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

Mohamed Hatmi* — **Christine Jacquin*** — **Sylvain Meignier**** — **Emmanuel Morin*** — **Solen Quiniou***

* *Université de Nantes, LINA UMR CNRS 6241*
2 rue de la Houssinière, BP 92208 F-44322 Nantes cedex 3
{mohamed.hatmi,christine.jacquin,emmanuel.morin,solen.quiniou}@univ-nantes.fr
** *Université du Maine, LIUM*
Institut d'Informatique Claude Chappe, avenue Laënnec F-72085 Le Mans cedex 9
sylvain.meignier@lium.univ-lemans.fr

RÉSUMÉ. *Nous nous intéressons à la tâche de reconnaissance des entités nommées pour la modalité orale. Cette tâche pose un certain nombre de difficultés qui sont inhérentes au traitement de l'oral. Dans ce travail, nous proposons d'étudier le couplage étroit entre la tâche de transcription de la parole et la tâche de reconnaissance des entités nommées. Dans ce but, nous détournons les fonctionnalités de base d'un système de transcription de la parole pour le transformer en un système de reconnaissance des entités nommées. Ainsi, en mobilisant les connaissances propres au traitement de la parole dans le cadre de la tâche liée à la reconnaissance des entités nommées, nous assurons une plus grande synergie entre ces deux tâches qui se traduit par une augmentation significative de la qualité de la reconnaissance des entités nommées.*

ABSTRACT. *We are interested in the recognition of named entities for the speech modality. Some difficulties may arise for this task due to speech processing. In this work, we propose to study the tight pairing between the speech recognition task and the named entity recognition task. For that purpose, we take away the basic functionnalities of a speech recognition system to turn it into a named entity recognition system. Therefore, by mobilising the inherent knowledge of the speech processing to the named entity recognition task, we ensure a better synergy between the two tasks. This leads to a significative rise in the quality of the named entity recognition task.*

MOTS-CLÉS : *reconnaissance des entités nommées, parole, modèle de langage.*

KEYWORDS: *named entity recognition, automatic speech recognition, language modeling.*

1. Introduction

Le traitement des entités nommées (EN) est une tâche incontournable dans de nombreuses applications relevant du traitement automatique des langues (TAL) notamment l'indexation, la recherche d'information ou encore la traduction. Si la reconnaissance des entités nommées (REN) peut être considérée comme un problème bien traité pour du texte bien formé[1] relevant de la langue générale (F-mesure supérieure à 0,9), cette tâche pose encore de nombreuses difficultés lorsque l'on s'éloigne de la langue générale pour aller en direction de langues de spécialité ou encore en passant du texte bien formé à d'autres modalités comme la parole ou les textes manuscrits.

En parole, la reconnaissance des entités nommées se heurte à un certain nombre de difficultés inhérentes aux caractéristiques de cette modalité. L'absence de capitalisation[2] et le manque de ponctuation induisent respectivement des ambiguïtés graphiques et de segmentation. La présence de disfluences et d'erreurs de reconnaissance constitue deux autres difficultés caractéristiques du traitement automatique de la parole (TAP) auxquelles se confrontent les systèmes de REN.

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

Dans la suite de cet article, nous commençons par dresser un état des lieux de la REN, en section 2, en mettant l'accent sur la modalité parole. La section 3 décrit

1. Nous entendons par « texte bien formé » un document relevant de l'écrit qui respecte les règles syntaxiques et typographiques.

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

les ressources radiophoniques et les mesures de performance utilisées dans le cadre de ce travail. La section 4 commence par présenter les différents systèmes de RAP et de REN que nous avons mobilisés avant de décrire l'approche que nous mettons en œuvre. La section 5 évalue les performances de notre proposition à travers différentes expérimentations et propose une analyse des erreurs rencontrées. Finalement, la section 6 vient conclure ce travail et propose quelques perspectives.

2. Reconnaissance des entités nommées en parole

Le concept d'entité nommée est apparu au milieu des années 90 à l'occasion de la sixième conférence américaine d'évaluation MUC (*Message Understanding Conference*) (Grishman et Sundheim, 1996). Les conférences MUC avaient pour objectif de promouvoir la recherche en extraction d'information. Les tâches proposées consistaient à remplir de façon automatique des formulaires relatifs à des événements (par exemple des fusions d'entreprises, des attentats...). Dans ce cadre, certains objets textuels, ayant une importance applicative particulière en TAL, ont été regroupés sous le nom d'entités nommées (EN). Le matériel textuel exploité était de langue anglaise et concernait des dépêches de presse écrite ciblées sur un thème spécifique. Au fil des années, les recherches concernant ces objets linguistiques se sont focalisées sur des problématiques de plus en plus complexes comme la désambiguïsation et l'annotation enrichie mais aussi sur leur reconnaissance dans des contextes différents (autres langues et autres modalités). Les campagnes d'évaluation CoNLL (*Conference on Natural Language Learning*), qui ont eu lieu dès 2002, s'intéressaient toujours à la modalité écrite (articles de presse) mais pour des langues cibles différentes (espagnol et hollandais en 2002 ; anglais et allemand en 2003) (Tjong Kim Sang, 2002 ; Tjong Kim Sang et De Meulder, 2003). En parallèle, des premiers travaux de REN relatifs à d'autres modalités ont vu le jour, comme ceux relatifs à la modalité parole (Kubala *et al.*, 1998). Cette nouvelle problématique a été prise en compte lors de nouvelles campagnes d'évaluation, notamment les campagnes ACE (*Automatic Content Extraction*) qui se sont déroulées entre les années 1999 et 2008. Celles-ci, tout en se focalisant sur différentes langues telles que l'anglais, l'espagnol et l'arabe, ont ciblé des documents issus du TAP (Doddington *et al.*, 2004). Suivant cette dynamique, la communauté scientifique francophone de la parole a aussi proposé une tâche relative à la REN sur des transcriptions de la parole lors des campagnes ESTER 1 (Gravier *et al.*, 2004), ESTER 2 (Galliano *et al.*, 2009) et ETAPE (Gravier *et al.*, 2012). Ce changement de contexte n'a d'ailleurs pas seulement concerné la modalité parole mais aussi d'autres modalités, comme la modalité texte imprimé. Les campagnes ACE en 2004 (Doddington *et al.*, 2004) ainsi que la campagne liée au projet Quaero en 2011 (Dinarelli et Rosset, 2012) ont comporté une tâche dédiée à la REN sur des documents océrisés (obtenus après reconnaissance par des systèmes d'OCR – pour *Optical Character Recognition* – sur des textes imprimés).

