée. Toutefois, ces métriques n'ont pas été très utilisées ni testées sur des données réelles pour prouver leur efficacité.

Afin d'aborder la problématique de l'évaluation des systèmes de RAP pour la REN, nous avons commencé par vérifier si le WER permettait effectivement d'évaluer la qualité des transcriptions automatiques pour la REN en utilisant les résultats des campagnes d'évaluation ETAPE et QUAERO qui incluent toutes les deux la même tâche de REN à partir de sorties de RAP. Nous avons ainsi montré que les mesures données par le WER ne permettent pas une bonne estimation de la qua-

lité des systèmes de RAP pour la REN puisque le classement des systèmes de RAP suivant le WER ne reflète pas celui obtenu suivant les performances des systèmes de REN sur les résultats des deux campagnes d'évaluation. Nous avons également étudié l'impact des erreurs de RAP sur la REN et nous avons montré que certaines sorties de RAP favorisent les erreurs d'omission d'entités alors que d'autres favorisent les erreurs d'insertion.

En nous appuyant sur les résultats de nos analyses et de notre étude de l'état de l'art, nous avons défini une nouvelle mesure ATENE qui permet d'évaluer la qualité des systèmes de RAP et l'impact de leurs erreurs pour des systèmes de REN appliqués en aval. Notre mesure vise à évaluer la perte d'informations causée par les erreurs de RAP en ciblant les passages de textes importants pour les systèmes de RAP. Elle est fondée sur l'utilisation de modèles statistiques permettant de prendre en compte les relations de dépendance qui relient les différentes informations contextuelles nécessaires pour la REN.
ATENE consiste à comparer les probabilités de présence d'entités sur les transcriptions de référence et d'hypothèse plutôt qu'une comparaison directe des graphèmes. Elle est composée de deux mesures élémentaires :

- $ATENE_{DS}$ qui permet l'évaluation de risque d'erreur d'omission et de substitution d'entités ;
- $ATENE_I$ qui permet d'évaluer le risque d'erreur d'insertion d'entités causé par les erreurs de RAP.

Pour la validation de notre mesure, nous avons mis en place une expérience qui permet de la comparer avec les autres mesures proposées dans la littérature sur la base de données réelles tirées de campagnes d'évaluation. Il s'agit de comparer la corrélation des classements des systèmes de RAP suivant les mesures d'évaluation et les classements fondés sur les performances des systèmes de REN. Les résultats montrent que ce sont les mesures de ATENE qui permettent le mieux d'évaluer la qualité des systèmes de RAP pour la REN sur les données des deux campagnes d'évaluation. En revanche, les résultats observés pour les mesures tirées de la littérature (que nous avons testées) ne sont pas stables. Alors que les mesures de NE-WER obtiennent des meilleures corrélations que WER, WIL et F-mesure sur les données ETAPE, ce sont ces derniers (WER, WIL et F-mesure) qui obtiennent les meilleures corrélations sur les données QUAERO. Le résultat montre également que le WIL et la F-mesure ne sont pas de meilleures mesures que le WER dans le contexte applicatif étudié.

Notre méthodologie de validation inclut également des mesures de corrélation entre les classements des systèmes de RAP donnés par les mesures élémentaires $ATENE_{DS}$ et $ATENE_I$ et les classements données respectivement par $ETER_{DS}$ (taux d'erreur d'omission et de substitution d'entités) et $ETER_I$ (taux d'erreur d'insertion d'entités). Les corrélations assez élevées obtenues pour les deux me-

sures élémentaires confirment qu'elles ont les comportements voulus et valident ainsi notre approche.

Perspectives

Nous donnons ici quelques pistes pour aller plus loin dans la continuité de ces travaux ou sur des problématiques connexes liées à l'évaluation des systèmes de RAP en contextes applicatifs.

Optimisation des systèmes de RAP Le développement des systèmes de RAP nécessite une phase d'optimisation qui permet de fixer un certain nombre de paramètres, comme le *fudge factor* (poids du modèle acoustique par rapport au modèle de langage), le poids d'ajout d'un mot [Watanabe et Le Roux, 2014] et le nombre des gaussiennes dans les approches fondées sur les HMM. Le choix de la valeur optimale de ces paramètres est empirique et nécessite une phase d'optimisation. Classiquement, durant cette phase différentes valeurs des paramètres à fixer sont testées pour garder celles qui permettent d'obtenir le WER optimal.
Nous pensons que ATENE peut être utilisée pour optimiser les systèmes de RAP développés pour faire de la REN. Toutefois, ceci reste à prouver et à tester notamment pour répondre à certaines questions telles que la détermination des critères d'arrêt de l'optimisation.

Caractérisation des erreurs de RAP et de leur impact La caractérisation des erreurs de RAP est à la fois intéressante pour une meilleure compréhension des erreurs et de leur prise en compte pendant le développement des systèmes de RAP et de REN, et pour la simplification de la procédure de calcul de ATENE. L'étude des erreurs à laquelle nous nous intéressons vise à caractériser les erreurs de RAP selon leur gravité pour la REN, et permettre ainsi de mettre en place une taxonomie d'erreurs de RAP fondée sur leur impact pour la REN. L'obtention d'une telle taxonomie permettra aux développeurs des systèmes de RAP de reconnaître les erreurs qu'il faut éviter et aux développeurs des systèmes de REN de mieux comprendre les causes de la baisse des performances due aux erreurs de RAP afin d'augmenter la robustesse de leurs systèmes face à ces erreurs. Une telle taxonomie permettra également de simplifier la procédure d'évaluation qui consistera alors à repérer, parmi les erreurs de RAP, celles qui sont graves pour la REN et à calculer, par la suite, un taux d'erreur ou de mesure d'une performance en se fondant sur le nombre de ces erreurs et leur gravité. Ainsi, une telle procédure d'évaluation ne nécessite ni l'existence d'une référence annotée en entités nommées ni de modèle statistique.

