



Adapter le vocabulaire d'un système de transcription automatique de la parole aux thèmes abordés

Rapport de stage
Master 2 Recherche en Informatique, 2010-2011

Auteur : Florent TISSIER <florent.tissier@laposte.net>

Encadrants : Guillaume GRAVIER <guillaume.gravier@irisa.fr>
Pascale SÉBILLOT <pascale.sebillot@irisa.fr>

Équipe : TexMex

Table des matières

Remerciements **3**

Résumé **4**

Introduction **5**

1 État de l'art **6**
1.1 Système de reconnaissance automatique de la parole 7
1.1.1 Vocabulaire 7
1.1.2 Modèle acoustique 8
1.1.3 Modèle de langue 8
1.2 Adaptation thématique 9
1.2.1 Documents d'adaptation 10
1.2.2 Adaptation du modèle de langue 11
1.3 Adaptation du vocabulaire 12
1.3.1 Recherche de mots candidats 13
1.3.2 Intégration des nouveaux mots dans le système de reconnaissance automatique de la parole 15
1.4 Conclusion 16

2 Données **16**
2.1 Système de reconnaissance de la parole 17
2.2 Corpus d'adaptation thématique 18

3 Sélection de mots candidats **19**
3.1 Mesures préliminaires 19
3.2 Évaluation 20
3.3 Filtre phonétique 21
3.3.1 Principe 21
3.3.2 Méthode de distance d'édition 21
3.3.3 Seuil et résultats 23
3.4 Filtre grammatical 25
3.4.1 Principe 25
3.4.2 Méthode d'estimation de probabilités avec un lissage par backoff . . 26
3.4.3 Seuil et résultats 28

4 Travaux en cours **29**
4.1 Mots candidats générés par flexion 29
4.1.1 Principe 29
4.1.2 Méthode 30
4.2 Mesures de confiance et filtrage 30
4.2.1 Principe 31

Conclusion et perspectives **31**

Annexe A **33**

Annexe B **34**

Références **35**

Table des figures

1 Représentation du lexique phonétisé sous la forme d'un arbre lexical. Cerclés en gras, les nœuds correspondant à la fin de la transcription phonétique d'un mot [Lecorvé 2010] 8
2 Représentation d'une séquence de mots W sous la forme de modèles de Markov cachés pour le calcul de la vraisemblance p(Y|W) du signal de parole Y [Lecorvé 2010] 9
3 Stratégie d'adaptation thématique d'un système de reconnaissance automatique de la parole pour un document multimédia donné [Lecorvé 2010] . 10
4 Sortie textuelle du système de transcription automatique de la parole . . . 17
5 Sortie phonétique du système de transcription automatique de la parole . . 18
6 Sortie du système de transcription automatique de la parole phonétisée . . 21
7 Matrice des scores pour chaque phonème entre une requête et une sous-chaîne du document 22
8 Courbe rappel-précision permettant de déterminer le seuil du filtre phonétique 24
9 Exemple de fichier étiqueté avec TreeTagger. 27
10 Courbe rappel-précision des filtres grammatical et phonétique 28
11 Liste des *tags* des étiqueteurs morphosyntaxique disambig et TreeTagger . 34

Liste des tableaux

1 Tableau récapitulatif du nombre de mots hors vocabulaire sélectionnés avec le filtre phonétique 25
2 Tableau récapitulatif des valeurs de WER et LER pour les transcriptions de référence et celles pour lesquelles nous avons rajouté des mots hors vocabulaire 33

Remerciements

Je souhaite remercier tout d'abord l'équipe pédagogique de l'ESIR et du Master Recherche en Informatique de Rennes 1 pour m'avoir permis de suivre des enseignements dans ces deux diplômes. J'adresse également mes remerciements à mes maîtres de stage, Guillaume Gravier et Pascale Sébillot, pour les conseils qu'ils ont pu me donner, ainsi que pour les nombreuses relectures et corrections de ce rapport. Enfin, je tiens à remercier toute l'équipe TexMex pour son accueil au cours de ces mois de stage.

Résumé

Pour accéder à la sémantique de documents multimédias (flux de TV, vidéos), nous utilisons des systèmes de reconnaissance de la parole qui produisent un texte correspondant à ce qui a été prononcé. Ces systèmes fonctionnent, entre autres, grâce à un vocabulaire contenant l'ensemble figé des mots qu'ils peuvent reconnaître, et à un modèle de langue regroupant les probabilités de succession des mots du vocabulaire. Bien qu'en règle générale les transcriptions ainsi obtenues soient fiables, ces systèmes ne sont pas spécialisés pour des thèmes bien précis (le sport, la guerre en Irack, *etc.*). Ce manque de spécialisation conduit à des erreurs de transcription sur des termes propres à chaque thème abordé. Une adaptation thématique est donc nécessaire, tant au niveau du modèle de langue qu'à celui du vocabulaire.

C'est sur ce dernier point que porte le stage. En effet, il s'agit de sélectionner un ensemble de mots candidats, dans l'idéal les mots absents des transcriptions. Une solution immédiate consiste à sélectionner, dans un ensemble de textes traitant du thème d'une transcription (corpus d'adaptation), tous les mots absents du vocabulaire et à les intégrer dans le système afin de produire de nouvelles transcriptions. Cette méthode n'est pas satisfaisante car elle peut conduire à un afflux de mots, sans intérêt réel et dégradant éventuellement les performances du système.

Ainsi, pour éviter ces problèmes, deux méthodes de filtrage, l'une phonétique et l'autre grammaticale, ont été mises en place et permettent de récupérer, pour une transcription, 63.6% des mots manquants sur les 50% possibles de retrouver dans des corpus d'adaptation. Un second objectif est donc d'augmenter le nombre de mots récupérables en se basant sur les racines morphologiques des mots présents dans les transcriptions.

Mots-clés : transcription automatique de la parole, adaptation thématique, sélection de mots hors vocabulaire

Introduction

Avec le développement des systèmes de télécommunication et d'Internet, énormément de documents multimédias sont produits et enregistrés, comme par exemple les flux de télévision. Les différentes chaînes de télévision, les sites de partage de vidéos tels que YouTube ou Dailymotion, regroupent et proposent des milliards de vidéos. Il devient alors vital de pouvoir accéder à la sémantique de ces documents, c'est-à-dire, réussir à extraire les éléments qui leur donnent du sens. Par exemple, les informations ainsi récupérées peuvent servir à produire des résumés ou faciliter l'indexation et la navigation dans un ensemble de documents. Parmi les différents médias (son, image, vidéo, *etc.*) contenus dans un document multimédia, la parole est, dans la majorité des cas, celui qui est le plus riche en information sémantique.

Un grand nombre de travaux de recherche porte sur la transcription automatique de documents audio ou multimédias à l'aide de systèmes de reconnaissance de la parole. Ces systèmes sont basés majoritairement sur des méthodes statistiques qui permettent d'associer des mots à des sons présents dans un signal audio. Les mots que le système peut reconnaître sont regroupés dans ce qu'on appelle le vocabulaire. Parmi les composants d'un système de reconnaissance de la parole qui utilisent des méthodes statistiques, un premier donne la probabilité qu'un son soit associé à un mot ; c'est ce qui est appelé modèle acoustique. Un second permet de déterminer les séquences de mots les plus probables ; c'est ce qui est appelé modèle de langue. Le vocabulaire et le modèle de langue sont appris une fois pour toute sur un ensemble de textes variés pour obtenir une connaissance générale et étendue de la langue.

En règle générale, les systèmes de reconnaissance de la parole permettent de transcrire plutôt correctement n'importe quel document multimédia. En revanche, ces systèmes ne sont pas aussi efficaces pour transcrire des documents qui ont un thème précis, tel que la guerre en Irak ou le sport. Dans ce cas en effet, certains des mots importants sont mal transcrits car ils sont absents du vocabulaire, ou ils ne sont pas pris en compte dans le modèle de langue parce que ce sont des mots rares. Une manière de résoudre ce problème serait de rendre un système de reconnaissance de la parole spécifique à un thème donné. Il faudrait alors être sûr du thème abordé dans le document avant transcription. De plus, il ne serait pas envisageable de se servir du système sur des documents parlant d'un sujet différent. C'est un usage peu probable dans la réalité car les systèmes de reconnaissance de la parole sont, en règle générale, utilisés pour transcrire des documents traitant de sujets variés comme par exemple un journal télévisé.

Pour pallier la trop grande généralité des systèmes de reconnaissance de la parole, des chercheurs travaillent sur des solutions afin que ces systèmes s'ajustent aux thèmes de chaque document à transcrire. C'est ce qui s'appelle l'adaptation thématique. Pour effectuer cette tâche, des travaux existent pour adapter soit le modèle acoustique, soit le modèle de langue, soit le vocabulaire. Tandis que certains traitent déjà le cas du modèle de langue [Seymore & Rosenfeld 1997, Lecorvé 2010], aucune méthode efficace n'a encore été trouvée pour résoudre le cas du vocabulaire. C'est donc sur ce point que se focalise le sujet du stage. L'objectif est de trouver les mots à ajouter au vocabulaire d'un système de transcription pour s'adapter à un thème donné. Une solution immédiate consisterait à sélectionner tous les mots inconnus du système dans un corpus traitant du thème concerné et à les ajouter directement dans le vocabulaire du système. Toutefois, cette solution n'est pas satisfaisante car elle peut conduire à un afflux de mots sans intérêt réel pouvant dégrader les performances du système. De plus, rien ne peut garantir que les mots à ajouter sont exactement ceux qui seront prononcés dans les documents multimédias à transcrire. En effet, le locuteur peut employer des synonymes des mots sélectionnés ou encore employer des conjugaisons de verbes qui n'auront pas été apprises par le système. Pour ces raisons, nous présentons dans ce rapport le système de filtrage que nous avons mis en place afin de réduire au maximum le nombre de mots hors vocabulaire récupérés tout en sélectionnant les mots les plus pertinents, mots absents des transcriptions.

La première partie de ce document a pour objectif de faire un état de l'art sur les travaux existant dans le domaine de la reconnaissance de la parole tout en insistant sur l'intérêt de l'adaptation thématique des systèmes de reconnaissance de la parole. Ensuite, nous décrivons les données sur lesquelles se basent notre travail, comme les sorties d'un système de transcription automatique de la parole par exemple. Dans une troisième partie, nous présentons les méthodes que nous avons mises en place pour la sélection des mots hors vocabulaire et les résultats obtenus avec ces méthodes. Le stage se terminant 3 semaines après la remise de ce document, une quatrième partie est dédiée aux travaux en cours de réalisation qui ont pour but d'améliorer les résultats obtenus jusqu'à présent. Finalement, nous discutons de nos résultats, des limites de notre travail et des améliorations qui pourraient être mises en place par la suite.

1 État de l'art

Dans cette partie, nous avons pour objectif de faire un état de l'art sur les travaux existant dans le domaine de la reconnaissance automatique de la parole tout en insistant sur l'intérêt de l'adaptation thématique des systèmes de reconnaissance de la parole. Tout d'abord, nous présentons la théorie régissant un système de reconnaissance automatique de la parole (SRAP). Ensuite, nous étudions les travaux sur l'adaptation thématique d'un SRAP, notamment l'adaptation du modèle de langue. Enfin, nous décrivons les recherches menées sur l'adaptation thématique du vocabulaire au sein des systèmes de reconnaissance de la parole.

1.1 Système de reconnaissance automatique de la parole

On utilise les systèmes de reconnaissance automatique de la parole dans le but de transcrire sous forme textuelle les mots prononcés dans un document audio ou vidéo. Ci-dessous est présentée la théorie régissant de tels systèmes.