Comme nous l'avons exposé précédemment, la tâche de REN s'est complexifiée au fil des années, non seulement au niveau des exigences relatives à la granularité

de la reconnaissance mais aussi au niveau de son contexte d'application (passage du texte écrit bien formé monolingue à celui du texte multilingue et aussi à d'autres modalités). Une composante commune à ce nouveau contexte de travail est la place de la REN par rapport à la tâche globale de la transcription du signal. Celle-ci est très souvent considérée uniquement comme un post-traitement appliqué aux sorties du processus de transcription du signal, quelle que soit la modalité étudiée (Dinarelli et Rosset, 2012 ; Galliano *et al.*, 2009). Ceci a amené les chercheurs à se heurter à de nouveaux problèmes liés aux nouvelles modalités des textes que nous allons expliciter dans le cadre de la modalité parole. Nous pouvons distinguer les problèmes liés à l'ambiguïté graphique (manque de capitalisation), les problèmes de segmentation (manque de ponctuation) ainsi que des problèmes intrinsèques à la modalité parole (disfluences, par exemple) (Ji-hwan et Woodland, 2000 ; Béchet et Charton, 2010). Une autre difficulté liée à la reconnaissance est aussi apparue : le bruit dans les transcriptions automatiques, qui est dû aux erreurs de reconnaissance et aux mots hors vocabulaire. Miller *et al.* (2000) ont montré que les erreurs de reconnaissance qui apparaissent dans les mots constituant les EN ou dans les mots inclus dans leur contexte ont un impact direct sur les performances de la REN. Une augmentation de 1 % du taux d'erreur sur les mots induit une diminution presque identique du taux de reconnaissance des EN, que ce soit d'ailleurs sur la modalité parole ou texte imprimé. Pour pallier ces problèmes, certains restaurent les ponctuations et la capitalisation dans la transcription (Gravano *et al.*, 2009). D'autres, comme Sudoh *et al.* (2006), détectent les EN qui ont été mal retranscrites afin qu'elles ne soient plus prises en compte par le processus de REN. Pour ce faire, un score de confiance est calculé et indique si un mot a été correctement reconnu ou non par le système de RAP.

Les approches précédentes se positionnent en tant que post-traitement au processus de transcription du signal ; elles n'ont en général pas donné une totale satisfaction concernant l'amélioration de la REN. D'autres approches sont apparues pour réaliser un couplage plus ou moins étroit entre le processus de transcription du signal et la REN. Parada *et al.* (2011) ont proposé d'inclure des caractéristiques propres aux mots hors vocabulaire dans le système de RAP. Un étiqueteur à base de CRF (*Conditional Random Field*) exploite ainsi les sorties d'un détecteur de mots hors vocabulaire pour identifier ou ignorer des régions dans lesquelles les EN sont incorrectement transcrites. D'autres travaux exploitent des sorties intermédiaires des systèmes de reconnaissance comme les n meilleures hypothèses (Zhai *et al.*, 2004), les graphes de mots (Favre *et al.*, 2005) ou les réseaux de confusion (Kurata *et al.*, 2012). Zhai *et al.* (2004) annotent les n meilleures hypothèses en sortie du système de RAP à l'aide d'un système qui s'appuie sur le principe d'entropie maximale. Un vote à partir des scores obtenus par le processus de transcription du signal et le système de REN est ensuite mis en place pour déterminer l'EN la plus probable (même si elle n'était pas présente dans la meilleure hypothèse du système de RAP). Les travaux de Favre *et al.* (2005), quant à eux, proposent de récupérer les EN directement dans le graphe de mots en sortie du système de RAP. La grammaire liée aux EN intègre les mots du lexique du système de RAP et exploite le graphe de mots complet afin d'extraire la liste des n meilleures hypothèses concernant les EN. Kuruta *et al.* (2012) exploitent les réseaux de confusion

construits en sortie du système de RAP. Les nœuds du réseau sont d'abord regroupés selon leur similarité (*clustering*) puis, à l'aide d'un algorithme s'appuyant sur le principe d'entropie maximale, le modèle relatif au système de REN est appris en prenant en compte les clusters précédemment déterminés. Ces différentes approches ont permis d'améliorer la REN sur la modalité parole. Cependant, il n'est pas évident de comparer les résultats obtenus car les expérimentations n'ont pas toujours été menées dans des contextes similaires (transcription du signal plus ou moins difficile selon la tâche ciblée et détection des entités nommées plus ou moins difficile suivant le type de média audio : dialogues enregistrés dans des centres d'appels, émissions radiophoniques...). Il est à noter que des méthodes similaires ont également été mises en œuvre dans d'autres modalités (sur des textes océrisés, par exemple) et ont donné des résultats comparables (Subramanian *et al.*, 2011).

Les types d'approches présentés précédemment se focalisent principalement sur l'amélioration des performances de la REN qui est la tâche finale visée mais ne cherchent aucunement à améliorer le système de RAP en termes de WER par exemple. En ce sens, Wang *et al.* (2003) ont montré que, dans le contexte d'une autre tâche qui est celle de la compréhension de la parole, la recherche en premier lieu de la diminution du WER du système de RAP est souvent moins efficace que la prise en compte dans les modèles des spécificités de la tâche traitée. D'autres auteurs ont choisi de chercher à améliorer conjointement les performances du système de RAP (WER) et la tâche de compréhension. Servan *et al.* (2006) et Deoras *et al.* (2013) préconisent de commencer par apprendre de manière indépendante les différents modèles puis de construire le modèle final à travers une combinaison pondérée des différents modèles initiaux. Dans ce contexte, les premiers auteurs utilisent une approche générative (automates à états finis) mais qui ne permet pas de capturer toutes les caractéristiques des différents modèles. Deoras *et al.* (2013) mettent aussi en œuvre une approche semblable. Ils conjuguent aussi, mais de manière différente, les scores obtenus par les différents modèles sur le graphe de mots en sortie du système de RAP. Ils utilisent une approche discriminative (à base de CRF) qui leur permet de capturer plus de caractéristiques des modèles qu'une approche à base d'automates à états finis. Avec cette stratégie, ils obtiennent à la fois une amélioration du WER du système de RAP ainsi que des performances de la tâche de compréhension.

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

par rapport à leur détection dans des transcriptions d'émissions radiophoniques par exemple (les EN sont moins complexes). Cependant, cette approche originale nous semble intéressante et, dans cet article, nous avons comme objectif de l'étudier de manière approfondie pour déterminer son réel impact sur la transcription du signal et sur la REN. L'idée générale étant que le système de RAP devient ainsi un système dédié à la tâche de REN en produisant en sortie des transcriptions étiquetées en entités nommées. Le système sera moins performant au niveau du taux d'erreur global sur les mots. En revanche, cette approche améliorera la détection et la catégorisation des EN.

3. Présentation des corpus et des mesures de performance

Dans cette section, nous présentons le corpus utilisé dans ce travail ainsi que les mesures de performance permettant d'évaluer notre approche.

3.1. *Corpus ESTER 2*

Pour nos différentes expérimentations, nous avons utilisé le corpus distribué dans le cadre de la campagne d'évaluation ESTER 2 (Galliano *et al.*, 2009). Ce corpus est constitué de transcriptions manuelles d'émissions radiophoniques et de transcriptions manuelles rapides de radios africaines. La plupart des émissions enregistrées correspondent à des journaux d'information et à des émissions conversationnelles provenant de quatre sources : France Inter, Radio France International (RFI), Africa number one (Africa 1) et Radio Télévision du Maroc (TVME).