5.5.9 Conclusion

Pour mettre en place l'étude des erreurs de RAP, nous proposons d'examiner plus en détail les causes de la baisse de Δ_M (la différence entre la marge calculée sur les sorties de RAP et celle calculée sur la transcription de référence) lors du calcul de ATENE. Ceci est possible en examinant la variation des poids des traits servant à calculer les probabilités et par la suite Δ_M. Nous rappelons ici que, pour les modèles de maximum d'entropie, la probabilité pour un mot m d'appartenir à une classe c d'entités est donnée par l'équation 5.15 :

$$P(c|m) = \frac{exp(\sum_{n=1}^{n} \lambda_i \times f_i(t,c))}{Z(m)} \tag{5.15}$$

Avec $f_1..f_k$ des fonctions caractéristiques liées aux différents traits considérés et $\lambda_1..\lambda_k$ leurs poids associés et $Z(m)$ un facteur de normalisation sur le mot.

Ainsi, identifier les λ_i qui causent la baisse de Δ_M permettra d'identifier les erreurs ayant modifié le traits ou les traits qui engendrent l'augmentation du risque d'erreur de REN dans un empan d'erreurs de RAP. Ceci nous permettra de déterminer les types d'erreur de RAP les plus graves pour la REN et de mettre en place une taxonomie.

Généralisation pour d'autres contextes applicatifs Nous avons exploré et testé notre mesure d'évaluation pour la tâche de REN à partir de la parole, mais son utilisation pour l'évaluation des systèmes de RAP, dans d'autres contextes applicatifs similaires, reste possible. Nous pensons particulièrement à tout ce qui est recherche et extraction d'informations, à l'analyse sémantique des documents audio et aux systèmes de dialogue oraux. Toutefois, l'adaptation de ATENE à de nouvelles tâches reste conditionnée par la disposition de données d'apprentissage et de test nécessaires pour le calcul des scores. En revanche, l'adaptation de ATENE à des tâches telles que la traduction automatique de la parole pose de nombreuses questions et nous pensons que le WER reste dans ce cadre applicatif une mesure adaptée puisque les systèmes de traduction sont le plus souvent évalués en utilisant la mesure BLEU [Papineni *et al.*, 2002] qui consiste à comparer les n-grammes des mots de la traduction de référence aux traductions de l'hypothèse pour calculer un taux d'erreur sur les mots. Ceci fait du BLEU une mesure très proche du WER. En revanche, ATENE cherche à mesurer la perte d'information ciblée sur des structures de données importantes pour une tâche cible. Ces structures de données ne sont pas faciles à définire (identifier) dans le cas de la traduction. Toutefois, la traduction automatique reste une tâche qui dépend fortement des langues visées (langue source et langue cible) et il se peut que, pour un couple de langues données, certaines erreurs sur des particules, des articles ou des traits morphologiques ne soient pas graves pour la tâche de traduction. La prise en compte de ce type de caractéristique pendant la phase d'évaluation des systèmes de RAP pourrait permettre une meilleure estimation de la qualité des transcriptions automatiques pour la traduction automatique.

En revanche, nous pensons que ATENE peut être adaptée pour évaluer la qualité des traductions automatiques pour faire de la recherche d'informations ou de l'analyse sémantique si les corpus de référence de la langue cible sont annotés selon la tâche visée.

Bibliographie

[Adda, 2011] ADDA, G. (2011). *Approches empiriques et modélisation statistique de la parole*. Thèse de doctorat, Université Paris Sud-Paris XI.

[Allauzen et Bonneau-Maynard, 2008] ALLAUZEN, A. et BONNEAU-MAYNARD, H. (2008). Training and Evaluation of POS Taggers on the French MULTITAG Corpus. *In Proceedings of the sixth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA).

[Babych et Hartley, 2003] BABYCH, B. et HARTLEY, A. (2003). Improving machine translation quality with automatic named entity recognition. *In Proceedings of the 7th International EAMT workshop on MT and other Language Technology Tools, Improving MT through other Language Technology Tools : Resources and Tools for Building MT*, pages 1–8. Association for Computational Linguistics.

[Bazillon *et al.*, 2008] BAZILLON, T., JOUSSE, V., BÉCHET, F., ESTÈVE, Y., LINARÈS, G. et LUZZATI, D. (2008). La parole spontanée : transcription et traitement. *Revue TAL. Volume*, 49(3).

[Béchet et Charton, 2010] BÉCHET, F. et CHARTON, E. (2010). Unsupervised knowledge acquisition for extracting named entities from speech. *In Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, pages 5338–5341. IEEE.