Le but d'un SRAP est de déterminer la séquence de mots W^* la plus probable, parmi toutes les séquences de mots W réalisables avec les mots du vocabulaire, sachant une observation acoustique Y :

$$W^* = \underset{W}{\operatorname{argmax}}\ \mathrm{P[W|Y]}\ .$$

Grâce à la règle de Bayes, cette formule peut être réécrite sous la forme suivante [Jelinek 1998] :

$$W^* = \underset{W}{\operatorname{argmax}}\ \frac{p(Y|W)\cdot P[W]}{p(Y)}\ ,$$

où P[W] est la probabilité que la séquence de mots W soit prononcée, p(Y) la probabilité que la séquence sonore Y soit émise et p(Y|W) la probabilité de la séquence d'observations acoustiques sachant une séquence de mots testée. Comme Y est donnée, la probabilité p(Y) sera donc la même pour n'importe quelle séquence de mots W testée. On peut alors simplifier l'équation par :

$$W^* = \underset{W}{\operatorname{argmax}}\ \mathrm{p(Y|W)}\cdot \mathrm{P[W]}\ .$$

Ce principe général met en évidence les différents composants d'un système de reconnaissance de la parole :

- un module de caractérisation du signal, qui, à partir d'un signal audio, produit une séquence d'observations acoustiques Y, aussi appelée vecteur acoustique ;
- un vocabulaire qui regroupe les mots connus par le système ainsi que leurs prononciations ;
- un modèle acoustique permettant de calculer la probabilité p(Y|W) ;
- un modèle de langue permettant, quant à lui, de déterminer les probabilités d'obtenir les séquences de mots W.

Dans la suite de cette sous-section, nous décrivons succinctement certains de ces composants.

1.1.1 Vocabulaire

Le vocabulaire définit, pour un système de reconnaissance de la parole, l'ensemble des mots qu'il connaît et qu'il peut donc manipuler. Ceci a d'importantes conséquences car si un mot n'est pas présent dans le vocabulaire, il ne pourra pas apparaître dans une transcription. En général, les mots retenus pour faire partie du vocabulaire sont ceux qui ont une fréquence d'apparition élevée dans un corpus de textes d'apprentissage. Pour des raisons pratiques, et notamment de place mémoire et de coût de calculs, le vocabulaire est en général limité à quelques dizaines de milliers de mots.

À chaque mot du vocabulaire, on associe une prononciation sous la forme de phonèmes. Les phonèmes constituent les plus petites unités de son dans la parole. Le plus souvent, cette action est réalisée grâce à des outils de phonétisation automatique comme ILPho [de Mareüil *et al.* 2000] ou LIA_PHON [Béchet 2001]. Cela permet de faire le lien entre le modèle acoustique (les sons qui sont prononcés dans le document audio) et le vocabulaire

de la langue. Il faut noter que seule la prononciation d'un mot est modélisée et non son sens ou son contexte d'utilisation. Le verbe *« mérite »* et le nom *« mérite »* auront la même prononciation mais un sens différent dans une phrase. Il en est de même pour le terme *« avocat »* qui peut à la fois désigner un métier ou un fruit [Huet 2007]. De plus, chaque item du vocabulaire peut être prononcé de façons différentes en fonction de la personne qui parle ou encore des termes qui le suivent ou le précèdent. Par exemple pour le mot *« clans »*, le *« s »* n'est prononcé que lorsque le mot suivant commence par une voyelle. Pour des raisons pratiques, l'ensemble des prononciations du vocabulaire est représenté sous la forme d'un arbre comme le montre la figure 1 [Lecorvé 2010].

1.1.2 Modèle acoustique

Le modèle acoustique sert à déterminer la probabilité p(Y|W). Pour ce faire, l'outil statistique le plus couramment employé est le modèle de Markov caché ou MMC, qui a maintes fois été utilisé avec succès dans des systèmes de reconnaissance de la parole [Rabiner 1989].

Ce modèle est divisé en plusieurs niveaux comme le présente la figure 2 [Lecorvé 2010]. Au premier niveau, chaque phonème est modélisé par un MMC à trois états, représentant le début, le milieu et la fin du phonème. Les probabilités reliant les observations aux états sont données par un mélange de gaussiennes appartenant au même espace que les vecteurs acoustiques de Y. Au second niveau du modèle, on représente les mots grâce à un arbre lexical contenant l'ensemble des prononciations pour les mots identifiés.

1.1.3 Modèle de langue

Le modèle de langue permet, pour chaque séquence de mots W, de calculer sa probabilité *a priori* P[W]. Concrètement, prenons une séquence de mots $W = w_1 w_2 ... w_n$, où w_i est le mot de rang i dans la séquence. La probabilité P[W] est estimée en calculant pour chaque mot w_i de la séquence sa probabilité d'apparaître en fonction des mots $w_1 ... w_{i-1}$ qui le précèdent :

$$P[W] = P[w_1 w_2 ... w_n] = P[w_1] * \prod_{i=2}^{N} P[w_i | w_1 ... w_{i-1}] \ .$$

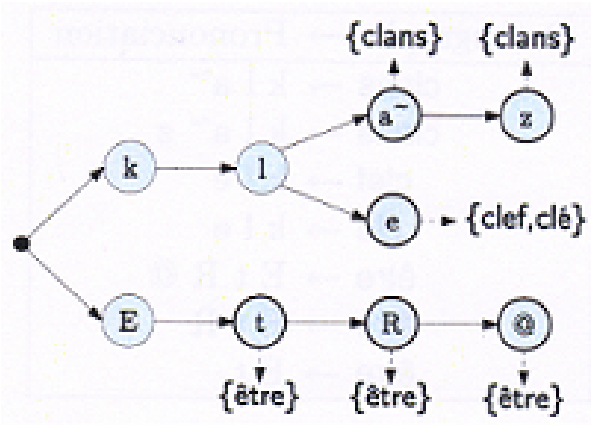


FIGURE 1 – Représentation du lexique phonétisé sous la forme d'un arbre lexical. Cerclés en gras, les nœuds correspondant à la fin de la transcription phonétique d'un mot [Lecorvé 2010]

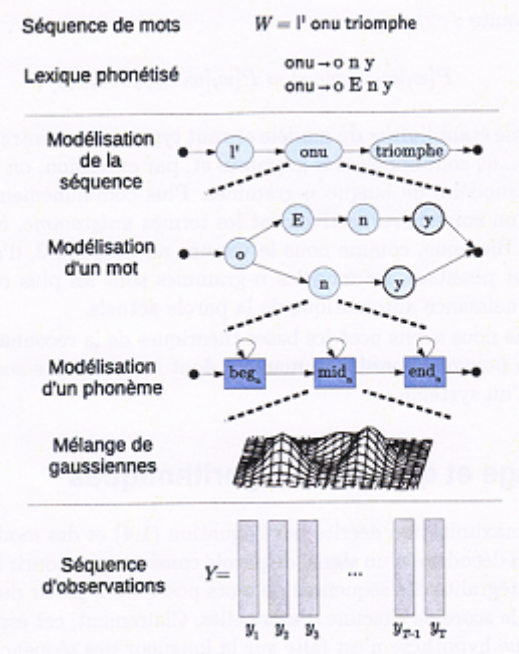


FIGURE 2 – Représentation d'une séquence de mots W sous la forme de modèles de Markov cachés pour le calcul de la vraisemblance p(Y|W) du signal de parole Y [Lecorvé 2010]

Ces probabilités sont estimées à partir d'un corpus d'apprentissage. Cependant, on peut remarquer que si le nombre de prédécesseurs d'un mot est important, l'efficacité de cette formule pourrait être remise en question. C'est pour cette raison que, dans la pratique, on prend l'hypothèse, certes fausse dans la réalité, que seuls les n prédécesseurs de w_i ont un impact sur sa probabilité d'apparition. Le nombre n détermine ainsi l'ordre du modèle. En règle générale n vaut entre 1 et 4. Les séquences de mots $w_{i-n+1}...w_i$ sont appelées n-grammes. Bien qu'il existe d'autres types d'approches statistiques, les n-grammes sont les plus utilisés dans le domaine de la reconnaissance de la parole car ils apportent le meilleur compromis entre performance et coût calculatoire [Jelinek 1976].

1.2 Adaptation thématique

Les systèmes de reconnaissance de la parole apprennent leur vocabulaire et leur modèle de langue sur une grande quantité de textes traitant de sujets variés, appelée corpus d'apprentissage. Cela leur permet d'avoir une connaissance très large et très générale de la langue. Dans la plupart des cas, ces systèmes produisent de bonnes transcriptions. Cependant, si l'on souhaite les utiliser sur des documents multimédias plus spécifiques, traitant de thèmes bien particuliers, tels que l'économie, l'éducation en Suède ou encore des concours de beauté, les résultats obtenus ne sont pas aussi bons. En effet, chacun de ces thèmes contient un vocabulaire qui lui est propre et une façon d'agencer les mots de ce vocabulaire qui lui est propre également. Les travaux actuels sur l'adaptation thématique des systèmes de reconnaissance de la parole cherchent à résoudre ce problème.

Le principe de fonctionnement des méthodes sur l'adaptation thématique se fonde sur les travaux initiés par [Bellegarda 2004]. Il s'agit d'adapter soit le modèle de langue, soit le vocabulaire généraliste, et parfois même, faire les deux comme le montre la figure 3. Afin de générer le modèle de langue ou le vocabulaire spécifique à un thème donné, il faut

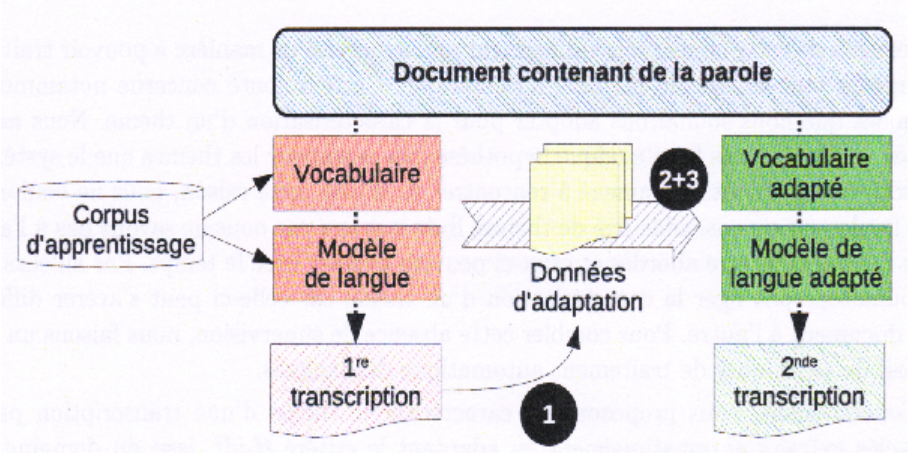


FIGURE 3 – Stratégie d'adaptation thématique d'un système de reconnaissance automatique de la parole pour un document multimédia donné [Lecorvé 2010]

obtenir des documents traitant de ce thème (partie 1.2.1). Ensuite, on peut adapter le modèle de langue avec l'une des trois méthodes existantes présentées en partie 1.2.2. Il est également possible de récupérer des mots hors vocabulaire traitant d'un thème spécifique, en utilisant différents critères. Cette façon de procéder fait l'objet d'une sous-section à part entière (partie 1.3) dans ce document car le stage porte sur ce problème.

1.2.1 Documents d'adaptation

Comme il a été dit précédemment, la première étape, afin de pouvoir adapter un système de reconnaissance de la parole à un thème donné, est de trouver les données qui vont servir pour la réalisation de cette tâche. Ces données sont appelées documents d'adaptation. Ce sont des documents qui traitent du thème concerné.

Il existe deux écoles pour retrouver ces documents. La première consiste à travailler sur un corpus textuel, dit hors-ligne, traitant de sujets variés. Les archives de journaux ou les corpus du français parlé provenant de la campagne ESTER sont de bons exemples de corpus hors-ligne. Le but est de rechercher les documents susceptibles de correspondre à la tâche d'adaptation parmi tous ceux présents dans le corpus. Le problème est qu'on ne peut pas être sûr que ces documents contiennent toutes les informations pertinentes pour un thème donné. La seconde école, à l'inverse, se propose d'exploiter Internet comme un corpus. Les documents se trouvant sur Internet permettent de récupérer beaucoup plus de vocabulaire que les documents hors-ligne. Ils apportent également une plus grande proximité avec la langue parlée [Bulyko *et al.* 2007]. Par exemple, les pronoms personnels des première et deuxième personnes du singulier sont souvent employées [Vaufreydaz *et al.* 1999]. De plus, sur Internet, on peut trouver des documents parlant de n'importe quel sujet et régulièrement mis à jour. Ce sont eux qui sont le plus souvent utilisés pour adapter les systèmes de reconnaissance de la parole [Geutner *et al.* 1998a, Palmer & Ostendorf 2005, Lecorvé 2010], même si d'autres systèmes d'adaptation se basent sur les documents hors-ligne [Seymore & Rosenfeld 1997].