Une première partie de ce corpus a été utilisée comme corpus d'apprentissage pour construire les modèles de langage du système de transcription. Une deuxième partie a été utilisée comme corpus de développement pour adapter le vocabulaire (corpus de développement *Dev1*) et pour ajuster certains paramètres du système proposé (corpus de développement *Dev2*). La dernière partie est utilisée pour évaluer les performances de notre système[3]. Les transcriptions manuelles des ressources radiophoniques composant les corpus de développement et de test sont enrichies par une annotation manuelle des EN. Le guide d'annotation adopté est celui de la campagne ESTER 2[4]. Les EN sont ainsi réparties en sept catégories principales (personne, lieu, organisation, fonction, production humaine, montant et date et heure), elles-mêmes divisées en 37 sous-catégories. Seules les catégories principales sont considérées dans ce travail. Le tableau 1 présente les principales caractéristiques de ces quatre corpus.

3. Le corpus d'apprentissage et le corpus de développement *Dev1* sont les mêmes que ceux utilisés pour la campagne ESTER 2. En revanche, le corpus de test utilisé pour la campagne ESTER 2 a été découpé en deux corpus – notre corpus de développement *Dev2* et notre corpus de test – afin de disposer d'un corpus de développement indépendant du corpus *Dev1* pour l'ajustement des paramètres de notre système (cf. section 5.1).

4. `www.afcp-parole.org/camp_eval_systemes_transcription/docs/Conventions_EN_ESTER2_v01.pdf`

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

Tableau 1. *Description des corpus utilisés (issus d'ESTER 2)*

3.2. *Mesures de performance*

La qualité de la RAP est évaluée en termes de taux d'erreur sur les mots (WER, pour *Word Error Rate*). Cette mesure indique le pourcentage de mots incorrectement reconnus dans la transcription par rapport au texte de référence. Nous considérons un mot comme étant un token. Les types d'erreurs considérés sont définis comme suit :

– substitution (S) : le nombre de mots de la référence incorrectement reconnus dans la transcription ;

– suppression (D) : le nombre de mots de la référence omis dans la transcription ;

– insertion (I) : le nombre de mots ajoutés dans la transcription et absents de la référence.

Le WER est donné par l'équation 1 :

$$WER = \frac{S + D + I}{Nombre\ de\ mots\ de\ la\ référence} \quad [1]$$

La qualité de la REN est évaluée en termes de F-mesure et de coût d'appariement (Makhoul *et al.*, 1999) (SER, pour *Slot Error Rate*). La F-mesure (cf. équation 4) représente la moyenne harmonique pondérée du rappel et de la précision. Le rappel et la précision sont calculés sur la base du nombre de mots contenus dans les EN. Le rappel (cf. équation 2) est défini comme étant le nombre de mots correctement étiquetés dans la transcription au regard du nombre de mots étiquetés dans la référence. La précision (cf. équation 3) correspond au nombre de mots correctement étiquetés dans la transcription au regard du nombre de mots correctement et incorrectement étiquetés dans la transcription.

$$Rappel = \frac{Nombre\ de\ mots\ correctement\ étiquetés}{Nombre\ de\ mots\ étiquetés\ de\ la\ référence} \quad [2]$$

$$Précision = \frac{Nombre\ de\ mots\ correctement\ étiquetés}{Nombre\ de\ mots\ étiquetés\ de\ la\ transcription} \quad [3]$$

$$F{-}mesure = \frac{2 * Rappel * Précision}{Rappel + Précision} \quad [4]$$

Le SER combine et pondère différents types d'erreurs. Plusieurs variantes du SER ont été définies en fonction des types d'erreurs pris en compte, des poids attribués à chaque type d'erreur et de la base d'évaluation (nombre d'EN ou nombre de mots contenus dans les EN). Dans nos expérimentations, nous utilisons la variante de SER, définie pour la campagne d'évaluation ESTER 2, pour évaluer les performances des systèmes de REN. La base de calcul est le nombre d'EN. Les types d'erreurs considérés sont définis comme suit :

– suppression (D) : le nombre d'EN totalement manquées par le système ;

– insertion (I) : le nombre d'EN dans la transcription n'ayant aucun mot commun avec une EN de la référence ;

– erreur de type (T) : le nombre d'EN détectées dans la transcription avec des frontières correctes mais une catégorie incorrecte ;

– erreur de frontière (F) : le nombre d'EN détectées dans la transcription avec une catégorie correcte mais des frontières incorrectes ;

– erreur de type et de frontière (TF) : le nombre d'EN détectées dans la transcription avec une catégorie et des frontières incorrectes.

Le SER est alors donné par l'équation 5 :

$$SER = \frac{D + I + 0,5 * T + 0,5 * F + 0,8 * TF}{Nombre\ d'entités\ étiquetées\ de\ la\ référence} \quad [5]$$

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

4. Intégration de la REN au système de transcription

Notre approche consiste à intégrer la tâche de REN au niveau du système de transcription. Dans cette optique, nous nous sommes appuyés sur le système de RAP du LIUM et sur le système de REN LIA_NE. Nous commençons par présenter ces deux systèmes avant d'introduire notre nouveau système.

4.1. *Système de transcription automatique de la parole du LIUM*

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

Le système de transcription repose sur trois ressources : i) un modèle acoustique qui calcule la suite de phonèmes la plus probable, ii) un modèle de langage qui propose la séquence de mots la plus correcte syntaxiquement et iii) un vocabulaire de 122 981 mots contenant les phonétisations des mots connus du système. Ce dernier fait le lien entre la séquence de phonèmes générée par le modèle acoustique et la suite de mots donnée par le modèle de langage durant le décodage.

La transcription est obtenue en enchaînant cinq passes de décodage. La première passe permet d'obtenir une hypothèse qui est utilisée pour adapter les modèles acoustiques aux locuteurs. La passe 2 génère un graphe qui servira à guider le décodage de la passe 3. La réduction de l'espace de recherche au graphe obtenu à l'issue de la passe 2 permet d'utiliser les triphones intermots des modèles acoustiques. L'utilisation des contextes intermots permet un décodage plus précis que celui des deux premières passes, là où seuls les triphones intramots sont exploités. Mais cette méthode est plus gourmande en ressources (en temps processeur comme en occupation mémoire), d'où la nécessité de réduire l'espace de recherche à un graphe prédéfini. La passe 3 crée ainsi un graphe qui est réévalué à l'aide d'un modèle de langage quadrigramme dans la passe 4 (dans les trois passes précédentes, le modèle de langage est un modèle trigramme). Finalement, la cinquième et dernière passe transforme le graphe généré en fin de passe 4 en un réseau de confusion duquel la transcription finale est extraite.

De cette description technique[5], il faut retenir le fait que deux modèles de langage sont utilisés : un modèle trigramme (dans les trois premières passes) et un modèle quadrigramme (dans les deux dernières passes). Ces modèles de langage partagent le même vocabulaire, qui est identique dans toutes les passes de la transcription. Dans la suite de cet article, nous référencerons ce système par SRAP_SANS_REN.