[Bellegarda, 2014] BELLEGARDA, J. R. (2014). Spoken language understanding for natural interaction : The SIRI experience. *In Natural Interaction with Robots, Knowbots and Smartphones*, pages 3–14. Springer.

[Berger *et al.*, 1996] BERGER, A. L., PIETRA, V. J. D. et PIETRA, S. A. D. (1996). A maximum entropy approach to natural language processing. *Computational linguistics*, 22(1):39–71.

[Bikel *et al.*, 1997] BIKEL, D. M., MILLER, S., SCHWARTZ, R. et WEISCHEDEL, R. (1997). Nymble : a high-performance learning name-finder. *In Proceedings of the fifth conference on Applied natural language processing*, pages 194–201. Association for Computational Linguistics.

[Bodenreider et Zweigenbaum, 2000] BODENREIDER, O. et ZWEIGENBAUM, P. (2000). Identifying proper names in parallel medical terminologies. *Studies in health technology and informatics*, 77:443.

[Brun *et al.*, 2010] BRUN, C., EHRMANN, M. et MAUPERTUIS, C. (2010). Un système de détection d'entités nommées adapté pour la campagne d'évaluation ester 2. *In Actes de la 17ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'10)*.

[Brun et Hagege, 2004] BRUN, C. et HAGEGE, C. (2004). Intertwining deep syntactic processing and named entity detection. *In Advances in Natural Language Processing*, pages 195–206. Springer.

[Bruno, 2009] BRUNO, I. (2009). La recherche scientifique au crible du benchmarking. *Revue d'histoire moderne et contemporaine*, (5):28–45.

[Canavan *et al.*, 1997] CANAVAN, A., GRAFF, D. et ZIPPERLEN, G. (1997). Callhome american english speech. *Linguistic Data Consortium*.

[Chinchor et Sundheim, 1993] CHINCHOR, N. et SUNDHEIM, B. (1993). MUC-5 evaluation metrics. *In Proceedings of the 5th conference on Message understanding*, pages 69–78. Association for Computational Linguistics.

[Chinchor, 1998] CHINCHOR, N. A. (1998). Overview of MUC-7/MET-2.

[Chok, 2010] CHOK, N. S. (2010). *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data*. Thèse de doctorat, University of Pittsburgh.

[Chomsky, 2002] CHOMSKY, N. (2002). *On nature and language*. Cambridge University Press.

[Cunningham *et al.*, 1995] CUNNINGHAM, H., GAIZAUSKAS, R. J. et WILKS, Y. (1995). *A General Architecture for Text Engineering (GATE) : A New Approach to Language Engineering R&D*. Citeseer.

[Daille et Morin, 2000] DAILLE, B. et MORIN, E. (2000). Reconnaissance automatique des noms propres de la langue écrite : les récentes réalisations. *TAL. Traitement automatique des langues*, 41(3):601–621.

[Demir *et al.*, 2013] DEMIR, C., SARACLAR, M. et CEMGIL, A. (2013). Single-channel speech-music separation for robust asr with mixture models. *IEEE T-ASLP*, 21(4):725–736.

[Dinarelli et Rosset, 2012] DINARELLI, M. et ROSSET, S. (2012). Tree representations in probabilistic models for extended named entities detection. *In Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 174–184. Association for Computational Linguistics.

[Doddington *et al.*, 2004] DODDINGTON, G. R., MITCHELL, A., PRZYBOCKI, M. A., RAMSHAW, L. A., STRASSEL, S. et WEISCHEDEL, R. M. (2004). The Automatic Content Extraction (ACE) Program-Tasks, Data, and Evaluation. *In LREC*. Citeseer.

[Doddington et Schalk, 1981] DODDINGTON, G. R. et SCHALK, T. B. (1981). Computers : Speech recognition : Turning theory to practice : New ICs have brought the requisite computer power to speech technology. *Spectrum, IEEE*, 18(9):26–32.

[Dudley, 1939] DUDLEY, H. (1939). Remaking speech. *The Journal of the Acoustical Society of America*, 11(2):169–177.

[Ehrmann, 2008] EHRMANN, M. (2008). *Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation*. Thèse de doctorat, Paris 7.

[Favre *et al.*, 2013] FAVRE, B., CHEUNG, K., KAZEMIAN, S., LEE, A., LIU, Y., MUNTEANU, C., NENKOVA, A., OCHEI, D., PENN, G., TRATZ, S., VOSS, C. et ZELLER, F. (2013). Automatic Human Utility Evaluation of ASR Systems : Does WER Really Predict Performance ? *In Interspeech, Lyon (France)*.

[Ferro *et al.*, 2001] FERRO, L., MANI, I., SUNDHEIM, B. et WILSON, G. (2001). TIDES Temporal Annotation Guidelines-Version 1.0. 2. *The MITRE Corporation, McLean-VG-USA*.

[Fort *et al.*, 2009] FORT, K., EHRMANN, M. et NAZARENKO, A. (2009). Vers une méthodologie d'annotation des entités nommées en corpus ? *In Traitement Automatique des Langues Naturelles 2009*.

[Fourour, 2002] FOUROUR, N. (2002). Nemesis, un système de reconnaissance incrémentielle des entités nommées pour le français. *In Actes de la 9ème conférence sur le Traitement Automatique des Langues Naturelles (TALN'02)*, pages 265–274.