Par la suite, nous allons expliquer le fonctionnement de la récupération de documents sur Internet car c'est ce type d'adaptation qui est le plus choisi et qui fournit les meilleurs

résultats, en dépit des potentielles fautes d'orthographe contenues dans ces documents qui faussent l'apprentissage du modèle de langue et du vocabulaire adaptés.

Il existe deux approches pour obtenir les documents en ligne. L'une consiste à mettre à jour de temps en temps une base de documents hors-ligne. Cette mise à jour se fait en récupérant des documents sur des sites d'actualité [Kemp & Waibel 1998, Allauzen & Gauvin 2005a, Martins *et al.* 2006]. Cette technique a pour principal problème de ne pouvoir s'adapter à des transcriptions qui ne sont pas en relation avec des faits d'actualité. L'autre approche consiste à récupérer les documents par le biais de requêtes sur des moteurs de recherche tels que Google ou Yahoo par exemple [Oger *et al.* 2008]. La recherche de documents s'effectue en cherchant dans une première transcription les mots-clés qui sont représentatifs d'un thème. Une des façons de les obtenir est de se servir d'un critère employé dans le domaine de la recherche d'information, le critère *tf-idf* [Salton 1989]. Ce critère permet d'attribuer à chaque mot de la transcription un score en fonction de sa fréquence d'apparition. Ensuite, des requêtes permettant de récupérer les documents d'adaptation à partir des mots-clés sont générées. L'inconvénient de cette méthode est que le choix des documents retournés est fait par le moteur de recherche.

Une fois l'ensemble des documents d'adaptation récupérés, nous pouvons adapter notre système de reconnaissance de la parole. Nous expliquons dans la partie suivante les méthodes permettant d'adapter le modèle de langue.

1.2.2 Adaptation du modèle de langue

Les systèmes de reconnaissance de la parole, comme nous l'avons dit précédemment, reposent sur un modèle de langue généraliste traduisant les probabilités d'enchaînements des mots. Le problème de ce modèle de langue est qu'il est conçu pour reconnaître le plus de séquences de mots d'une langue. Il n'est pas adapté pour transcrire des documents multimédias traitant d'un thème particulier contenant des mots rares. L'adaptation du modèle de langue consiste ainsi à rendre plus probable les séquences de mots spécifiques à un thème donné.

Il existe trois approches pour adapter un modèle de langue. Une première consiste à ré-estimer complètement le modèle en se servant du corpus généraliste et des documents thématiques trouvés après une première transcription. Cette méthode est peu employée car elle implique de recommencer tout le processus d'apprentissage du système de reconnaissance de la parole, ce qui est extrêmement coûteux en terme de temps. Ce n'est pas une méthode qui peut être utilisée pour transcrire une succession de documents traitant de thèmes différents. Une deuxième approche vise à générer un modèle de langue thématique sur les documents d'adaptation récupérés après une première transcription. Le modèle thématique contient moins d'éléments que le modèle généraliste. Ensuite, les modèles de langue généraliste et thématique sont fusionnés avec une méthode appelée interpolation linéaire [Seymore & Rosenfeld 1997]. Cette méthode a pour effet de créer un modèle de langue adapté au thème voulu, sans avoir à ré-estimer complètement le modèle de langue du système. Cependant, l'interpolation linéaire a comme inconvénient de ne pas pouvoir contrôler les modifications apportées par le mélange des deux modèles. Comme la précédente, la troisième approche vise tout d'abord à générer un modèle de langue thématique. Ensuite, au lieu de mélanger complètement les modèles de langue généraliste et thématique, une autre méthode d'adaptation est utilisée, appelée minimum d'information discriminante (MDI) [Lecorvé 2010]. Elle consiste à ré-estimer dans le modèle de

langue généraliste seulement les séquences de mots, représentatives du thème, présentes dans le modèle de langue thématique. Cette méthode a l'avantage de pouvoir contrôler les probabilités de séquences de mots qui sont re-calculées. Ainsi seules celles qui ont une importance pour le thème seront estimées à nouveau.

1.3 Adaptation du vocabulaire

Si une des façons d'adapter un système de reconnaissance de la parole est de ré-estimer les probabilités du modèle de langue, cette approche ne permet toutefois pas de résoudre les mauvaises transcriptions dues à des mots absents du vocabulaire. Une autre facette de l'adaptation thématique consiste à ajouter des mots qui ne sont pas présents dans le vocabulaire généraliste (on parle de mots hors vocabulaire ou OOV pour *Out-Of-Vocabulary*) du système mais qui pourraient être employés dans les documents audio à transcrire. Il est important de pouvoir trouver ces mots car, en général, ce sont des mots représentatifs du thème, comme des noms de personnes ou encore des termes techniques. De plus, il a été prouvé qu'une mauvaise transcription (les mots mal transcrits sont souvent remplacés par plusieurs mots courts et proches phonétiquement) des mots hors vocabulaire pouvait générer 1.5 à 2 erreurs de transcription en moyenne sur les mots adjacents au mot mal reconnu par le système [Rosenfeld 1995].

La recherche de mots hors vocabulaire a été traitée dans plusieurs domaines et suscite toujours autant d'interrogations. On peut retrouver ce problème bien évidemment dans des travaux sur la reconnaissance de la parole mais aussi en reconnaissance optique de caractères [Bazzi *et al.* 1999] par exemple.

Concrètement, la recherche de mots hors vocabulaire consiste à déterminer un sous-ensemble de mots pouvant être jugés comme les plus pertinents parmi tous les mots que l'on peut trouver en dehors du vocabulaire. Ces mots sont appelés mots candidats. En règle général, les mots hors vocabulaire sont des entités nommées (nom propre par exemple), des noms communs, des verbes ou encore des adjectifs. Alors que les verbes, les adjectifs et les noms communs sont des termes fixes dans une langue, les entités nommées sont, quant à elles, des termes particuliers qui changent avec l'actualité et qui s'avèrent donc difficile à retrouver [Bechet & Yvon 2000]. Il existe différentes façons de récupérer ces mots hors vocabulaire. Certains auteurs décident par exemple d'ajouter tous les OOV qu'ils rencontrent dans un ensemble de documents thématiques [Kemp & Waibel 1998, Schwarm *et al.* 2004]. Cette technique a pour effet d'augmenter considérablement le nombre de mots dans le vocabulaire. Ceci conduit à des taux de reconnaissance moindre et des besoins en espace mémoire et en temps de calcul plus importants. Une autre solution consiste à choisir entre différents critères pour filtrer les mots hors vocabulaire. Tandis que certains auteurs appliquent ces critères sur l'ensemble d'une transcription, d'autres tentent de détecter des zones d'erreur, passages où un mot absent du vocabulaire devrait être présent, afin d'être plus précis et de réduire le volume de mots pouvant être obtenu. Pour ce faire, des études [Wessel *et al.* 2001, Jiang 2004, Burget *et al.* 2008] portent sur le développement de mesures, appelées mesures de confiance, permettant au système de donner une probabilité représentant pour lui la certitude d'avoir reconnu le bon mot. D'autres auteurs utilisent des méthodes [Bazzi 2002, Wang 2009], où ils tentent de repérer les zones erronées pendant la transcription. Ces méthodes sont donc intégrées dans le système de reconnaissance de la parole et s'appliquent surtout au niveau du mo-

dèle acoustique. D'autres encore [Rastrow *et al.* 2009], ont mis en place une approche hybride, combinant les deux techniques citées précédemment, pour détecter les mots hors vocabulaire. Dans la suite de ce document, nous allons présenter quelques critères que nous trouvons pertinents parmi tous ceux rencontrés dans la littérature.

Une fois les mots candidats sélectionnés, le problème qui se pose ensuite est de savoir comment les intégrer au sein du système de reconnaissance de la parole. Ces deux aspects – recherche de mots et intégration dans le SRAP – sont successivement présentés ci-dessous.

1.3.1 Recherche de mots candidats

Afin de rechercher les mots hors du vocabulaire intéressant à prendre en compte, on peut utiliser différents critères.

Critère phonétique. Certains travaux cherchent à déterminer les mots candidats qui pourraient être prononcés dans les documents à transcrire. Dans ce but, certains auteurs tentent de récupérer les erreurs d'une première transcription grâce à une méthode *Hypothesis Driven Lexical Adaptation* [Geutner *et al.* 1998b]. En général, ces erreurs sont transcrites comme une suite de petits mots proches phonétiquement du mot hors vocabulaire. Ensuite, [Geutner *et al.* 1998b] et [Palmer & Ostendorf 2005] proposent de rechercher dans un corpus de textes phonétisés, autre que le corpus utilisé par le SRAP, tous les mots candidats qui sont proches phonétiquement de la séquence mal transcrite. La méthode employée consiste à rechercher une suite de phonèmes (appelée la requête) dans un document audio (appelé document). L'objectif de ce type de recherche n'est pas de retrouver exactement la même suite de phonèmes dans le document, mais plutôt une séquence de phonèmes qui s'en approche le plus [Wechsler *et al.* 1998, Muscariello *et al.* 2009]. Cela est dû au fait que les documents audio contiennent, en général, des parasites (bruits de fond, accent du locuteur, *etc.*). Cette recherche peut être faite en utilisant une mesure de similarité basée sur une adaptation de la formule de la distance d'édition qui est souvent utilisée pour calculer la distance entre deux chaînes de caractères [Muscariello *et al.* 2009, Guinaudeau 2008]. Les mots candidats ainsi trouvés sont ajoutés au vocabulaire. Bien que cette méthode réduise le nombre d'OOV, elle peut avoir des conséquences néfastes sur l'espace mémoire et les temps de calcul à cause du nombre important de mots ajoutés au vocabulaire.

Critère morphologique. On s'intéresse ici à la forme des mots. Les systèmes de reconnaissance de la parole utilisent un vocabulaire fermé, ne pouvant contenir toutes les flexions et dérivations de tous les mots de la langue. Une partie conséquente des OOV est due au fait que le vocabulaire ne contient pas, par exemple, l'ensemble des conjugaisons possibles de tous les verbes ou encore toutes les formes en genre et en nombre des noms et adjectifs. Certains travaux cherchent alors à déterminer les mots candidats en générant toutes les formes fléchies (verbe conjugué, adjectif féminin singulier, *etc.*) des mots présents dans une transcription automatique.

Des auteurs tels que [Geutner *et al.* 1998a] essaient de récupérer, dans un corpus de textes variés, des mots candidats qui sont morphologiquement proches de mots présents dans la transcription. D'autres [Martins *et al.* 2006] proposent l'idée de générer automatiquement des conjugaisons pour les verbes. Cette idée intéressante pourrait être exploitée pour générer toutes les flexions et les dérivations d'un mot. Le risque, une fois encore, est

d'avoir à ajouter trop de mots au vocabulaire mais aussi, de manière erronée, des mots qui n'existent pas dans la langue.

Critère syntaxique. Des auteurs s'intéressent à la façon dont les mots se combinent pour former des phrases. Dans [Oger *et al.* 2008], les auteurs cherchent à trouver les zones d'erreur lors d'une première transcription automatique. Ils déterminent ainsi, à la main, les mots mal transcrits dans les différentes phrases d'une transcription. Ensuite, ils retirent les OOV des phrases et se servent d'un moteur de recherche pour effectuer des recherches de corpus avec les mots restants. Cette approche permet d'obtenir 7.7% des mots hors du vocabulaire. Nous pouvons nous demander si ce taux faible n'est pas provoqué par la qualité des documents récupérés qui, parfois, contiennent des fautes d'orthographe par exemple. Une autre interrogation est envisageable quant à la pertinence des documents retournés par le moteur de recherche par rapport à la séquence de mots utilisés pour faire la recherche.

Critère thématique. Ici, on s'intéresse à l'importance des mots par rapport au thème de la transcription. Des études récentes cherchent à déterminer les mots hors vocabulaire par rapport au thème d'une première transcription. Dans [Marin *et al.* 2009], les auteurs proposent de trouver les mots représentatifs du thème abordé dans la transcription automatique, grâce au score *tf-idf*. Ensuite, ils récupèrent un corpus d'adaptation contenant des textes en adéquation avec les mots-clés en générant des requêtes sur Internet. L'ensemble des mots hors vocabulaire de ce corpus forme la liste des mots candidats qu'ils ajoutent au SRAP. Toutefois, l'insertion de ces mots dans le SRAP n'a abouti à aucune amélioration de leurs transcriptions.