4.2. *Système de reconnaissance des entités nommées* LIA_NE

Le système LIA_NE (Béchet et Charton, 2010) utilisé dans ce travail a été développé par le Laboratoire Informatique d'Avignon (LIA). Ce système repose sur une méthode à base d'apprentissage supervisé permettant de déterminer conjointement les frontières et les catégories des EN.

Le système LIA_NE utilise, dans un premier temps, un modèle génératif à base de modèles de Markov cachés (*Hidden Markov Models*, HMM) afin d'annoter le texte en parties du discours. Il applique ensuite un modèle discriminatif à base de champs conditionnels aléatoires (*Conditional Random Field*, CRF) pour annoter les EN. Les résultats du modèle permettent au modèle CRF d'éliminer un certain nombre d'ambiguïtés et de construire des règles d'une portée plus large en se fondant sur les traits contextuels des mots et sur leurs parties du discours. Le CRF a été appris sur le corpus d'apprentissage de la campagne d'évaluation ESTER 1 (Galliano *et al.*, 2005).

5. Le lecteur pourra se référer à l'article de Deléglise *et al.* (2009) pour une description complète du système.

Ce corpus – qui représente une sous-partie du corpus d'apprentissage ESTER 2 (cf. tableau 1) – est constitué d'une centaine d'heures d'émissions radiophoniques francophones manuellement transcrites et annotées en EN. La figure 1 montre les résultats de la REN pour quelques phrases extraites du corpus ESTER 2.

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

Figure 1. *Exemple de texte annoté par* LIA_NE *au format d'ESTER 2*

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

4.3. *Couplage de la REN au processus de transcription*

Contrairement aux approches traditionnelles séparant la tâche de transcription de celle de REN, notre objectif est d'intégrer la tâche de REN au sein du processus de transcription. De cette manière, le système de transcription devient un système dédié à la tâche de REN et produit en sortie des transcriptions étiquetées en entités nommées. Le problème revient donc à trouver la séquence de mots étiquetés en EN la plus probable $(\widehat{W}, \widehat{E}) = (w_1, e_1), (w_2, e_2), (w_3, e_3), ..., (w_k, e_k)$, étant donné la séquence d'observations acoustiques $(X = x_1, x_2, x_3, ..., x_p)$:

$$(\widehat{W}, \widehat{E}) = \arg\max_{w} \ \mathbb{P}(W, E|X) \qquad [6]$$

Pour parvenir à cet objectif, notre démarche consiste à nous appuyer sur les ressources mobilisées par le système de transcription pour les enrichir et ainsi construire notre système de REN. Les deux ressources exploitées sont, d'une part, les mots du vocabulaire, et, d'autre part, les modèles de langage construits à partir des corpus d'apprentissage du système de transcription. Ces corpus doivent avoir une couverture importante par rapport au domaine ciblé car si un mot n'apparaît pas dans le vocabulaire, il ne pourra pas apparaître dans les transcriptions. Nous supposons également que les EN qui apparaissent dans les transcriptions automatiques conservent

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

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

La méthode proposée consiste alors à exploiter les connaissances mobilisées pour la tâche de transcription en intégrant les informations relatives aux EN dans les modèles de langage et dans le vocabulaire. Cette méthode se décompose en quatre étapes : i) annotation des corpus d'apprentissage en EN, ii) annotation du vocabulaire du système de transcription, iii) adaptation du dictionnaire phonétisé et iv) adaptation des modèles de langage. Le système qui en résulte sera nommé SRAP_REN dans la suite de l'article.

4.3.1. *Annotation des corpus d'apprentissage*

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

Après l'annotation des corpus d'apprentissage, nous avons transformé les annotations obtenues pour les mettre au format BI. L'encodage BI permet de préciser la catégorie ainsi que les frontières des EN en indiquant, pour chaque mot appartenant à une EN, s'il correspond au début (B) ou à l'intérieur de l'entité (I). Les mots n'appartenant à aucune EN ne sont pas annotés. La figure 2 présente un exemple de texte annoté en EN en utilisant l'encodage BI.

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

Tableau 2. *Corpus d'apprentissage pour la construction des modèles de langage*

(a) le journal Denis**-pers-b** Astagneau**-pers-i** bonsoir Denis**-pers-b**

(b) son directeur**-fonc-b** général**-fonc-i** Philibeaux**-pers-b** Dillon**-pers-i** démissionne

(c) sur place pour France**-org-b** Inter**-org-i** Frédéric**-pers-b** Barrère**-pers-i**

(d) depuis vingt cinq**-time-b** ans**-time-i** dans le cinquième**-loc-b** arrondissement**-loc-i** Benoît**-pers-b** Collombat**-pers-i**

(e) autour de moins**-amount-b** trois**-amount-i** à moins**-amount-b** cinq**-amount-i** degrés**-amount-i** en général

(f) il est vingt**-time-b** heures**-time-i** à Johannesburg**-loc-b**

(g) deux**-amount-b** milliards**-amount-i** d'**-amount-i** euros**-amount-i** ont été mobilisés

Figure 2. *Exemple de texte annoté au format BI*

4.3.2. *Annotation du vocabulaire du système de transcription*

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

4.3.3. *Adaptation du dictionnaire phonétisé*

L'adaptation du dictionnaire phonétisé consiste à associer une ou plusieurs phonétisations aux mots étiquetés en EN. L'ajout des étiquettes n'ayant aucun effet sur la phonétisation, les mots étiquetés conservent alors les mêmes phonétisations que celles attribuées aux mots sans étiquette dans le dictionnaire du système SRAP_SANS_REN. Par exemple, la phonétisation associée à « nantes**-loc-b** » et à « nantes**-org-b** » est « nn an tt ». Nous supposons ici que la prononciation d'un mot n'est pas dépendante de la catégorie de l'entité nommée ; cette hypothèse est générale-

ment vérifiée. On notera cependant que certains prénoms correspondant à des noms de villes peuvent avoir une prononciation différente de celle de la ville. C'est le cas, par exemple, pour Paris Hilton ou encore pour Milan Milutinovi (président de la Serbie).

4.3.4. *Adaptation des modèles de langage*

Pour permettre au système de transcription de générer des sorties annotées en EN, nous avons réappris les modèles de langage pour qu'ils prennent également en compte les étiquettes des EN. Le rôle du modèle de langage est ainsi de contraindre le système de transcription à générer des transcriptions syntaxiquement correctes et annotées en EN. Aucune coupe (*cut-off*) n'a été appliquée, c'est-à-dire que tous les *n*-grammes du vocabulaire observés dans les corpus d'apprentissage sont pris en compte (même ceux n'apparaissant qu'une seule fois). Les deux modèles de langage exploités dans le système de transcription sont des modèles d'ordre 3 (trigramme) et 4 (quadrigramme), estimés avec l'outil SRILM (Stolcke, 2002). Un modèle trigramme et un modèle quadrigramme sont tout d'abord créés pour chaque corpus d'apprentissage. Ces modèles sont ensuite fusionnés par interpolation linéaire[6].