[Friburger, 2002] FRIBURGER, N. (2002). *Reconnaissance automatique des noms propres : application à la classification automatique de textes journalistiques*. Thèse de doctorat, Tours.

[Fukuda *et al.*, 1998] FUKUDA, K.-i., TSUNODA, T., TAMURA, A., TAKAGI, T. *et al.* (1998). Toward information extraction : identifying protein names from biological papers. *In Pac Symp Biocomput*, volume 707, pages 707–718. Citeseer.

[Galibert *et al.*, 2014] GALIBERT, O., LEIXA, J., ADDA, G., CHOUKRI, K. et GRAVIER, G. (2014). The ETAPE speech processing evaluation. *In Proc of LREC*, Reykjavik, Iceland. ELRA.

[Galibert *et al.*, 2010] GALIBERT, O., QUINTARD, L., ROSSET, S., ZWEIGENBAUM, P., NÉDELLEC, C., AUBIN, S., GILLARD, L., RAYSZ, J.-P., POIS, D., TANNIER, X., DELÉGER, L. et LAURENT, D. (2010). Named and Specific Entity Detection in Varied Data : The QUAERO Named Entity Baseline Evaluation. *In Proc of LREC*, Valletta, Malta. ELRA.

[Galibert *et al.*, 2011] GALIBERT, O., ROSSET, S., GROUIN, C., ZWEIGENBAUM, P. et QUINTARD, L. (2011). Structured and Extended Named Entity Evaluation in Automatic Speech Transcriptions. *In Proc of IJCNLP*, Chiang Mai, Thailand.

[Galliano *et al.*, 2006] GALLIANO, S., GEOFFROIS, E., GRAVIER, G., BONASTRE, J.-F., MOSTEFA, D. et CHOUKRI, K. (2006). Corpus description of the ESTER evaluation campaign for the rich transcription of french broadcast news. *In Proceedings of LREC*, volume 6, pages 315–320.

[Galliano *et al.*, 2009a] GALLIANO, S., GRAVIER, G. et CHAUBARD, L. (2009a). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. *In Proc of Interspeech 2009*.

[Galliano *et al.*, 2009b] GALLIANO, S., GRAVIER, G. et CHAUBARD, L. (2009b). The ESTER 2 evaluation campaign for the rich transcription of French radio broadcasts. *In Interspeech*, volume 9, pages 2583–2586.

[Garofolo *et al.*, 2000] GAROFOLO, J. S., AUZANNE, C. G. et VOORHEES, E. M. (2000). The TREC Spoken Document Retrieval Track : A Success Story. *NIST SPECIAL PUBLICATION SP*, 500(246):107–130.

[Garofolo *et al.*, 1999] GAROFOLO, J. S., VOORHEES, E. M., AUZANNE, C. G., STANFORD, V. M. et LUND, B. A. (1999). 1998 TREC-7 spoken document retrieval track overview and results. *In Broadcast News Workshop'99 Proceedings*, page 215. Morgan Kaufmann Pub.

[Godfrey *et al.*, 1992] GODFREY, J. J., HOLLIMAN, E. C. et MCDANIEL, J. (1992). SWITCHBOARD : Telephone speech corpus for research and development. *In Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520. IEEE.

[Gravier *et al.*, 2012] GRAVIER, G., ADDA, G., PAULSSON, N., CARRÉ, M., GIRAUDEL, A. et GALIBERT, O. (2012). The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. *In* CHAIR), N. C. C., CHOUKRI, K., DECLERCK, T., ur DO AN, M. U., MAEGAARD, B., MARIANI, J., ODIJK, J. et PIPERIDIS, S., éditeurs : *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

[Grishman et Sundheim, 1996] GRISHMAN, R. et SUNDHEIM, B. (1996). Message Understanding Conference - 6 : A Brief History. *In Proc. of COLING*, pages 466–471.

[Grouin, 2013] GROUIN, C. (2013). *Anonymisation de documents cliniques : performances et limites des méthodes symboliques et par apprentissage statistique*. Thèse de doctorat, Université Pierre et Marie Curie-Paris VI.

[Grouin *et al.*, 2011] GROUIN, C., GALIBERT, O., ROSSET, S., QUINTARD, L. et ZWEIGENBAUM, P. (2011). Mesures d'évaluation pour entités nommées structurées. *Évaluation des méthodes d'Extraction de Connaissances dans les Données, Brest, France*.

[Hatmi *et al.*, 2013] HATMI, M., JACQUIN, C., MORIN, E., MEIGNER, S. *et al.* (2013). Named Entity Recognition in Speech Transcripts following an Extended Taxonomy. *In Workshop on Speech, Language and Audio in Multimedia (SLAM)*.

[Hirschman, 1998] HIRSCHMAN, L. (1998). Language understanding evaluations : lessons learned from MUC and ATIS. *In Proceedings of the first international conference on language resources and evaluation (LREC)*, pages 117–122.

[Isozaki, 2001] ISOZAKI, H. (2001). Japanese named entity recognition based on a simple rule generator and decision tree learning. *In Proceedings of the 39th Annual Meeting on Association for Computational Linguistics*, pages 314–321. Association for Computational Linguistics.

[Isozaki et Kazawa, 2002] ISOZAKI, H. et KAZAWA, H. (2002). Efficient support vector classifiers for named entity recognition. *In Proceedings of the 19th international conference on Computational linguistics-Volume 1*, pages 1–7. Association for Computational Linguistics.