Une démarche exploratoire étudiée dans [Lecorvé 2010] cherche également à déterminer les mots candidats par rapport à leur importance dans le thème de la transcription en choisissant différents critères de filtrage parmi ceux vus précédemment. Dans ces travaux expérimentaux, l'auteur cherche à ne récupérer que les OOV prononcés dans la transcription automatique afin de réduire le nombre de mots à ajouter au vocabulaire, et donc ne pas trop affecter la fiabilité du système.

Critère temporel. Ici, on cherche à récupérer des mots hors vocabulaire sur un corpus qui est temporellement proche de la transcription [Auzanne *et al.* 2000, Federico & N.Bertoldi 2001]. Pour ce faire, les auteurs utilisent des sites Internet d'actualité. Ils vont alors rechercher un ensemble de documents qui sont proches (quelques jours d'intervalles) avec la date de création du document audio que l'on cherche à transcrire. D'autres encore [Allauzen & Gauvain 2003], se servent, comme précédemment, des informations thématiques d'une première transcription ainsi que de la date des documents pour générer un corpus d'adaptation à partir de sites Internet d'actualité. L'ensemble des mots hors vocabulaire qu'ils obtiennent avec ces textes forme la liste des mots candidats qu'ils ajouteront au SRAP.

Le choix de ces différents critères pour sélectionner les mots hors vocabulaire est une étape importante pour l'adaptation thématique du vocabulaire au sein d'un système de reconnaissance de la parole. Dans le meilleur des cas, les méthodes les plus performantes arrivent à récupérer 50% de mots hors vocabulaire pertinents. Mais en pratique, l'efficacité des différentes méthodes dépend de la langue et des thèmes à traiter [Lecorvé 2010].

1.3.2 Intégration des nouveaux mots dans le système de reconnaissance automatique de la parole

L'étape suivant la sélection des mots candidats est leur intégration dans le système de reconnaissance de la parole. Pour réaliser cette intégration, on peut agir en deux temps. Une première étape consiste à générer, pour chaque mot à ajouter, ses prononciations. Ceci peut se faire, comme nous l'avons vu au début de ce document (sous-section 1.1.1), grâce à des outils automatiques. La seconde étape consiste, quant à elle, à intégrer ces nouveaux mots au sein du modèle de langue. Dans la littérature, on retrouve, pour ce second point, différentes méthodes similaires à celles employées pour l'adaptation du modèle de langue.

Certains auteurs [Kemp & Waibel 1998, Allauzen & Gauvin 2005a] ont ainsi choisi de ré-apprendre complètement le modèle de langue en se basant sur le corpus d'apprentissage ainsi que sur les documents où les mots candidats ont été trouvés. D'autres [Geutner *et al.* 1998a] préfèrent construire un modèle de langue à partir des documents où ils ont trouvé les mots candidats et, seulement ensuite, mélanger ce dernier avec le modèle de langue généraliste en se servant de l'interpolation linéaire. Bien que la seconde méthode soit intéressante car elle ne nécessite pas de régénérer un modèle de langue complet pour le SRAP, elle ne permet pas de définir, pour chaque mot, des probabilités qui sont propres à son contexte d'utilisation. La première méthode n'a pas ce défaut, mais elle implique le risque que toutes les probabilités d'enchaînements des mots ne soient pas prises en compte car le corpus d'adaptation est plutôt petit et l'apparition des mots à ajouter peut être rare. Il faut noter également que ces deux méthodes induisent forcément de refaire une transcription pour que les nouveaux mots soient pris en compte. C'est pour cela que certains auteurs ont réalisé des travaux sur une adaptation *a posteriori* qui tente de corriger les erreurs de transcription par le remplacement du terme mal reconnu [Palmer & Ostendorf 2005].

Les deux méthodes précédentes fonctionnent pour des SRAP à vocabulaire fermé et obligent ainsi à ré-estimer le modèle de langue. Cependant, il existe des SRAP à vocabulaire ouvert qui peuvent se passer de cette étape. En effet, un vocabulaire ouvert possède, en plus des mots qui le composent, une ou plusieurs classes symbolisant les OOV. Ces classes sont considérées comme des parties intégrantes du vocabulaire et sont donc prises en compte dans le calcul des probabilités des séquences de mots du modèle de langue. Dans le plus simple des cas, une seule classe est considérée [Ohtsuki *et al.* 2005]. Lorsqu'un nouveau mot lui est ajouté, il apparaît pour le système comme une variante de prononciation. L'emploi d'une seule classe a pour effet de considérer tous les OOV comme étant interchangeables, ce qui est faux en réalité. C'est pour cela qu'en général plusieurs classes sont utilisées [Allauzen & Gauvin 2005b, Oger *et al.* 2008]. Toutefois le découpage en plusieurs classes est difficile car une classe ne peut représenter qu'un type de mots (verbe, nom, *etc.*) sans prendre en compte le sens et le contexte des mots qui y sont ajoutés.

1.4 Conclusion

Aujourd'hui, les documents multimédias sont une mine d'informations difficile d'accès en raison de leur nombre et du manque de techniques pour les exploiter de manière optimale. La première partie de ce document a présenté un domaine de la recherche qui vise à récupérer la sémantique des documents multimédias par le biais de transcriptions. L'état de l'art réalisé ici montre qu'il est vital d'adapter un SRAP au thème du document à transcrire. Pour cela, on peut procéder de deux façons. La première vise à recalculer les probabilités des séquences de mots par rapport au thème concerné. La seconde consiste à ajouter au vocabulaire de nouveaux mots importants pour caractériser un thème. Il existe ainsi plusieurs critères pour sélectionner les mots candidats. Dans cette dernière partie, nous allons présenter comment ces critères peuvent être utilisés dans le cadre de notre stage sur l'adaptation thématique du vocabulaire.

Comme point de départ pour notre étude, nous ne disposons que des sorties d'un système de reconnaissance de la parole. Or, nous avons besoin d'une base de mots hors vocabulaire pour pouvoir faire notre adaptation. L'utilisation du critère thématique combiné aux sorties du SRAP semble être pertinente, comme première étape, pour récupérer un ensemble de mots hors vocabulaire en accord avec les thèmes des différentes transcriptions. Cependant la masse de mots récupérée par une telle méthode risque d'être trop importante et d'avoir des répercussions sur la qualité des transcriptions. Il paraît alors indispensable de réduire cet ensemble d'OOV. Pour cela, l'utilisation des critères phonétique, morphologique et syntaxique s'avère adaptée puisqu'ils peuvent agir comme des filtres en ne sélectionnant que les mots qui semblent pertinents dans les transcriptions.

Nous avons fait le choix, dans le cadre de notre stage, d'utiliser les différents critères présentés dans la partie 1.3.1 pour, à tour de rôle, sélectionner et filtrer les mots candidats et ne récupérer que ceux qui sont essentiels. Notre méthode originale effectue un double filtrage phonétique et grammatical sur une liste de mots candidats obtenue à partir des corpus générés par G. Lecorvé pendant sa thèse sur l'adaptation des modèles de langue. Ainsi, avant de décrire en détail les travaux que nous avons entrepris, nous présentons dans la section suivante l'ensemble des outils dont nous disposons.

2 Données

Notre objectif est d'obtenir une liste de mots hors vocabulaire la plus petite possible et contenant le plus d'OOV pertinents pour une transcription automatique générée par un système de reconnaissance de la parole. Nous exposons ici, l'ensemble des outils mis à notre disposition. Nous présentons, tout d'abord, le système de reconnaissance de la parole que nous allons utiliser. Puis nous décrivons les corpus d'adaptation thématique dont nous disposons.

```
20030418_0700_0800_inter_dga 1 159.850000 0.150000 <s> 0.99
20030418_0700_0800_inter_dga 1 160.000000 0.240000 en 0.0498660978623291
20030418_0700_0800_inter_dga 1 160.240000 0.470000 irak 0.99
20030418_0700_0800_inter_dga 1 160.710000 0.160000 l' 0.125840757974992
20030418_0700_0800_inter_dga 1 160.870000 0.460000 ancien 0.912788315532362
20030418_0700_0800_inter_dga 1 161.330000 0.470000 régime 0.132970096443191
20030418_0700_0800_inter_dga 1 161.800000 0.130000 s' 0.506146786177548
20030418_0700_0800_inter_dga 1 161.930000 0.370000 effrite 0.323523700754101
20030418_0700_0800_inter_dga 1 162.300000 0.220000 chaque 0.526804778637689
20030418_0700_0800_inter_dga 1 162.520000 0.240000 jour 0.518953633337861
20030418_0700_0800_inter_dga 1 162.760000 0.050000 un 0.524419544705337
20030418_0700_0800_inter_dga 1 162.810000 0.170000 peu 0.519172545762208
20030418_0700_0800_inter_dga 1 162.980000 0.420000 plus 0.525775318282395
20030418_0700_0800_inter_dga 1 163.400000 0.180000 les 0.530330752556494
20030418_0700_0800_inter_dga 1 163.580000 0.420000 américains 0.515427659419056
20030418_0700_0800_inter_dga 1 164.000000 0.160000 ont 0.552558837817575
20030418_0700_0800_inter_dga 1 164.160000 0.380000 arrêté 0.482351394750825
20030418_0700_0800_inter_dga 1 164.540000 0.280000 hier 0.535576866973124
20030418_0700_0800_inter_dga 1 164.820000 0.210000 un 0.511604632886345
20030418_0700_0800_inter_dga 1 165.030000 0.370000 autre 0.534166860619067
20030418_0700_0800_inter_dga 1 165.400000 0.240000 demi 0.404158557139464
20030418_0700_0800_inter_dga 1 165.640000 0.310000 frère 0.443632934746254
20030418_0700_0800_inter_dga 1 165.950000 0.100000 de 0.500523879770746
20030418_0700_0800_inter_dga 1 166.050000 0.300000 saddam 0.545234532204805
20030418_0700_0800_inter_dga 1 166.350000 0.360000 hussein 0.500523879770746
20030418_0700_0800_inter_dga 1 166.710000 0.060000 </s> 0.0457584119755515
20030418_0700_0800_inter_dga 1 166.750000 0.160000 <s> 0.99
20030418_0700_0800_inter_dga 1 166.910000 0.620000 barzane 0.904190733280068
20030418_0700_0800_inter_dga 1 167.530000 0.240000 al 0.755154701031801
20030418_0700_0800_inter_dga 1 167.770000 0.850000 tikriti 0.950737442788115
20030418_0700_0800_inter_dga 1 168.620000 0.020000 </s> 0.705892143819917
20030418_0700_0800_inter_dga 1 168.580000 0.140000 <s> 0.99
```

FIGURE 4 – Sortie textuelle du système de transcription automatique de la parole

2.1 Système de reconnaissance de la parole

Nous travaillons avec un système de reconnaissance de la parole appelé IRENE [Gravier *et al.* 2005]. Son principe de fonctionnement est similaire à ce que nous avons décrit dans la section 1.1 de ce document. Concrètement, IRENE prend en entrée un document audio, dans notre cas des émissions de radio, et rend à la fin du processus plusieurs sorties exploitables. On peut ainsi obtenir une transcription sous forme de texte (figure 4) et une transcription phonétique (figure 5). Nous ne décrivons que ces dernières car ce sont celles que nous utilisons. La sortie textuelle donne le nom du segment concerné (« 20030418_0700_0800_inter_dga » pour la $3^{ème}$ ligne par exemple), le temps dans le document audio auquel les mots sont prononcés (« 160.240000 0.470000 »), les mots transcrits (« irak ») et une mesure de confiance (« 0.99 ») donnée par le système et qui lui permet de faire une estimation sur le fait qu'un mot soit correct ou non. Les *« <s> »* et *« </s> »* représentent respectivement le début et la fin d'un groupe de souffle. Un groupe de souffle est un ensemble de mots prononcés entre deux blancs (respiration par exemple). La sortie phonétique donne, quant à elle, la séquence de phonèmes dite avec le moment où elle a été prononcée, puis les mots transcrits correspondants avec également les moments de début et de fin de prononciation.