5. Expérimentations et résultats

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

5.1. *Détermination de la taille optimale du vocabulaire*

5.1.1. *Constitution des différents ensembles de vocabulaire*

Dans un premier temps, nous avons associé aux mots du vocabulaire du système SRAP_SANS_REN, toutes les étiquettes assignées dans les corpus annotés. Comme aucune coupe n'a été effectuée dans les modèles de langage, tout mot annoté apparais-

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

sant au moins une fois dans les corpus d'apprentissage est inclus dans le vocabulaire. Cela augmente la taille du vocabulaire de 122 981 à 503 192 mots. Il en résulte que certains mots sont annotés de manière erronée en raison des erreurs commises par le système LIA_NE lors de l'étiquetage des corpus d'apprentissage. Ces erreurs pourront alors se retrouver dans la sortie finale du système de transcription.

Afin de filtrer les étiquettes erronées ou rares (ce qui permettra également de diminuer la taille du vocabulaire), nous nous sommes appuyés sur l'hypothèse suivante : les mots incorrectement étiquetés apparaissent avec une fréquence beaucoup moins élevée que les mêmes mots correctement étiquetés. Par exemple, le mot « footballistique » apparaît 88 fois sans étiquette et une seule fois avec l'étiquette personne (« footballistique**-pers-b** »). Dans ce cas, nous considérons que le mot « footballistique**-pers-b** » a été incorrectement étiqueté. La méthode utilisée pour réaliser le filtrage est inspirée de celle proposée par Allauzen et Gauvain (2004) dans le cadre de la détermination du vocabulaire d'un système de transcription. Elle consiste à :

1) entraîner un modèle unigramme pour chaque corpus d'apprentissage utilisé pour créer les modèles de langage, en utilisant le vocabulaire étiqueté (503 192 mots) ;

2) calculer les coefficients d'interpolation entre ces modèles unigrammes sur le corpus de développement *Dev1* ;

3) construire le modèle de langage global unigramme avec les coefficients d'interpolation calculés à l'étape 2 ;

4) extraire les n mots étiquetés les plus probables du modèle unigramme en fonction d'un seuil choisi.

Notre but étant d'étudier l'impact du vocabulaire sur le WER et le SER du système SRAP_REN, nous avons constitué différents ensembles de mots. Chaque ensemble est construit en faisant varier le seuil relatif aux probabilités du modèle unigramme. Les seuils ont été choisis afin de déterminer des ensembles de mots dont la taille augmente de manière uniforme (chaque ensemble de mots constituant ainsi un vocabulaire). Il est à noter que, dans tous les cas, le vocabulaire sélectionné couvre l'ensemble des mots qui apparaissent dans le vocabulaire du système SRAP_SANS_REN (nous ajoutons ainsi, dans ces ensembles obtenus, les mots qui appartiennent au vocabulaire du système SRAP_SANS_REN s'ils n'y figuraient pas déjà). Nous obtenons ainsi, douze ensembles de mots[7].

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

Nous avons commencé par adapter les modèles de langage et le dictionnaire phonétisé, en fonction de chaque vocabulaire sélectionné. Nous obtenons ainsi douze systèmes de transcription avec des vocabulaires différents. Ces douze systèmes vont nous

7. Nous avons au préalable découpé de manière empirique le vocabulaire en huit ensembles distincts afin de réaliser l'étude. Afin de bien visualiser les variations de WER en fonction de la taille du vocabulaire, nous avons été amenés à redécouper quatre de ces ensembles.

permettre d'évaluer l'influence du choix du vocabulaire sur la qualité de la transcription (WER) ainsi que sur la qualité de la REN (SER et F-mesure). Dans ces expérimentations, nous utilisons le corpus de développement *Dev2*.

Étude relative à la qualité de la transcription

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

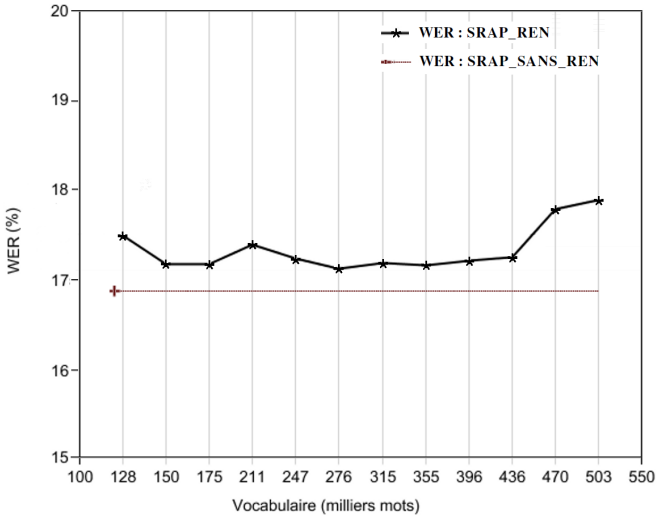

Figure 3. *Influence sur le WER du choix du vocabulaire annoté du système de transcription (corpus de développement* Dev2*)*

La figure 3 montre la variation du WER en fonction de la taille N de chaque vocabulaire annoté sélectionné. Chaque point de la première courbe correspond ainsi à un système SRAP_REN avec un des vocabulaires annotés sélectionné. Il est à noter que le système SRAP_SANS_REN affiche un WER de 16,88 % sur le corpus *Dev2*, ce qui est représenté par la ligne droite sur la figure. Nous pouvons observer que, sans aucun filtrage du vocabulaire étiqueté ($N = 503\,192$ mots), l'intégration de la REN entraîne une augmentation du WER d'environ 1 %. Cela est principalement dû à l'augmentation de la taille de l'espace de recherche. Nous remarquons ensuite que la diminution de la taille du vocabulaire permet le plus souvent une amélioration du WER pour les différents N sélectionnés. Le WER le plus proche de celui du système SRAP_SANS_REN est obtenu en utilisant un vocabulaire comportant

$N = 276\,757$ mots (WER de 17,11 %, soit 0,23 % d'augmentation). Nous voyons donc que l'intégration de la REN au sein du système de transcription a un faible impact sur la qualité du système de RAP.

Étude relative à la qualité de la REN

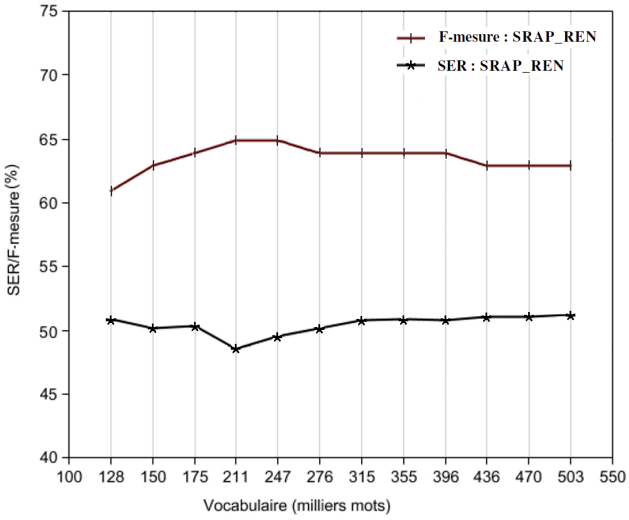

Figure 4. *Influence, sur la qualité de la REN, du choix du vocabulaire annoté du système de transcription (corpus de développement* Dev2*)*