[Ji-Hwan Kim, 2000] JI-HWAN KIM, P. W. (2000). A rule-based named entity recognition system for speech input. *In The Proceedings of the 6˜(th) International Conference on Spoken Language Processing (Volume I)*.

[Jones et Galliers, 1996] JONES, K. S. et GALLIERS, J. R. (1996). *Evaluating natural language processing systems : An analysis and review*, volume 1083. Springer Science & Business Media.

[Karanasou et Lamel, 2010] KARANASOU, P. et LAMEL, L. (2010). Comparing SMT methods for automatic generation of pronunciation variants. *In Advances in Natural Language Processing*, pages 167–178. Springer.

[Kendall, 1948] KENDALL, M. G. (1948). Rank correlation methods.

[Ketabdar *et al.*, 2007] KETABDAR, H., HANNEMANN, M. et HERMANSKY, H. (2007). Detection of out-of-vocabulary words in posterior based ASR. *In Interspeech*, pages 1757–1760.

[King, 1984] KING, M. (1984). When is the next Alpac report due ? *In Proceedings of the 10th international conference on Computational linguistics*, pages 352–353. Association for Computational Linguistics.

[Lavergne *et al.*, 2010] LAVERGNE, T., CAPPÉ, O. et YVON, F. (2010). Practical Very Large Scale CRFs. *In Proceedings the 48th Annual Meeting of the Association for Computational Linguistics*, pages 504–513.

[Le Meur *et al.*, 2004] LE MEUR, C., GALLIANO, S. et GEOFFROIS, E. (2004). Conventions d'annotations en Entités nommées-ESTER. *Rapport technique de la campagne Ester*.

[Levenshtein, 1966] LEVENSHTEIN, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8):707–710.

[Li *et al.*, 2014] LI, J., DENG, L., GONG, Y. et HAEB-UMBACH, R. (2014). An overview of noise-robust automatic speech recognition. *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, 22(4):745–777.

[Liberman, 2010] LIBERMAN, M. (2010). The future of computational linguistics : or, what would Antonio Zampolli do. *Antonio Zampolli Prize speech, presented at LREC2010*.

[Luzzati *et al.*, 2014] LUZZATI, D., GROUIN, C., VASILESCU, I., ADDA-DECKER, M., BILINSKI, E., CAMELIN, N., KAHN, J., LAILLER, C., LAMEL, L. et ROSSET, S. (2014). Human annotation of ASR error regions : Is "gravity" a sharable concept for human annotators ? *In International Conference on Language Resources and Evaluation*.

[M., 2003] M., S. (2003). Word Sense Disambiguation. The case for Combinations of Knowledge Sources.

[Maier, 2002] MAIER, V. (2002). Evaluating RIL as basis of automatic speech recognition devices and the consequences of using probabilistic string edit distance as input. *Univ. of Sheffield, third year project*.

[Makhoul *et al.*, 1999] MAKHOUL, J., KUBALA, F., SCHWARTZ, R. et WEISCHEDEL, R. (1999). Performance Measures For Information Extraction. *In Proc. of DARPA Broadcast News Workshop*, pages 249–252.

[Makino *et al.*, 2007] MAKINO, S., SAWADA, H. et ARAKI, S. (2007). Frequency-domain blind source separation. *In Blind Speech Separation*, pages 47–78. Springer.

[Maurel *et al.*, 2011] MAUREL, D., FRIBURGER, N., ANTOINE, J.-Y., ESHKOL, I. et NOUVEL, D. (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Traitement automatique des langues*, 52(1):69–96.

[McCowan *et al.*, 2004] MCCOWAN, I. A., MOORE, D., DINES, J., GATICA-PEREZ, D., FLYNN, M., WELLNER, P. et BOURLARD, H. (2004). On the use of information retrieval measures for speech recognition evaluation. Rapport technique, IDIAP.

[McDonald, 1996] MCDONALD, D. (1996). Internal and external evidence in the identification and semantic categorization of proper names. *Corpus processing for lexical acquisition*, pages 21–39.

[Merchant *et al.*, 1996] MERCHANT, R., OKUROWSKI, M. E. et CHINCHOR, N. (1996). The multilingual entity task (MET) overview. *In Proceedings of a workshop on held at Vienna, Virginia : May 6-8, 1996*, pages 445–447. Association for Computational Linguistics.

[Miller, 1955] MILLER, G. A. (1955). Note on the bias of information estimates. *Information theory in psychology : Problems and methods*, 2:95–100.

[Morris, 2002] MORRIS, A. (2002). An information theoretic measure of sequence recognition performance. Rapport technique, IDIAP.

[Morris *et al.*, 2004] MORRIS, A. C., MAIER, V. et GREEN, P. (2004). From WER and RIL to MER and WIL : improved evaluation measures for connected speech recognition. *In INTERSPEECH*.

[Munteanu *et al.*, 2006] MUNTEANU, C., BAECKER, R., PENN, G., TOMS, E. et JAMES, D. (2006). The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives. *In Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 493–502. ACM.

[Nadeau et Sekine, 2007] NADEAU, D. et SEKINE, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[Naylor et Gaubitch, 2010] NAYLOR, P. A. et GAUBITCH, N. D. (2010). *Speech dereverberation*. Springer Science & Business Media.