Pour évaluer les performances des systèmes de transcription de la parole, nous disposons de mesures telles que le WER (*Word Error Rate*) ou encore le LER (*Lemma Error Rate*). Le WER indique le taux de mots incorrectement reconnus par rapport à ce qui est dit dans le document audio. Plus le taux est faible, meilleure est la reconnaissance. Pour calculer le LER, on lemmatise la transcription. Pour cela, on ramène les formes fléchies (conjuguées, plurielles) à des formes standards (infinitif, singulier) appelées lemmes. Par exemple, on transforme tous les verbes à l'infinitif et tous les noms au masculin singulier. Ensuite on calcule le taux de lemmes incorrectement reconnus par rapport à la transcription de référence lemmatisée.

2.2 Corpus d'adaptation thématique

En plus du système de reconnaissance de la parole mis à notre disposition, ainsi que des diverses mesures permettant d'en évaluer les sorties, nous disposons, pour nos expérimentations, de documents audio issus du corpus ESTER. Ce corpus rassemble un ensemble d'émissions de radio, datant de 1998 à 2004, accompagnées de leur transcription textuelle – appelée *transcription initiale* par la suite. Ces émissions traitent de sujets variés. Pour sa thèse, G. Lecorvé a sélectionné 6 heures d'enregistrement qu'il a découpé à la main en segments thématiques indépendants – par la suite, nous référons à ces documents par le terme *segment*. Nous réutilisons le découpage en ensembles de développement, sur lequel nous effectuons nos expériences, et de test, sur lequel nous validons nos travaux, effectués par G. Lecorvé. Au total nous disposons de 42 segments dans l'ensemble de développement et de 37 segments dans celui de test.

Nous avons également accès à des corpus d'adaptation pour chacun des segments des documents audio mis à notre disposition. Ces corpus proviennent des travaux de G. Lecorvé sur l'adaptation thématique des modèles de langue et sont générés grâce à un score *tf * idf* adapté aux particularités des transcriptions pour faire émerger les mots caractéristiques des thèmes des divers segments. Une fois l'ensemble des termes clés récupérés, il a généré des requêtes pour interroger des moteurs de recherche tels que Yahoo ! ou Bing afin d'obtenir un ensemble de pages web. Ensuite, ces pages ont été normalisées, c'est-à-dire que tout le contenu non textuel a été retiré (balise HTML, ponctuations, *etc.*). Pour plus de détails, nous invitons le lecteur à se référer à [Lecorvé 2010].

Ces corpus sont la première étape de notre travail. En effet, ils représentent la mise en application du critère thématique décrit à la section 1.3.1. À partir de ces corpus, nous récupérons une liste d'une dizaine de milliers de mots hors vocabulaire pour chaque segment que nous allons par la suite épurer.

```
phones_sequence 0 0.0999999999999943 t
phones_sequence 0.0999999999999943 0.199999999999989 i
phones_sequence 0.199999999999989 0.310000000000002 t
phones_sequence 0.310000000000002 0.340000000000003 e
phones_sequence 0.340000000000003 0.370000000000005 9
phones_sequence 0.370000000000005 0.47999999999999 a~
phones_sequence 0.47999999999999 0.590000000000003 f
phones_sequence 0.590000000000003 0.629999999999995 E
phones_sequence 0.629999999999995 0.680000000000007 t
phones_sequence 0.680000000000007 0.740000000000009 S
phones_sequence 0.740000000000009 0.800000000000011 e
phones_sequence 0.800000000000011 0.909999999999997 d
phones_sequence 0.909999999999997 0.97999999999999 j
phones_sequence 0.97999999999999 1.03 l
phones_sequence 1.03 1.11000000000001 e
phones_sequence 1.11000000000001 1.25 Z
phones_sequence 1.25 1.30000000000001 9
phones_sequence 1.30000000000001 1.44999999999999 a~
word 0.00 0.17 <s> 0.990000
word 0.17 0.34 qui 0.547239
word 0.34 0.57 était 0.475453
word 0.57 0.68 en 0.712676
word 0.68 0.89 fait 0.761769
word 0.89 1.00 les 0.712676
word 1.00 1.69 dirigeants 0.761769
word 1.69 1.74 </s> 0.474446
```

FIGURE 5 – Sortie phonétique du système de transcription automatique de la parole

3 Sélection de mots candidats

La quantité de mots hors vocabulaire contenue dans chaque corpus d'adaptation est importante. Ainsi, l'ajout d'un tel volume de mots, dans le système, dégraderait surement les transcriptions par l'insertion de nouvelles erreurs. Afin de réduire la taille de ces ensembles de mots, nous allons effectuer deux filtrages successifs. Notons toutefois, que les filtrages sont effectués sur l'ensemble d'une transcription et non sur des sous-parties, de cette dernière, reconnues erronées par le système. En effet, nous ne prenons pas en compte les mesures de confiance générées par le SRAP IRENE car elles ne sont pas assez précises.

Tout d'abord, nous présentons dans cette partie les mesures préliminaires réalisées dans le but de déterminer l'impact de l'insertion d'une quantité plus ou moins conséquente de mots hors vocabulaire dans le SRAP IRENE. Ensuite, nous décrivons les mesures utilisées pour analyser l'efficacité de nos filtrages. Puis, nous présentons les filtres mis en place pour réduire la masse d'OOV de chaque corpus. Pour chacun de ces filtres, nous expliquons leur principe de fonctionnement, puis les méthodes utilisées pour leur réalisation et finalement les résultats obtenus.

3.1 Mesures préliminaires

Dans un premier temps, nous avons étudié le potentiel des corpus d'adaptation générés pour chaque segment, c'est-à-dire, le nombre d'OOV pertinents possibles de sélectionner. Ainsi, nous avons estimé que l'ensemble des transcriptions contenait 505 mots hors vocabulaire. Parmi ces mots absents, la proportion d'OOV pertinents qu'il est possible d'obtenir s'élève à 252 mots sur la totalité des corpus d'adaptation. Cela représente environ 50% de la masse totale d'OOV manquants dans les transciptions.

Après avoir déterminé le potentiel des corpus d'adaptation, nous avons souhaité connaître l'impact, sur les sorties du système de reconnaissance de la parole, de l'ajout des mots hors vocabulaire pertinents et de l'ensemble des mots hors vocabulaire des corpus.

Tout d'abord, nous avons sélectionné, dans le corpus d'adaptation de chaque segment, tous les mots hors vocabulaire pertinents que nous pouvions retrouver. Ces mots ont été intégrés dans le système, en ajoutant les nouveaux mots dans le vocabulaire, puis en adaptant le modèle de langue par interpolation linéaire (voir partie 1.3.2), pour générer de nouvelles transcriptions. La figure 2 (en annexe A) regroupe les résultats obtenus pour cette expérience. Nous constatons, pour certains segments (segments 1 et 6 par exemple), une légère amélioration des sorties du SRAP avec une diminution des scores WER et LER pour les nouvelles transcriptions par rapport aux transcriptions initiales. Dans d'autres cas (segments 3 et 4), les valeurs du WER et du LER sont restées identiques car, bien que les OOV aient été ajoutés dans le système, ils n'ont pas été retenus pour faire partie de la transcription. Nous remarquons également, dans des cas plus rares, l'augmentation de certaines mesures du WER (segments 20 et 26). L'analyse des sorties concernées montrent que les mots hors vocabulaire ont été pris en compte par le système, mais que leur intégration a généré de nouvelles erreurs sur la reconnaissance des mots qui les entourent. Dans le même esprit, nous avons voulu réaliser une expérience où nous ajoutons la totalité des mots hors vocabulaire présents dans les différents corpus d'adaptation. Les ensembles ainsi constitués sont d'une dizaine de milliers de mots par segment. Pour des raisons techniques, le système ne supporte pas l'insertion de ce volume conséquent de mots. Nous

ne pouvons donc connaitre l'impact réel sur les transcriptions. Cependant, nous pensons probable qu'une des conséquences soit la perte en qualité des sorties du SRAP. En effet, intégrer de nouveaux mots dans le système implique la génération d'un nouveau modèle de langue et donc d'amoindrir les probabilités d'enchainement des séquences de mots du modèle d'origine avec les probabilités d'enchainement des séquences contenant les OOV. Les résultats préliminaires nous permettent donc de conclure sur l'importance d'un filtrage afin de ne pas provoquer de nouvelles erreurs dans les transcriptions. Le but est alors de réduire au minimum cette liste de mots candidats, l'idéal étant de ne sélectionner que les mots hors vocabulaire pertinents pour s'approcher des résultats de la première expérience. Pour effectuer cette sélection, nous avons employé deux critères étudiés dans la partie 1.3.1 qui nous ont permis de réaliser deux systèmes de filtrage, l'un phonétique et l'autre grammatical, que nous présentons respectivement dans les parties 3.3 et 3.4. mais dans un premier temps, nous décrivons les mesures utilisées pour évaluer leur performance.

3.2 Évaluation

Les mesures donnant un retour précis sur la pertinence des ensembles de mots que nous intégrons dans le système sont le WER et le LER car ils permettent d'évaluer les sorties du SRAP. Pour des raisons de temps mais aussi à cause de problèmes techniques, nous n'avons pas pu réaliser ces mesures. Nous avons utilisé d'autres métriques, que nous présentons ici, pour jauger de la qualité des filtrages mis en place.

Le rappel est défini par le nombre de mots hors vocabulaire pertinents récupéré par rapport à la totalité des mots hors vocabulaire pertinents qu'il est possible d'obtenir. Ainsi pour chaque segment, on calcule :

$$Rappel_i = \frac{\text{Nombre d'OOV pertinents récupérés}}{\text{Nombre d'OOV pertinents attendus}} .$$

Après avoir calculé le rappel pour chaque segment, on calcule le rappel global en faisant la moyenne de tous les rappels pour les n segments.

$$Rappel = \frac{\sum_{i=1}^{n} Rappel_i}{n} .$$

La précision correspond, pour un segment donné, au nombre de mots hors vocabulaire pertinents obtenus par rapport à celui sélectionné :

$$Précision_i = \frac{\text{Nombre d'OOV pertinents récupérés}}{\text{Nombre total d'OOV récupérés}} .$$

Comme pour le rappel, la précision globale se calcule en faisant la moyenne des précisions des n segments utilisés :

$$Précision = \frac{\sum_{i=1}^{n} Précision_i}{n} .$$

La section suivante présente le filtre phonétique mis en place, ainsi que les résultats évalués à l'aide du rappel et de la précision.

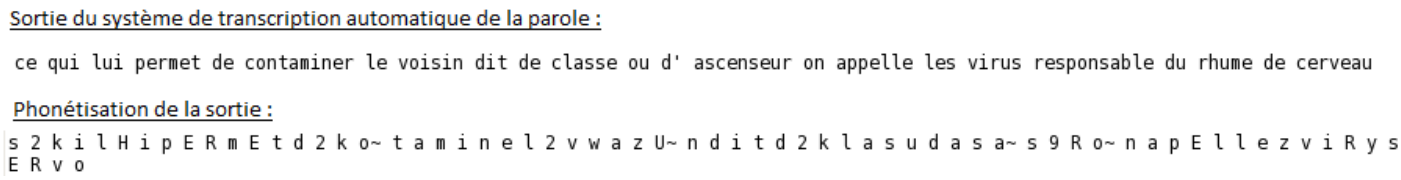


FIGURE 6 – Sortie du système de transcription automatique de la parole phonétisée

3.3 Filtre phonétique

Dans l'optique de nous focaliser sur les seuls mots intéressants à ajouter au système de reconnaissance de la parole, afin de s'approcher des résultats des expériences préliminaires, nous présentons le premier filtre réalisé. Il s'agit du filtre phonétique.