À partir des transcriptions annotées obtenues par chacun des douze systèmes de transcription, nous avons calculé le SER et la F-mesure. La figure 4 montre les différents scores obtenus. En utilisant l'ensemble du vocabulaire étiqueté ($N = 503\,192$ mots), le système affiche un SER de 51,23 % et une F-mesure de 63,00 %. La diminution de la taille du vocabulaire permet de constater une amélioration progressive du SER et de la F-mesure jusqu'à une certaine valeur de N, puis ensuite une dégradation de ces mêmes indicateurs. L'amélioration s'explique par le fait que le filtrage de certaines étiquettes erronées permet de contrôler l'étiquetage de certains mots et donc d'éviter la reproduction de certaines erreurs commises par le système LIA_NE lors de l'annotation des corpus d'apprentissage. Par exemple, si le mot « DSK » apparaît dans le lexique comme étant une personne (sous la forme « DSK**-pers-b** »), il sera toujours étiqueté dans les transcriptions comme étant une personne et ce quel que soit son contexte ou sa position. Il existe cependant une taille N du vocabulaire à partir de laquelle le nombre de mots annotés comme étant des EN devient trop faible et dégrade la qualité des résultats. Les meilleurs résultats sont obtenus en utilisant un vocabulaire de taille $N = 211\,576$ mots. Pour cette valeur, le SER est égal à 48,56 % et la F-mesure à 65,00 %.

5.1.3. *Choix de la taille du vocabulaire*

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

5.2. *Évaluation des performances du système* SRAP_REN *sur le corpus de test*

Nous évaluons ici l'influence de l'intégration de la REN au sein du système de RAP sur la qualité de transcription et sur la REN, en utilisant le corpus de test. Le corpus de test est constitué de 51 836 mots. Parmi ceux-ci, 6 158 mots appartiennent à des entités nommées (soit 8,41 % des mots de ce corpus), ce qui représente un total de 3 289 EN.

5.2.1. *Influence de l'intégration de la REN sur la qualité de transcription*

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

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

Tableau 3. *Taux d'erreur sur les mots (WER) avant et après l'intégration de la REN dans le processus de RAP (corpus de test)*

Le tableau 3 donne le WER obtenu pour les deux versions du décodeur. Nous remarquons tout d'abord que le WER sur les mots correspondant à des EN est plus élevé que le WER sur tous les mots, pour les deux versions du décodeur. Cette augmentation du WER peut provenir d'une part d'une faible représentation de la séquence de mots constituant les EN dans les corpus d'apprentissage des modèles de langage. D'autre part, les noms de personnes, les noms d'organisations et plus généralement les mots d'origine étrangère ont fréquemment des phonétisations erronées générant des erreurs lors du décodage. Nous constatons également que l'intégration de la REN au sein du

décodeur entraîne une augmentation de 0,94 % pour le WER sur tous les mots des transcriptions et de 2,58 % pour le WER uniquement sur les mots correspondant à des EN. Cela est dû à l'intégration des marqueurs d'EN sur les mots de vocabulaire qui a pour effet d'augmenter le nombre de mots du vocabulaire mais aussi de modifier les probabilités des modèles de langage, ce qui les rend moins performants en termes de WER.

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

Tableau 4. *Nombre de mots et WER par catégorie d'entité nommée, avant et après l'intégration de la REN dans le processus de RAP (corpus de test)*

Le tableau 4 présente le nombre de mots correspondant à des EN, pour chaque catégorie d'EN, ainsi que les performances obtenues par les deux versions du décodeur, pour chacune de ces catégories. Nous pouvons tout d'abord constater que la qualité de transcription se trouve légèrement dégradée pour les différentes catégories. Les résultats obtenus par les deux systèmes montrent également que les noms de personnes sont les plus difficiles à transcrire, avec un WER élevé d'environ 46 %. En effet, en dehors des journalistes qui sont présents à l'antenne quotidiennement, les noms des personnes sont généralement cités pour un fait d'actualité ponctuel. Ces noms correspondent fréquemment aux invités de l'émission ou aux personnes relatives aux faits traités. La sélection d'un nombre limité d'entités de cette classe à inclure dans le vocabulaire du système de RAP est donc une tâche délicate (Dufour *et al.*, 2012). En ce qui concerne les noms d'organisations, de lieux, de fonctions et de productions humaines, leur transcription s'avère satisfaisante. Ces entités évoluent moins rapidement dans le temps que les entités de type personne. La transcription des montants ainsi que des dates et heures est également satisfaisante. Cela est dû au fait que ces entités sont généralement composées de noms communs qui sont présents dans le vocabulaire du système de RAP.

5.2.2. *Influence de l'intégration de la REN sur la qualité de la REN*

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

transcriptions résultantes. D'autre part, nous avons décodé les données de test avec le système SRAP_REN. Nous avons également utilisé SRAP_REN uniquement pour l'étape de transcription suivi de LIA_NE pour la tâche de REN.

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

Tableau 5. *Performances de la REN avant et après l'intégration de la REN dans le processus de RAP (corpus de test) en termes de SER, F-mesure (F), précision (P) et rappel (R)*

Les deux premières lignes du tableau 5 montrent les résultats obtenus pour les deux versions du décodeur. Malgré la légère dégradation de la qualité de la transcription (et ce de manière plus marquée pour les EN que pour les autres mots du vocabulaire), nous constatons que l'intégration de la REN au sein du système de RAP permet une amélioration importante du SER et de la F-mesure : une diminution de 4,79 % du SER et une augmentation de 5,00 % de F-mesure. En effet, nous cherchons une meilleure reconnaissance des EN et non pas une meilleure transcription. Nous attribuons le gain obtenu au fait qu'assigner des étiquettes d'EN au niveau du système de RAP permet de contrôler l'étiquetage des mots composant les EN, notamment quand il y a des erreurs de transcription dans le contexte de ceux-ci. La dernière ligne du tableau 5 montre les résultats obtenus en utilisant SRAP_REN uniquement pour l'étape de transcription puis en utilisant LIA_NE pour la REN. Comme le système de reconnaissance utilisé a un moins bon WER que SRAP_SANS_REN, nous constatons que le taux de rappel relatif aux mots composant les entités est légèrement moins bon que pour l'exploitation de SRAP_SANS_REN et LIA_NE en cascade. Le taux de précision est cependant légèrement meilleur. Le taux de SER est, quant à lui, stable. Il est à noter que la composition de ces deux derniers systèmes conduit aussi à de moins bonnes performances pour la REN que celles obtenues par SRAP_REN (variation de 5 % du SER et de la F-mesure).