[Nirenburg *et al.*, 2003] NIRENBURG, S., SOMERS, H. L. et WILKS, Y. (2003). *Readings in machine translation*. MIT Press.

[NIST, 1999] NIST (1999). The ACE website.

[NIST, 2000] NIST (2000). ACE Pilot Study Task Definition, 2000.

[NIST, 2002] NIST (2002). The ace 2002 evaluation plan.

[NIST, 2004] NIST (2004). The ACE 2004 (ACE04) Evaluation plan.

[NIST, 2005] NIST (2005). The ACE 2005 (ACE05) Evaluation plan.

[NIST, 2007] NIST (2007). The ACE 2007 (ACE07) Evaluation plan.

[NIST, 2008] NIST (2008). The ACE 2008 (ACE08) Evaluation plan.

[Nouvel, 2012] NOUVEL, D. (2012). Reconnaissance des entités nommées par exploration de règles d'annotation. *These de doctorat, Université François Rabelais de Tours, Tours, France*.

[Osenova et Kolkovska, 2002] OSENOVA, P. et KOLKOVSKA, S. (2002). Combining the named-entity recognition task and NP chunking strategy for robust preprocessing. *In Proceedings of the Workshop on Treebanks and Linguistic Theories, September*, pages 20–21.

[Pallett *et al.*, 1995] PALLETT, D., FISCUS, J. G., FISHER, W. M., GAROFOLO, J. S., LUND, B. A. et PRZYBOCKI, M. A. (1995). 1994 Benchmark Tests for the ARPA Spoken Language Program. *In Proceedings of the SLST Workshop*.

[Pallett, 1985] PALLETT, D. S. (1985). Performance assessment of automatic speech recognizers. *J. Res. Natl. Bureau of Standards*, 90:371–387.

[Pallett, 2003] PALLETT, D. S. (2003). A look at NIST's benchmark ASR tests : past, present, and future. *In Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 483–488. IEEE.

[Papineni *et al.*, 2002] PAPINENI, K., ROUKOS, S., WARD, T. et ZHU, W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. *In Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

[Papoulis et Pillai, 2002] PAPOULIS, A. et PILLAI, S. U. (2002). *Probability, random variables, and stochastic processes*. Tata McGraw-Hill Education.

[Paroubek, 2013] PAROUBEK, P. (2013). *De l'évaluation en Traitement Automatique des Langues*. Habilitation à diriger des recherches, Université Paris Sud.

[Plamondon *et al.*, 2004] PLAMONDON, L., LAPALME, G. et PELLETIER, F. (2004). Anonymisation de décisions de justice. *In XIe Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2004)*, pages 367–376.

[Poibeau, 1999] POIBEAU, T. (1999). Le repérage des entités nommées, un enjeu pour les systemes de veille. *In Terminologies Nouvelles, actes du colloque Terminologie et Intelligence Artificielle (TIA ?99)*, volume 2, pages 43–51.

[Poibeau, 2001] POIBEAU, T. (2001). Deconstructing harry : une évaluation des systemes de repérage d'entités nommées. *Revue de la société d'électronique, d'électricité et de traitement de l'information*, 5:25–33.

[Przybocki *et al.*, 1999] PRZYBOCKI, M. A., FISCUS, J. G., GAROFOLO, J. S. et PALLETT, D. S. (1999). 1998 Hub-4 information extraction evaluation. *In Proc. DARPA Broadcast News Workshop,(Herndon, Va, USA)*, pages 13–18.

[Pustejovsky *et al.*, 2003] PUSTEJOVSKY, J., CASTANO, J. M., INGRIA, R., SAURI, R., GAIZAUSKAS, R. J., SETZER, A., KATZ, G. et RADEV, D. R. (2003). TimeML : Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

[QUAERO, 2008] QUAERO (2008). QUARO web site.

[Raymond, 2013] RAYMOND, C. (2013). Robust tree-structured named entities recognition from speech. *In Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 8475–8479. IEEE.

[Riccardi et Gorin, 1998] RICCARDI, G. et GORIN, A. L. (1998). Stochastic language models for speech recognition and understanding. *In ICSLP*.

[Rosset *et al.*, 2011] ROSSET, S., GROUIN, C. et ZWEIGENBAUM, P. (2011). *Entités nommées structurées : guide d'annotation QUAERO*. LIMSI-Centre national de la recherche scientifique.

[SAIC, 1998] SAIC (1998). Proceedings of the Seventh Message Understanding Conference (MUC-7).

[Salton et Buckley, 1988] SALTON, G. et BUCKLEY, C. (1988). Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523.

[San José *et al.*, 2010] SAN JOSÉ, H. B., RECOURCÉ, G., COUTO, J., SAGOT, B., STERN, R. et TEYSSOU, D. (2010). Traitement des inconnus : une approche systématique de l'incomplétude lexicale. *In Traitement Automatique des Langues Naturelles : TALN 2010*.

[Sang, 2002] SANG, T. K. (2002). Introduction to the CoNLL-2002 shared task : Language-independent named entity recognition. *In Proceedings of the 6th conference on Natural language learning at Morristown, NJ, USA*. Association for Computational Linguistics.