3.3.1 Principe

Comme nous l'avons expliqué dans la partie 1.3.1, le filtrage phonétique consiste à rechercher une petite suite de phonèmes, communément appelée la requête, dans une plus grande suite de phonèmes, appelée le document. Nous ne cherchons pas à retrouver exactement la suite de phonèmes de notre requête dans le document. En effet, la requête est phonétisée à partir d'un texte, elle n'est donc pas altérée par des bruits de fond ou autres parasites. Le document est, quant à lui, une transcription phonétique d'une bande sonore. Cela implique que différents éléments tels que les marques d'hésitation, ou les accents des locuteurs, peuvent perturber la prononciation des termes recherchés. De plus, comme les mots inconnus sont transcrits par des séquences de mots plus courtes mais phonétiquement proches, de nouvelles dégradations dans la phonétisation du document audio sont ajoutées. Dans notre cas, nous récupérons pour chaque segment l'ensemble des mots hors vocabulaire présents dans son corpus d'adaptation. Chacun de ces mots est alors phonétisé grâce à LIA_PHON [Béchet 2001]. LIA_PHON est un système développé par le Laboratoire Informatique d'Avignon permettant de phonétiser des textes. Nous avons choisi d'utiliser cet outil car, contrairement à d'autres outils de phonétisation tel que ILPho [de Mareüil *et al.* 2000], ce système n'est pas basé sur un vocabulaire fermé mais sur un ensemble de règles, ce qui constitue un réel avantage car nous ne connaissons pas, à l'avance, les mots que nous allons intégrer. Chacun des mots candidats phonétisés constitue alors une requête. Le document est quant à lui la transcription phonétisée du segment. Cette sortie phonétisée ne contient aucun signe de ponctuation ou de séparateur (espace) pour différencier les mots comme le montre la figure 6. Au final, notre objectif est de rechercher dans notre document la séquence de phonèmes qui minimise le plus la distance avec la requête. Cette distance est une adaptation de la distance d'édition que nous décrivons dans la section suivante. Puis, nous montrons les résultats obtenus grâce à cette méthode.

3.3.2 Méthode de distance d'édition

Le principe de la méthode de la distance d'édition que nous utilisons est le suivant. Pour chacune des requêtes, nous cherchons une correspondance phonétique la plus proche possible d'une sous-chaîne de phonèmes dans le document. Pour ce faire, nous calculons, pour chaque phonème de la requête, le coût d'une transformation (omission, substitution,

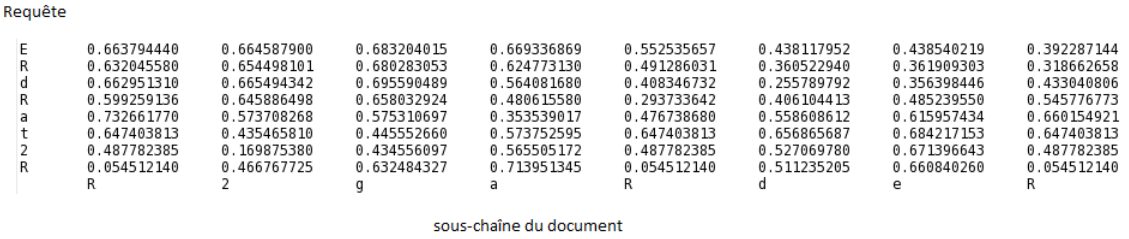


FIGURE 7 – Matrice des scores pour chaque phonème entre une requête et une sous-chaîne du document

insertion) par rapport à une sous-chaîne de phonèmes du document. Ce calcul est effectué sur toutes les sous-chaînes du document. La meilleure correspondance est déterminée par le coût le plus faible entre la suite de phonèmes de la requête et une sous-chaîne du document. En effet, un score faible implique qu'il y a eu peu de transformations, donc que les deux suites de phonèmes se ressemblent. Par exemple la figure 7 nous montre que les deux ensembles de phonèmes sont semblables avec un score de 0.39 et que 3 phonèmes diffèrent entre les deux chaînes (ici, 3 substitutions).

Nous avons repris la distance d'édition réalisée par Camille Guinaudeau pendant son stage de Master 2 [Guinaudeau 2008]. L'algorithme initial de distance d'édition permet de donner un score de similitude entre un document et une requête en fonction de points de départ préétablis. Les modifications apportées par C. Guinaudeau autorisent la recherche de similarités entre n'importe quelles sous-chaînes du document et une requête. Ainsi, la requête peut se situer n'importe où dans le document. Pour cela, elle a modifié le programme original pour ajouter automatiquement de nouveaux points de départ le long de la séquence de phonèmes du document.

Formellement, on remplit une matrice des coûts $d(i, j)$ avec $1 \leq i \leq Longueur_document$ et $1 \leq j \leq Longueur_requête$. Nous définissons alors $L(i, j)$ comme la longueur du chemin parcouru entre un point de départ $(1, j)$ jusqu'au point (i, j) et $D(i, j)$ comme la distance accumulée du point de départ $(1, j)$ au point (i, j). Nous définissons également $W(i, j)$ comme étant le coût moyen : $\frac{D(i,j)}{L(i,j)}$.

Comme l'explique C. Guinaudeau, le début d'une correspondance, entre la requête et le document, peut être modifiée pour créer de nouveaux points de départ. Concrètement, un nouveau point de départ sera admis si le coût d'un appariement entre la requête et le document au point $(1, j)$ est inférieur au coût de transformation entre la requête et le document en ce point. Si ce n'est pas le cas, alors nous continuons d'augmenter le coût de la correspondance entre la requête et le document. Formellement, nous avons :

$$\text{Pour } i = 1 \text{ et } 1 \leq j \leq N : \begin{cases} \text{D(1,j)} = \text{d(1,j) si d(1,j)} < \text{W(1,j)} \\ \text{L(1,j)} = 1 \\ \\ \text{D(1,j)} = \text{D(1,j-1)} + \text{d(1,j) sinon} \\ \text{L(1,j)} = \text{L(1,j-1)} + 1 \end{cases}$$

Pour $i \neq 1$, on calcule chaque chemin en cherchant à minimiser, pour chaque point (i, j), le coût $W(i, j)$:

$$W(i,j) = min\left[\frac{Co\hat{u}t_{insi}}{L(i-1,j)+1}, \frac{Co\hat{u}t_{subi,j}}{L(i-1,j-1)+1}, \frac{Co\hat{u}t_{omisj}}{L(i,j-1)+1}\right] ,$$

avec les différents coûts définis à partir de scores de confusion. Pour plus de détails sur la mise en place technique de ces changements, nous invitons le lecteur à se référer à [Guinaudeau 2008].

Le programme que nous avons repris donne, en sortie, juste un score de similarité entre une requête et la sous-chaîne la plus proche dans un document. Cependant, pour la poursuite de nos travaux, nous avons besoin d'avoir plus d'informations sur les sorties du filtre. La première amélioration que nous avons apportée est de retourner l'ensemble des sous-chaînes du document qui sont reconnues, par la méthode, comme étant similaire à la requête, et cela en fonction d'un score donné. En effet, nous ne connaissons pas les zones erronées d'une transcription, aussi une recherche sur toutes les sous-chaînes possibles est donc entreprise. Puis, seules les parties de la transcription, avec un score inférieur à celui demandé, sont conservées. Dans la deuxième modification réalisée, nous cherchons à retrouver les séquences de mots correspondantes aux séquences de phonémes récupérées précédemment. En effet, nous disposons pour une transcription (comme le montre la figure 6), à la fois, de sa forme textuelle (où chaque mot est séparé par un espace), et de la séquence de phonèmes qui lui est associée (où aucune distinction entre les mots n'est faite). Ainsi, nous avons mis en place une fonction d'association entre les phonèmes de la séquence du document, reconnus comme étant similaires à la requête, et la suite de mots correspondants dans sa transcription textuelle. Pour ce faire, à partir de chaque point de départ $(1, j)$, repéré dans le document, nous parcourons dans la matrice $d(i, j)$ le chemin contenant les coûts minimums pour atteindre le point (i, j). Nous collectons, dans le même temps, les phonèmes correspondants présents sur ce trajet – par la suite, la séquence de phonèmes ainsi obtenue est désignée par S_1. Ensuite, nous phonétisons, grâce à LIA_PHON, la transcription textuelle en gardant la séparation en mots. Puis nous cherchons la suite de phonèmes S_1 associée dans celle générée précédemment. Une fois la séquence identifiée, nous récupérons la séquence de mots qui y correspond car nous avons gardé la séparation en mots. Il faut noter que, pour le filtre phonétique en lui même, ces améliorations n'ont pas d'incidence. En fait, elles permettent de fournir les valeurs d'entrées (mot candidat / séquence de mots) à notre filtre grammatical que nous présentons dans la suite de ce document (partie 3.4).

Avec les modifications décrites ci-dessus, nous sélectionnons pour un seuil donné un ensemble de mots candidats. En fonction du seuil que nous fournissons, nous obtenons plus ou moins de mots candidats. Nous cherchons alors à déterminer le seuil optimal pour récupérer le moins de mots candidats possible en ayant le plus de mots hors vocabulaire pertinents.

3.3.3 Seuil et résultats

La méthode de filtrage phonétique que nous avons mis en place fournit un score sur la probabilité que deux séquences de mots soient similaires. Ainsi, afin de déterminer la valeur de seuil à choisir pour avoir un ensemble de mots le moins imposant possible et contenant le plus de mots hors vocabulaire pertinents, nous calculons les indices de rappel et de précision pour différents niveaux de seuils (figure 8). À partir de la courbe représentative des performances de notre filtre, nous choisissons le seuil pour lequel nous

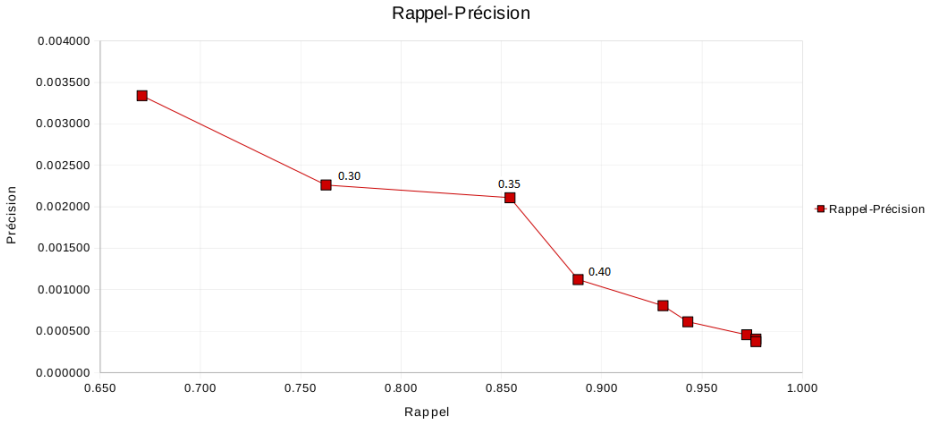


FIGURE 8 – Courbe rappel-précision permettant de déterminer le seuil du filtre phonétique

pensons avoir le meilleur compromis entre le nombre d'OOV pertinents récupérés et la volume total de mots hors vocabulaire sélectionné. Puis nous commentons les résultats ainsi obtenus pour notre valeur de seuil.

Idéalement, nous aimerions obtenir un seuil avec un rappel fort et une précision élevée (l'idéal étant de ne récupérer que les mots hors vocabulaire pertinents). D'après la courbe (figure 8), le seuil qui nous semble intéressant est celui à 0.35 car nous constatons, qu'après ce point, la courbe s'effondre. De plus, nous étudions les seuils de niveaux inférieurs et supérieurs à 0,35, respectivement 0,30 et 0,40. Ainsi, avec un seuil de 0.30, nous sélectionnons 72,7% d'OOV pertinents avec un ensemble de mots candidats réduit à 11.6% de son volume d'origine. Bien que le volume de mots candidats semble d'une taille acceptable (moins de 1000 mots par segment), le nombre d'OOV pertinents obtenu est faible comparé à celui récupéré pour un seuil de 0.35 (78.78%) et de 0.40 (80.1%). Pour un seuil de 0.40, nous diminuons la taille de l'ensemble de mots sélectionnés à seulement 25% de son volume d'origine alors que pour un seuil de 0.35, nous sommes à 14.75% de l'ensemble d'origine. L'analyse de ces différents niveaux de seuils nous conforte dans le choix de 0.35, que nous considérons par la suite comme le seuil optimal pour notre filtre phonétique.