Le tableau 6 présente le nombre d'EN de chaque catégorie, ainsi que les performances de la REN pour chacune des catégories, avant et après l'intégration de la REN dans le processus de RAP. Nous pouvons tout d'abord remarquer une diminution du SER pour les différentes catégories. Cette diminution est de l'ordre de 5 % pour les personnes, les organisations et les lieux, qui correspondent aux EN les plus représentées dans le corpus de test (environ 72 % des EN). En ce qui concerne les fonctions et les productions humaines (dont le nombre est assez faible par rapport aux autres catégories), la baisse du SER est plus importante : 15,16 % et 14,58 %, respectivement. Pour expliquer cette baisse, nous pouvons regarder le détail des nombres d'erreurs de chaque type qui sont utilisés pour le calcul du SER.

Catégorie	Nombre d'EN	SER SRAP_SANS_REN puis LIA_NE	SER SRAP_REN
Personne	701	69,90 %	64,76 %
Organisation	807	76,64 %	71,56 %
Lieu	864	60,19 %	54,46 %
Fonction	300	65,33 %	50,17 %
Production humaine	24	112,50 %	97,92 %
Montant	156	53,21 %	51,28 %
Date et heure	437	57,44 %	55,38 %

Tableau 6. *Nombre d'EN et SER par catégorie d'entité nommée avant et après l'intégration de la REN dans le processus de RAP (corpus de test)*

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

Tableau 7. *Nombre d'erreurs pour les différents types d'erreurs par catégorie d'entité nommée avant et après l'intégration de la REN dans le processus de RAP (corpus de test) : insertion (I), suppression (S), erreur de type (ET) et erreur de frontière (EF)*

Le tableau 7 montre les différents types d'erreurs comptabilisés dans le SER. En utilisant SRAP_REN, nous constatons une diminution des suppressions pour les personnes et les organisations, une diminution des insertions pour les personnes ainsi qu'une diminution des erreurs de frontière pour les lieux. En ce qui concerne les fonctions, la baisse du SER est due aux suppressions qui sont presque deux fois moins présentes (123 suppressions pour SRAP_SANS_REN puis LIA_NE contre 75 pour SRAP_REN) ainsi qu'à une diminution des erreurs de frontière. Pour les productions humaines, la baisse du SER concerne principalement les suppressions et les insertions. Il y a en revanche une légère augmentation du nombre d'erreurs de frontière mais les EN de type production humaine sont les plus longues avec une moyenne de 3,4 mots par EN d'où une plus grande difficulté à correctement identifier leurs frontières. Fi-

nalement, pour les montants et les dates et heures, la baisse du SER est de l'ordre de 2 %. Le nombre d'erreurs de type et de frontière reste constant pour les montants mais ces EN sont également parmi les plus longues avec en moyenne 2,7 mots par EN. En ce qui concerne les dates et heures, la baisse porte essentiellement sur les erreurs de frontière.

5.2.3. *Influence de la prise en compte des majuscules*

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

5.3. *Analyse des erreurs*

Afin d'avoir une vue plus fine des problèmes liés à notre approche, nous avons réalisé une typologie des erreurs commises par notre système. Notre corpus de test de référence comprend 6 158 mots appartenant à des EN. Parmi ces mots, 2 537 sont soit mal retranscrits, soit pas du tout étiquetés comme EN ou alors mal étiquetés (mauvais typage) par notre système. Nous avons analysé manuellement un échantillon de 450 mots choisis aléatoirement parmi ces 2 537 mots afin d'étudier les faiblesses du système. Nous avons ainsi défini quatre classes d'erreurs que nous explicitons dans les paragraphes suivants.

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

séquence de mots correspondante (c'est-à-dire du *n*-gramme obtenu avec un modèle de langage sans EN). Le choix du vocabulaire adapté au domaine à traiter reste un point délicat comme dans tous les systèmes de transcription de la parole et de REN.

Une deuxième classe d'erreurs correspond à la reproduction d'erreurs commises par LIA_NE. Ces erreurs (qui touchent 20,44 % des mots que nous étudions) sont inhérentes à notre type d'approche. En effet, nous avons annoté les corpus d'apprentissage à l'aide d'un outil automatique qui, malgré ses bonnes performances, commet des erreurs d'annotation qui sont alors réitérées par notre système. Par exemple, pour l'entité « ministère**-org-b** de**-org-i** la**-org-i** défense**-org-i** française », le mot « française » n'est pas inclus dans l'entité. La correction de ce type d'erreur nécessite son traitement en amont, c'est-à-dire au niveau de l'annotation initiale des corpus d'apprentissage. Pour atteindre cet objectif, un traitement manuel est bien sûr à exclure vu la taille des corpus d'apprentissage mais un outil plus performant de REN pourrait être mis en œuvre.

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

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

6. Conclusion

Dans ce travail, nous nous sommes intéressés à la tâche de la reconnaissance des entités nommées en parole. La reconnaissance des entités nommées pour ce type de modalité pose un certain nombre de difficultés inhérentes à ses caractéristiques intrinsèques. Nous avons proposé de détourner les fonctionnalités de base d'un système de transcription de la parole pour en faire un système de reconnaissance des entités nommées (ou plus exactement un système de transcription de la parole dont les sorties sont annotées en entités nommées). Ainsi, en mobilisant, pour la tâche de reconnaissance des entités nommées, les connaissances exploitées pour la tâche de reconnaissance de la parole, nous assurons une plus grande synergie entre ces deux tâches.

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

Les deux éléments clés de notre système de reconnaissance des entités nommées sont le vocabulaire et les modèles de langage. Un avantage indéniable de cette approche provient du fait que le vocabulaire est unique et commun à l'ensemble de notre système, ce qui évite de créer d'éventuels effets de bord entre la transcription et la détection des entités nommées. En revanche, un mot hors vocabulaire ne pourra ni être transcrit, ni étiqueté en EN avec ce type d'approche. En outre, la portée des règles d'annotation est limitée. En effet, cette approche ne permet pas la désambiguïsation de certaines EN dont les formes qui les composent ne sont pas apparues dans le corpus d'apprentissage comme faisant partie d'une EN. Le choix du vocabulaire adapté au domaine à traiter reste un point délicat comme dans tous les systèmes de transcription de la parole et de REN. Les modèles de langage d'ordre 3 et 4 intégrant les étiquettes d'entités nommées sont appris à partir de textes annotés automatiquement en EN. Cette solution, peu coûteuse par rapport à une annotation manuelle, est fonctionnelle quoique imparfaite. Une des sources d'erreurs provient ainsi de la reproduction des erreurs les plus fréquentes de LIA_NE par notre système. Bien entendu, la méthode proposée nécessite de disposer d'un système de REN mais celui-ci n'a pas besoin d'être spécifique au traitement de l'oral. La majorité des corpus d'apprentissage pour les modèles de langage sont ainsi des journaux ou des dépêches avec des marqueurs forts d'entités nommées comme les ponctuations et les majuscules.

Nous avons aussi proposé dans cette étude une analyse des erreurs de reconnaissance des entités nommées, ce qui permet de mieux cerner les limites de notre approche qui sont autant de perspectives notamment en ce qui concerne les mots hors vocabulaire.

Remerciements

Ce travail a été réalisé dans le cadre du projet DEPART (Documents Ecrits et Paroles – Reconnaissance et Traduction), `www.projet-depart.org`, financé par la région des Pays de la Loire.