[Santos *et al.*, 2006] SANTOS, D., SECO, N., CARDOSO, N. et VILELA, R. (2006). HAREM : An Advanced NER Evaluation Contest for Portuguese. *In Proceedings of LREC*, pages 1986–1991.

[Sekine et Isahara, 2000] SEKINE, S. et ISAHARA, H. (2000). IREX : IR & IE Evaluation Project in Japanese. *In LREC*, pages 1977–1980.

[Sekine et Nobata, 2004] SEKINE, S. et NOBATA, C. (2004). Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. *In LREC*, pages 1977–1980.

[Sekine *et al.*, 2002] SEKINE, S., SUDO, K. et NOBATA, C. (2002). Extended Named Entity Hierarchy. *In LREC*.

[Speranza, 2007] SPERANZA, M. (2007). Evalita 2007 : the named entity recognition task. *Proc. of EVALITA*.

[Stern et Sagot, 2010] STERN, R. et SAGOT, B. (2010). Détection et résolution d'entités nommées dans des dépêches d'agence. *In Traitement Automatique des Langues Naturelles : TALN 2010*.

[Sundheim, 1991] SUNDHEIM, B. M. (1991). Overview of the third message understanding evaluation and conference. *In Proceedings of the 3rd conference on Message understanding*, pages 3–16. Association for Computational Linguistics.

[Sundheim, 1992] SUNDHEIM, B. M. (1992). Overview of the fourth message understanding evaluation and conference. *In Proceedings of the 4th conference on Message understanding*, pages 3–21. Association for Computational Linguistics.

[Sundheim, 1996] SUNDHEIM, B. M. (1996). Overview of results of the MUC-6 evaluation. *In Proceedings of a workshop on held at Vienna, Virginia : May 6-8, 1996*, pages 423–442. Association for Computational Linguistics.

[Tjong Kim *et al.*, 2003] TJONG KIM, S., F, E. et DE MEULDER, F. (2003). Introduction to the CoNLL-2003 shared task : Language-independent named entity recognition. *In Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics.

[Virtanen *et al.*, 2012] VIRTANEN, T., SINGH, R. et RAJ, B. (2012). *Techniques for noise robustness in automatic speech recognition*. John Wiley & Sons.

[Wacholder *et al.*, 1997] WACHOLDER, N., RAVIN, Y. et CHOI, M. (1997). Disambiguation of proper names in text. *In Proceedings of the fifth conference on Applied natural language processing*, pages 202–208. Association for Computational Linguistics.

[Wang *et al.*, 2003] WANG, Y.-Y., ACERO, A. et CHELBA, C. (2003). Is word error rate a good indicator for spoken language understanding accuracy. *In Automatic Speech Recognition and Understanding, 2003. ASRU'03. 2003 IEEE Workshop on*, pages 577–582. IEEE.

[Watanabe et Le Roux, 2014] WATANABE, S. et LE ROUX, J. (2014). Black box obtimization for automatic speech recognition. *In Acoustic, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 3256–3260. IEEE.

[Young *et al.*, 1997] YOUNG, S. J., ADDA-DEKKER, M., AUBERT, X., DUGAST, C., GAUVAIN, J.-L., KERSHAW, D. J., LAMEL, L., LEEUWEN, D., PYE, D., ROBINSON, A. J. *et al.* (1997). Multilingual large vocabulary speech recognition : the European SQALE project. *Computer Speech & Language*, 11(1):73–89.

[Young et Chase, 1998] YOUNG, S. J. et CHASE, L. (1998). Speech recognition evaluation : a review of the US CSR and LVCSR programmes. *Computer Speech & Language*, 12(4):263–279.

Annexes

Annexe A

Résultats de la Corrélations de Kendall

		NER-1	NER-2	NER-3	NER-4	NER-5	NER-6	NER-7	mean
	WER	0,48	0,41	0,22	0,43	0,42	0,36	0,41	0,39
	NE-WER	**0,68**	0,59	0,26	0,54	0,49	0,60	0,46	0,52
	WIL	0,40	0,33	0,16	0,37	0,37	0,28	0,33	0,32
	F-mesure	0,40	0,33	0,16	0,37	0,37	0,28	0,33	0,32
ATENE	*Baseline*	0,60	**0,72**	0,37	**0,69**	0,53	0,62	0,61	0,59
	Base+préf.+suf.	0,65	**0,72**	0,35	0,68	0,55	**0,66**	**0,63**	**0,61**
	Base+pos	0,61	0,68	0,32	0,68	0,55	0,60	0,61	0,58
	Base+préf.+suf.+pos	0,64	0,71	0,35	0,68	0,56	0,67	0,62	0,60
	Baseline+maj.	0,34	0,51	0,49	0,60	0,42	0,59	0,57	0,50
	Base+préf.+suf.+maj.	0,36	0,65	0,53	0,64	0,46	0,60	0,63	0,55
	Base+pos+maj.	0,32	0,55	0,45	0,58	0,40	0,54	0,59	0,49
	Base+préf.+suf.+pos+maj.	0,41	0,62	0,45	**0,65**	0,51	**0,59**	0,64	0,55

TABLEAU A.1 – Corrélations de Kendall moyennes entre les classements fondés sur les performances des systèmes de REN mesurés en $ETER$ sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, pour les données de la campagne d'évaluation ETAPE.