Notre seuil étant défini, nous pouvons étudier les résultats obtenus par notre filtre (tableau 1) sur un ensemble de 27 segments. C'est une partie de l'ensemble de développement que nous utilisons. Nous avons décidé de le réduire, selon le temps nécessaire pour obtenir et exploiter les résultats. Nous constatons, qu'avec le filtre, nous sélectionnons 78.78% des mots hors vocabulaire (52 OOV récupérés sur 66 possibles). Nous remarquons également que la taille des ensembles de mots candidats est réduite à 14.75% de leur taille originale. Cependant, cet ensemble reste tout de même important. En effet, nous avons, en règle générale, une liste de plus d'un millier de mots pour seulement un faible nombre d'OOV pertinents. Par la suite, nous allons donc continuer de réduire le volume de mots candidats sélectionnés tout en gardant le plus d'OOV pertinents.

3.4 Filtre grammatical

Le filtre grammatical vient se placer dans la chaîne des opérations après le filtre phonétique. En effet, nous récupérons les sorties du filtre phonétique, c'est-à-dire, les mots candidats ainsi que la/les séquence(s) de mots qu'ils peuvent remplacer – par la suite, nous référons à ces mots remplacés par *« séquence d'origine »* ou *« séquence initiale »*. Toutefois, la taille de l'ensemble obtenu est trop importante pour l'intégrer au SRAP. Nous souhaitons alors la réduire en regardant si le remplacement des séquences d'origine par leur substitut, dans une séquence de mots, forme toujours un enchaînement grammaticalement correct. Dans un premier temps, nous expliquons le principe de fonctionnement du filtre grammatical, puis nous décrivons la méthode que nous avons mise en place. Enfin, nous présentons les résultats que nous avons obtenus.

3.4.1 Principe

Le filtrage grammatical consiste à vérifier la validité d'une séquence d'étiquettes catégorielles. En d'autre termes, nous examinons que l'enchaînement des catégories grammaticales des mots (nom, adjectif, verbe, *etc.*) qui constituent cette suite de mots est

segments	**OOV pertinents récupérés**	**OOV pertinents attendus**	**T1 = Total OOV récupérés**	**T2 = Total OOV à l'origine**	**Pourcentage de T1 par rapport T2**
1	3	3	1363	6389	21.33%
2	1	1	2966	17380	17.07%
3	1	1	785	6667	11.77%
4	1	1	915	10642	8.60%
5	1	2	1217	12478	9.75%
6	4	6	1865	10946	17.04%
7	4	6	1911	16039	11.91%
8	1	1	5420	33491	16.18%
9	2	2	2981	17661	16.88%
10	2	2	5536	31238	17.72%
11	1	1	3175	23509	13.51%
12	1	1	3046	21161	14.39%
13	3	8	1536	8072	19.03%
14	1	2	3378	20527	16.46%
15	6	7	636	4130	15.40%
16	1	1	303	3210	9.44%
17	1	1	742	4635	16.01%
18	2	2	1076	4467	24.09%
19	1	1	674	11108	6.07%
20	3	3	776	6309	12.30%
21	1	2	687	5196	13.22%
22	0	1	476	7011	6.79%
23	1	1	1025	5321	19.26%
24	6	6	1821	6076	29.97%
25	2	2	281	3452	8.14%
26	1	1	3577	16937	21.12%
27	1	1	1118	23041	4.85%
Total	52	66			
Pourcentage moyen d'OOV récupérés					14.75%

TABLE 1 – Tableau récapitulatif du nombre de mots hors vocabulaire sélectionnés avec le filtre phonétique

cohérent. Par exemple une séquence du type *« Le jeune homme rédige son rapport »* (article adjectif nom verbe pronom nom) est correcte. En revanche, une du type *« Le homme rapport rédige »* (article nom nom verbe) n'est pas correcte.

Le principe de fonctionnement de notre filtre grammatical est le suivant. Nous générons une nouvelle séquence de mots (une succession de quatre mots sans ponctuation) avec l'introduction du mot hors vocabulaire substitut à la place de sa séquence initiale. Puis, grâce à un outil d'étiquetage grammatical (*POS tagging : part-of-speech-tagging* en anglais), nous associons aux mots leurs catégories grammaticales à l'aide de leurs contextes (c'est-à-dire, grâce aux autres mots présents dans la séquence de mots). Ensuite, à l'aide d'un modèle n-grammes, regroupant les probabilités d'enchaînements des séquences d'étiquettes, nous déterminons un score sur la probabilité d'avoir la séquence. Pour ce faire, nous utilisons deux logiciels d'étiquetage morphosyntaxique : disambig et TreeTagger. Disambig a été conçu par l'équipe TEXMEX. Un de ses atouts est qu'il permet d'annoter les documents avec 144 étiquettes (ou *tags*) différentes (voir la liste des *tags* en annexe B), en revanche, il est bâti avec un dictionnaire fermé. En réalité, il s'agit du même dictionnaire que celui utilisé par IRENE. Il peut ainsi étiqueter tous les mots de la transcription, excepté les mots hors vocabulaire que nous avons introduits. C'est la raison pour laquelle nous utilisons aussi TreeTagger. En effet, à l'inverse de disambig, il n'est pas construit autour d'un dictionnaire. Il est défini par un ensemble de règles [Schmid 1994, Schmid 1995] et permet donc d'attribuer une étiquette aux mots hors vocabulaire. Toutefois, il dispose d'une collection d'étiquettes moins importante (33 *tags*, voir annexe B). Nous avons donc réalisé une association entre les étiquettes de TreeTagger et disambig pour n'obtenir que des *tags* reconnus par ce dernier. Nous avons choisi de ne pas utiliser exclusivement TreeTagger pour deux raisons. La première est due au nombre réduit de *tags* qu'il possède, ce qui procure à TreeTagger une connaissance moindre sur les classes grammaticales des mots par rapport à disambig. La seconde est que contrairement à disambig, TreeTagger ne donne pas les probabilités d'enchaînements des différentes étiquettes. Une fois que la transcription est étiquetée (par exemple figure 9), nous calculons un score pour déterminer la validité du *tag* du mot candidat inséré par rapport aux étiquettes qui l'entourent. Pour estimer ce score, nous adaptons la méthode utilisée dans [Brants 2000a], méthode que nous présentons dans la sous-section suivante.

3.4.2 Méthode d'estimation de probabilités avec un lissage par backoff

Nous voulons déterminer la probabilité que la nouvelle étiquette s'intègre correctement dans une séquence d'étiquettes. Nous souhaitons prendre en compte les classes grammaticales des mots placés avant et après le mot candidat inséré, car selon [Brants 2000b], on obtient une meilleure précision sur le score indiquant si la phrase est cohérente avec le nouveau *tag*. Il faut savoir que la plupart des techniques d'estimation de ce type n'utilisent que les prédécesseurs du terme concerné. Ainsi, nous travaillons avec des séquences de quatre *tags* ayant la forme suivante :

$$\textit{séquence de tags} = \; t_{i-2}\; t_{i-1}\; t_i\; t_{i+1}$$

où t_i est l'étiquette du mot candidat que nous ajoutons, t_{i-1} et t_{i-2} représentent respectivement les classes grammaticales du prédécesseur de notre OOV et du prédécesseur de son prédécesseur. Le *tag* t_{i+1} représente quant à lui la classe grammaticale du mot placé

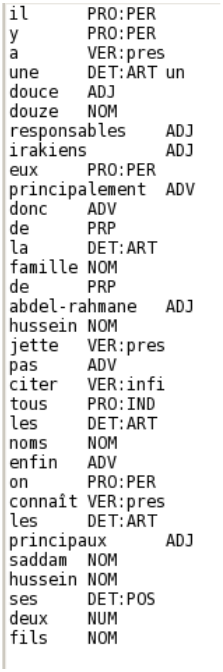


FIGURE 9 – Exemple de fichier étiqueté avec TreeTagger.

après le mot candidat.

Nous souhaitons pouvoir attribuer un score à notre étiquette t_i. [Brants 2000b] utilise une technique de lissage par interpolation qui consiste à combiner les n-grammes d'ordre inférieur avec un poids pour estimer la probabilité d'une séquence qu'elle soit connue ou non. Ce type de lissage permet d'obtenir de bons résultats, mais dans notre cas, il implique d'estimer les coefficients d'interpolation sur le corpus d'apprentissage de disambig. Or, pour des raisons de temps et comme le modèle 7-grammes de disambig intègre déjà un lissage par backoff, qui consiste à combiner les n-grammes d'ordre inférieur avec un poids pour estimer la probabilité d'une séquence inconnue, nous avons adapté le critère de [Brants 2000b] au cas du backoff.

Ainsi, nous souhaitons estimer la probabilité d'ajouter t_i dans une séquence. Cela se traduit par l'équation suivante :

$$score = \mathrm{P}(t_i \backslash t_{i-2},\, t_{i-1})\ \mathrm{P}(w_i \backslash t_i)\ \mathrm{P}(t_{i+1} \backslash t_i)\ ,$$

où $\mathrm{P}(t_i \backslash t_{i-2},\, t_{i-1})$ est la probabilité d'avoir t_i sachant les deux *tags* qui le précèdent, $\mathrm{P}(w_i \backslash t_i)$ est la probabilité d'avoir le mot candidat sachant l'étiquette, avec w_i représentant ce mot. Dans notre cas, cette probabilité sera égale à $\frac{1}{\text{Nombre de } tags}$, c'est-à-dire $\frac{1}{144}$. Et $\mathrm{P}(t_{i+1} \backslash t_i)$ est la probabilité d'avoir l'étiquette t_{i+1} sachant que l'on vient d'insérer l'étiquette t_i. C'est le calcul idéal du score dans le cas où, dans notre modèle, les probabilités $\mathrm{P}(t_i \backslash t_{i-2}\ ,\, t_{i-1})$ et $\mathrm{P}(t_{i+1} \backslash t_i)$ sont référencées.

En prenant en compte les séquences inconnues par l'introduction dans nos calculs du lissage par backoff, nous obtenons :

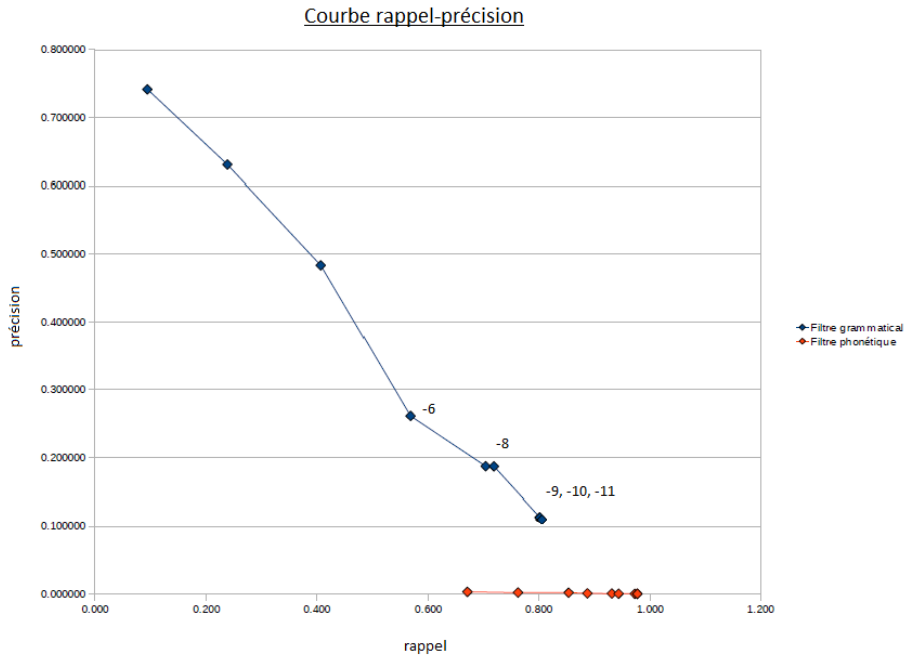


FIGURE 10 – Courbe rappel-précision des filtres grammatical et phonétique

$$\mathrm{P}(t_i \backslash t_{i-2}\ ,\ t_{i-1}) = \begin{cases} P(\mathrm{t}_i \backslash t_{i-2}\ ,\ t_{i-1}) & \text{si la séquence } t_{i-2}\ t_{i-1}\ t_i \text{ existe} \\ \alpha_1 P(\mathrm{t}_i \backslash t_{i-1}) & \text{si la séquence } t_{i-2}\ t_{i-1}\ t_i \text{ n'existe pas et la séquence } t_{i-1}\ t_i \text{ existe} \\ \alpha_2 P(\mathrm{t}_i) & \text{sinon} \end{cases}$$

avec α_1 et α_2 les coefficients de backoff respectivement associés aux probabilités $\mathrm{P}(t_i \backslash t_{i-1})$ et $\mathrm{P}(t_i)$.