7. Bibliographie

Allauzen A., Gauvain J.-L., « Construction automatique du vocabulaire d'un système de transcription », *Actes des XXV^es Journées d'étude sur la parole (JEP'04)*, Fès, Maroc, 2004.

Béchet F., Charton E., « Unsupervised knowledge acquisition for Extracting Named Entities from speech », *Proceedings of the 35th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'10)*, Dallas, TX, USA, p. 151-179, 2010.

Deléglise P., Estève Y., Meignier S., Merlin T., « Improvements to the LIUM french ASR system based on CMU Sphinx : what helps to significantly reduce the word error rate ? », *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech'09)*, Brighton, United Kingdom, p. 2123-2126, 2009.

Deoras A., Tür G., Sarikaya R., Hakkani-Tür D. Z., « Joint Discriminative Decoding of Words and Semantic Tags for Spoken Language Understanding », *IEEE Transactions on Audio, Speech & Language Processing*, vol. 21, n° 8, p. 1612-1621, 2013.

Dinarelli M., Rosset S., « Tree-Structured Named Entity Recognition on OCR Data : Analysis, Processing and Results », *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, p. 1266-1272, 2012.

Doddington G., Mitchelland A., Przybocki M., Ramshaw L., Strassel S., Weischedel R., « The Automatic Content Extraction (ACE) Program—Tasks, Data, and Evaluation », *Proceedings of the 4th international conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, p. 837-840, 2004.

Dufour R., Damnati G., Charlet D., Béchet F., « Automatic transcription error recovery for Person Name Recognition », *Proceedings of the 13th Annual Conference of the International Speech Communication Association (Interspeech'12)*, Portland, OR, USA, 2012.

Favre B., Béchet F., Nocéra P., « Robust Named Entity extraction from large spoken archives », *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05)*, Vancouver, British Columbia, Canada, p. 491-498, 2005.

Galliano S., Geoffrois E., Mostefa D., Choukri K., Bonastre J.-F., Gravier G., « The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News », *Proceedings of the 9th European Conference on Speech Communication and Technology (InterSpeech'05)*, Lisbon, Portugal, p. 1149-1152, 2005.

Galliano S., Gravier G., Chaubard L., « The Ester 2 Evaluation Campaign for the Rich Transcription of French Radio Broadcasts », *Proceedings of the 10th Annual Conference of the International Speech Communication Association (Interspeech'09)*, Brighton, UK, p. 2583-2586, 2009.

Gravano A., Jansche M., Bacchiani M., « Restoring punctuation and capitalization in transcribed speech », *Proceedings of the 34th IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'09)*, Taipei, Taiwan, p. 4741-4744, 2009.

Gravier G., Adda G., Paulsson N., Carré M., Giraudel A., Galibert O., « The ETAPE corpus for the evaluation of speech-based TV content processing in the French language », *Proceedings of the 8th international conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, p. 114-118, 2012.

Gravier G., Bonastre J.-F., Geoffrois E., Galliano S., Tait K. M., Choukri K., « ESTER, une campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en

français », *Actes des XXV^{es} Journées d'étude sur la parole (JEP'04)*, Fès, Maroc, p. 253-256, 2004.

Grishman R., Sundheim B., « Message Understanding Conference-6 : A Brief History », *Proceedings of the 16th conference on Computational linguistics (COLING'96)*, Copenhagen, Denmark, p. 466-471, 1996.

Hori A., Atsushi N., « An extremely-large-vocabulary approach to named entity extraction from speech », *Proceedings of the 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP'06)*, Toulouse, France, p. 973-976, 2006.

Huang X., Alleva F., Hwang M.-Y., Rosenfeld R., « An overview of the SPHINX-II speech recognition system », *Proceedings of the workshop on Human Language Technology (HLT'93)*, Princeton, NJ, USA, p. 81-86, 1993.

Ji-hwan K., Woodland P., « A Rule-Based Named Entity Recognition System for Speech Input », *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'00)*, Beijing, China, p. 521-524, 2000.

Kubala F., Schwartz R., Stone R., Weischedel R., « Named Entity Extraction From Speech », *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, Lansdowne, VA, USA, p. 287-292, 1998.

Kurata G., Itoh N., Nishimura M., Sethy A., Ramabhadran B., « Leveraging word confusion networks for named entity modeling and detection from conversational telephone speech », *Speech Communication*, vol. 54, n° 3, p. 491-502, 2012.

Makhoul J., Kubala F., Schwartz R., Weischedel R., « Performance measures for information extraction », *Proceedings of DARPA Broadcast News Workshop*, Herndon, VA, USA, p. 249-252, 1999.

Miller D., Boisen S., Schwartz R., Stone R., Weischedel R., « Named Entity Extraction from Noisy Input : Speech and OCR », *Proceedings of the 6th Applied Natural Language Processing Conference (ANLP'00)*, Seattle, WA, USA, p. 316-324, 2000.

Parada C., Dredze M., Jelinek F., « OOV Sensitive Named-Entity Recognition in Speech », *Proceedings of the 12th Annual Conference of the International Speech Communication Association (Interspeech'11)*, Florence, Italy, p. 2085-2088, 2011.

Servan C., Raymond C., Béchet F., Nocéra P., « Conceptual decoding from word lattices : application to the spoken dialogue corpus media », *Proceedings of the 7th Annual Conference of the International Speech Communication Association (Interspeech'06)*, Pittsburgh, PA, USA, p. 1614-1617, 2006.

Stolcke A., « SRILM - An Extensible Language Modeling Toolkit », *Proceedings of the International Conference of Spoken Language Processing (ICSLP'02)*, Denver, CO, USA, p. 901-904, 2002.

Subramanian K., Prasad R., Natarajan P., « Robust named entity detection from optical character recognition output », *Proceedings of the 11th International Conference on Document Analysis and Recognition (IJDAR'11)*, Beijing, China, p. 189-200, 2011.

Sudoh K., Tsukada H., Isozaki H., « Incorporating speech recognition confidence into discriminative named entity recognition of speech data », *Proceedings of the 44th Annual Meeting on Association for Computational Linguistics (ACL'06)*, Sydney, Australia, p. 617-624, 2006.

Tjong Kim Sang E. F., « Introduction to the CoNLL-2002 shared task : language-independent named entity recognition », *Proceedings of the 6th conference on Natural language learning (COLING'02)*, Taipei, Taiwan, p. 1-4, 2002.

Tjong Kim Sang E. F., De Meulder F., « Introduction to the CoNLL-2003 shared task : language-independent named entity recognition », *Proceedings of the 7th conference on Natural language learning (ConLL'03)*, Edmonton, Canada, p. 155-158, 2003.

Wang Y.-Y., Acero A., Chelba C., « Is word error rate a good indicator for spoken language understanding accuracy », *Automatic Speech Recognition and Understanding (ASRU'03)*, IEEE, p. 577-582, 2003.

Zhai L., Fung P., Schwartz R., Carpuat M., Wu D., « Using N-best lists for named entity recognition from Chinese speech », *Proceedings of Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics (HLT-NAACL'04)*, Boston, MA, USA, p. 37-40, 2004.