		NER-1	NER-2	NER-3	mean
WER		0,33	0,29	0,03	0,22
NE-WER		0,18	0,14	-0,03	0,09
WIL		0,33	0,29	0,03	0,22
F-mesure		0,33	0,29	0,03	0,22
$ATENE_{DS}$	*Baseline*	0,22	0,33	0,29	0,28
	Base+préf.+suf.	0,25	0,37	0,33	0,32
	Base+pos	0,29	0,11	0,29	0,23
	Base+préf.+suf.+pos	0,25	0,29	0,25	0,27
	Baseline+maj.	0,14	0,33	0,44	0,30
	Base+préf.+suf.+maj.	**0,37**	0,48	**0,51**	**0,45**
	Base+pos+maj.	0,18	0,14	0,48	0,27
	Base+préf.+suf.+pos+maj.	0,33	**0,51**	0,48	0,44

TABLEAU A.2 – Corrélations de Kendall moyennes entre les classements fondés sur performances des systèmes de REN mesurer en $ETER$ sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, pour les données de la campagne d'évaluation QUAERO.

		NER-1	NER-2	NER-3	NER-4	NER-5	NER-6	NER-7	mean
WER		0,40	0,56	0,56	0,56	0,49	0,64	0,55	0,54
NE-WER		**0,61**	0,57	0,32	0,40	0,35	0,57	0,58	0,49
WIL		0,32	0,51	**0,53**	0,55	0,50	0,65	0,48	0,51
Rappel		0,28	0,48	0,49	0,53	0,50	0,62	0,47	0,48
$ATENE_{DS}$	*Baseline*	0,52	0,63	0,41	0,63	0,46	0,64	0,55	0,55
	Base+préf.+suf.	0,60	**0,65**	0,40	0,56	0,46	0,65	**0,59**	0,56
	Base+pos	0,53	0,64	0,50	**0,64**	**0,51**	**0,71**	0,58	**0,59**
	Base+préf.+suf.+pos	0,58	**0,65**	0,36	0,54	0,44	0,61	0,57	0,54

TABLEAU A.3 – Corrélations de Kendall moyennes entre les classements fondés sur les taux d'erreur d'omission et de substitution des systèmes de REN mesurés en $ETER_{DS}$ sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, pour résultats de la campagne ETAPE.

		NER-1	NER-2	NER-3	mean
WER		0,66	0,44	**0,70**	0,60
NE-WER		**0,74**	0,51	**0,70**	**0,65**
WIL		0,66	0,44	**0,70**	0,60
Rappel		0,66	0,44	**0,70**	0,60
$ATENE_{DS}$	*Baseline+maj.*	0,29	0,44	0,40	0,38
	Base+préf.+suf.+maj.	0,59	0,51	0,55	0,55
	Base+pos+maj.	0,59	**0,59**	0,62	0,60
	Base+préf.+suf.+pos+maj.	0,59	**0,59**	0,62	0,60

TABLEAU A.4 – Corrélations de Kendall moyennes entre les classements fondés sur les taux d'erreur d'omission et de substitution des systèmes de REN mesurer en $ETER_{DS}$ sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, pour résultats de la campagne QUAERO.

		NER-1	NER-2	NER-3	NER-4	NER-5	NER-6	NER-7	mean
WER		0,24	0,27	-0,15	0,27	0,19	0,14	0,26	0,17
NE-WER		0,38	0,41	0,16	0,44	0,45	0,39	0,36	0,37
WIL		0,18	0,20	-0,23	0,20	0,14	0,06	0,18	0,10
Précision		0,19	0,22	-0,18	0,22	0,14	0,10	0,21	0,13
$ATENE_I$	*Baseline*	0,64	0,61	**0,24**	0,72	0,66	0,58	0,60	0,58
	Base+préf.+suf.	**0,65**	**0,64**	0,20	**0,75**	0,65	0,62	**0,61**	**0,59**
	Base+pos	0,61	0,57	0,21	0,71	**0,67**	0,59	0,56	0,56
	Base+préf.+suf.+pos	0,61	0,62	0,23	0,73	0,63	**0,64**	0,58	0,58

TABLEAU A.5 – Corrélations de Kendall moyennes entre les classements fondés sur les taux d'erreur d'insertion des systèmes de REN mesurées en $ETER_I$ sur les différentes sorties de RAP et les classements fondés sur les différentes mesures de d'évaluation, pour résultats de la campagne ETAPE.

		NER-1	NER-2	NER-3	mean
WER		0,07	-0,11	-0,33	-0,12
NE-WER		-0,14	-0,33	-0,48	-0,32
WIL		0,07	-0,11	-0,33	-0,12
Précision		0,07	-0,11	-0,33	-0,12
$ATENE_I$	*Baseline+maj.*	0,37	0,40	0,25	0,34
	Base+préf.+suf.+maj.	**0,48**	**0,51**	0,22	**0,40**
	Base+pos+maj.	0,37	0,33	**0,48**	0,39
	Base+préf.+suf.+pos+maj.	**0,48**	0,51	0,22	**0,40**

TABLEAU A.6 – Corrélations de Kendall moyennes entre les classements fondés sur les taux d'erreur d'insertions des systèmes de REN mesurer en $ETER_I$ sur les différentes sorties de RAP et les classements fondés sur les différentes mesures d'évaluation, pour résultats de la campagne QUAERO.