On a également :

$$\mathrm{P}(t_{i+1} \backslash t_i) = \begin{cases} P(\mathrm{t}_{i+1} \backslash t_i) & \text{si la séquence } t_i\ t_{i+1} \text{ existe} \\ \alpha_3 P(\mathrm{t}_{i+1}) & \text{sinon} \end{cases}$$

avec α_3 le coefficient de backoff associé à la probabilité $\mathrm{P}(t_{i+1})$.

Nous obtenons ainsi pour chaque étiquette un score représentant la probabilité que t_i soit cohérent dans la séquence de mots. Nous cherchons alors à déterminer pour quel score nous avons le meilleur compromis entre le nombre de mots hors vocabulaire pertinents récupérés et le nombre total d'OOV sélectionnés. Nous allons donc mesurer pour différents scores les ensembles que nous obtenons afin de déterminer le meilleur seuil.

3.4.3 Seuil et résultats

Comme pour le filtre phonétique, nous cherchons maintenant à déterminer le seuil du filtre grammatical pour lequel nous obtiendrons le meilleur rapport entre les mesures de rappel et de précision. Les valeurs utilisées en entrée du filtre grammatical sont les valeurs de sorties du filtre phonétique pour un niveau de seuil égal à 0.35. Nous avons ainsi déterminé la courbe de rappel-précision en fonction de différents seuils (figure 10). Lorsque nous analysons la courbe du filtre grammatical et les valeurs associées, on s'aperçoit que le seuil optimal pourrait être à -9 (les niveaux de seuils allant de -3 à -11). En effet, les

seuils inférieurs à -9, comme -6 et -8, permettent respectivement de récupérer 55.7% et 75% des OOV pertinents provenant de la sortie du filtre phonétique alors qu'avec un seuil de -9 ou supérieur, on en sélectionne 80.7%. Cela représente 43.9%, pour un seuil à -6, 59% pour un seuil à -8 et 63.6% pour celui à -9, des OOV pertinents en entrée du filtre phonétique. Un autre constat est que la masse de mots candidats récupérés a diminuée. Ainsi, pour un seuil de -6, nous obtenons 8% du volume total de mots candidats en entrée du filtre phonétique. Pour les seuils de -8 et -9 nous avons respectivement 10% et 11.2% du volume d'origine. Pour des seuils supérieurs à -9, le nombre d'OOV pertinents sélectionnés ne change pas, seule la taille des ensembles de mots candidats augmente.

Le double filtrage nous a permis de réduire considérablement notre ensemble de mots candidats d'origine tout en gardant un grand nombre de mots pertinents. Le problème maintenant est que les listes de mots en entrée du filtre phonétique ne contiennent que 50% des OOV pertinents que l'on souhaite sélectionner. Nous présentons ainsi, dans la partie suivante, les travaux en cours de réalisation et notamment ceux sur l'augmentation du volume de mots pertinents présents dans les ensembles d'entrées de nos filtres.

4 Travaux en cours

Le stage se termine 3 semaines après le rendu de ce rapport, c'est pour cette raison que dans cette partie, nous abordons les travaux qui sont en cours de développement au moment de la rédaction de ce document. L'analyse des corpus d'adaptation que nous utilisons nous a permis de constater que le volume de mots hors vocabulaire pertinents récupéré n'était pas très élevé. Ainsi, nous travaillons actuellement sur une méthode permettant d'accroître le nombre d'OOV pertinents que nous pouvons retrouver par un système de génération de flexion des mots. C'est ce que nous présentons dans la première section de cette partie. Ensuite, nous décrivons une piste de réflexion sur la détection des zones d'erreur et leurs utilisations dans le système de filtre mis en place précédemment.

4.1 Mots candidats générés par flexion

Nous expliquons dans la partie 2.2 de ce rapport qu'à partir des corpus d'adaptation thématiques générés par G. Lecorvé, il nous est possible de récupérer au mieux 50% de mots hors vocabulaire pertinents. Nous nous sommes alors intéressés à l'autre moitié. Ainsi, nous avons constaté qu'une part importante de ces mots est en fait des verbes, des noms ou encore des adjectifs pour lesquels on peut retrouver dans le dictionnaire du sytème de reconnaissance de la parole un mot partageant le même lemme.

4.1.1 Principe

Nous cherchons ici à augmenter notre ensemble de mots candidats pour que la quantité de mots hors vocabulaire possible de récupérer dépasse les 50%. Nous nous sommes aperçus, en regardant certaines parties erronées d'une transcription, que les mots mal transcrits sont parfois des verbes conjugués à la mauvaise personne, ou encore des noms ou des adjectifs mal accordés. En effet, le dictionnaire du système ne connaît pas toutes les flexions des mots qu'il possède et ainsi, lors des transcriptions, il peut choisir des mots mal accordés ou mal conjugés. Par exemple on retrouve le verbe *« statuer »* conjugué

à la troisième personne du singulier de l'imparfait (*« statuait »*), mot présent de le dictionnaire, alors que c'est la troisième personne du pluriel de l'imparfait qui est attendu (*« statuaient »*), mot inconnu du voabulaire.

Le principe de la méthode que nous souhaitons mettre en œuvre est le suivant. Dans un premier temps, nous récupérons dans une transcription tous les verbes, noms et adjectifs. Ceci peut être fait grâce aux outils d'étiquetage morphosyntaxique que nous avons présentés dans la partie 3.4.1. Ensuite, pour chacun de ces mots, nous générons toutes les flexions possibles. La liste de toutes ces formes fléchies représente alors notre groupe de mots candidats. Cela peut avoir pour effet de générer des milliers de mots candidats pour une transcription. Nous pouvons alors appliquer nos filtres phonétique et grammatical pour réduire la taille de cet ensemble.

Une fois les mots candidats sélectionnés, il faut les ajouter au système de reconnaissance de la parole. Contrairement aux mots candidats provenant de corpus d'adaptation, l'intégration de ces nouveaux mots ne peut se faire facilement. En effet, nous avons besoin d'un corpus où l'on puisse trouver le mot que l'on souhaite ajouter, car pour rappel, pour générer les nouveaux modèles de langue, nous avons besoin d'observer les OOV dans un document textuel. Ainsi, une solution pourrait être d'effectuer des requêtes sur des moteurs de recherche pour récupérer des corpus d'adaptation contenant nos OOV. Ces requêtes pourraient être créées à partir du mot candidat ainsi que d'autres mots présents dans la transcription.

4.1.2 Méthode

La génération des flexions des mots est assez complexe à mettre en place. Pour les verbes, la solution la plus simple est de parcourir des sites Internet fournissant pour un verbe donné l'ensemble de ses conjugaisons, cependant pour les noms et les adjectifs, la tâche s'avère plus délicate. En effet, à cause du nombre de règles mais surtout en raison du nombre d'exceptions présentes dans la langue française, il est très difficile de spécifier des règles de flexion des noms et adjectifs. Cependant, on peut trouver dans la littérature des auteurs ayant traité ce problème avec des systèmes dédiés à la génération de flexion, comme PILAF [Courtin *et al.* 1994] ou d'autres encore, ayant créé des dictionnaires électroniques qui permettent d'obtenir toutes les flexions d'un mot [Dendien & Pierrel 2003].

Les premiers tests effectués, avec ces dictionnaires en ligne, montrent qu'il est possible de récupérer environ 15% de mots hors vocabulaire pertinents parmi l'ensemble des mots générés. Ces premiers résultats sont encourageants.

4.2 Mesures de confiance et filtrage

Nous appliquons les filtres phonétique et grammatical, présentés précédemment, sur l'ensemble d'une transcription. Or, une transcription n'est pas entièrement mal transcrite, en règle générale, ce sont seulement des petites zones qui contiennent des erreurs. Dans cette partie, nous nous proposons d'exposer une idée que nous aimerions développer, et qui pourrait améliorer l'efficacité de nos filtres.

4.2.1 Principe

Comme nous l'avons expliqué dans la partie 1.3.1, les mesures de confiance permettent au système de reconnaissance de la parole d'indiquer s'il pense avoir bien transcrit un mot ou non. L'idée serait alors de ne sélectionner dans les transcriptions que les parties ayant une mesure de confiance faible. Ensuite, au lieu d'appliquer nos filtres sur la transcription complète, nous pourrions les appliquer sur les zones reconnues peu fiables par le système. On réduirait ainsi l'ensemble de recherche du filtre phonétique et nous nous concentrerions que sur les zones d'erreur potentielles. Une seconde façon d'utiliser les mesures de confiance serait, dans notre cas, de les prendre en compte directement dans le calcul du score du filtre phonétique.

Nous avons écarté l'idée dès le départ d'utiliser les mesures de confiance données par le système IRENE car elles ne sont pas assez fiables. Toutefois, un membre de l'équipe, J. Fayolle [Fayolle *et al.* 2010], travaille sur ce problème afin d'améliorer la qualité de ces mesures. Nous pourrons alors réutiliser son travail pour compléter le nôtre.

Conclusion et perspectives

L'objectif de ce stage était de mettre en place un système pour adapter le vocabulaire d'une transcription au thème principal de cette dernière. Pour ce faire, nous avons développé une méthode originale, utilisant une combinaison de plusieurs critères pour la sélection des OOV, basée sur un double filtrage s'appuyant sur les critères phonétique et syntaxique que nous avons pu étudier dans l'état de l'art. Ainsi, dans un premier temps, nous avons sélectionné tous les mots thématiques hors vocabulaire présents dans les corpus d'adaptation extraits à partir du critère thématique. La liste de mots ainsi obtenue étant trop importante (des dizaines de milliers de mots), nous avons réalisé deux filtres, un phonétique et un grammatical, permettant de réduire le volume de mots hors vocabulaire tout en gardant le maximum d'OOV pertinents. Nous arrivons alors à retrouver 63.6% de mots hors vocabulaire pertinents tout en réduisant la taille de l'ensemble de mots hors vocabulaire à 11.2% de son volume d'origine. Il faut cepeandant noter que nous travaillons sur un sous-ensemble limité des mots possibles de récupérer puisque les corpus d'adaptation que nous utilisons ne contiennent que 50% des OOV pertinents que nous recherchons.

Nous pensons qu'il est possible d'améliorer ces résultats. En effet, comme nous l'avons expliqué dans la partie 4, nous pouvons accroître le nombre de mots qu'il est possible de récupérer en générant les flexions des mots présents dans la transcription. De plus, on pourrait affiner les résultats de nos filtres en ne cherchant les mots hors vocabulaire que sur les parties potentiellement erronées des transcriptions. En effet, la part de mots inutiles récupérés reste élevée par rapport au nombre de mots hors vocabulaire pertinents sélectionnés. Il serait alors intéressant d'arriver à discerner dans la transcription les zones d'erreur pour réduire le champs de recherche des mots candidats seulement aux endroits qui ont besoin d'être adaptés par l'utilisation de mesures de confiance.

Bien qu'il soit possible d'augmenter le volume de mots hors vocabulaire pertinents par la génération de flexion, nous restons encore loin d'avoir la possibilité de tous les retrouver. Il existe une autre catégorie, certes plus rare, de mots qui, lorsqu'ils ne sont pas présents dans le dictionnaire du système, pourraient être obtenus. Il s'agit de mots

que nous sommes capables de générer par dérivation, c'est-à-dire, par ajout de préfixes et de suffixes dérivationnels au mot (par exemple ajouter *« pré »* à *« qualification »* pour créer le mot *« préqualification »*).

Nous souhaitons également avoir une meilleure estimation de nos résultats de sortie. En effet, nous avons analysé les sorties des filtres que nous avons mis en place. Les résultats obtenus nous montrent seulement la proportion d'OOV que nous pouvons sélectionner et la taille à laquelle ces ensembles des mots candidats peuvent être réduits. Pour avoir un retour concret sur l'efficacité de notre filtrage, il faudrait intégrer nos listes de mots candidats dans le système et regarder l'impact que cela aurait sur le WER et le LER.

Annexe A

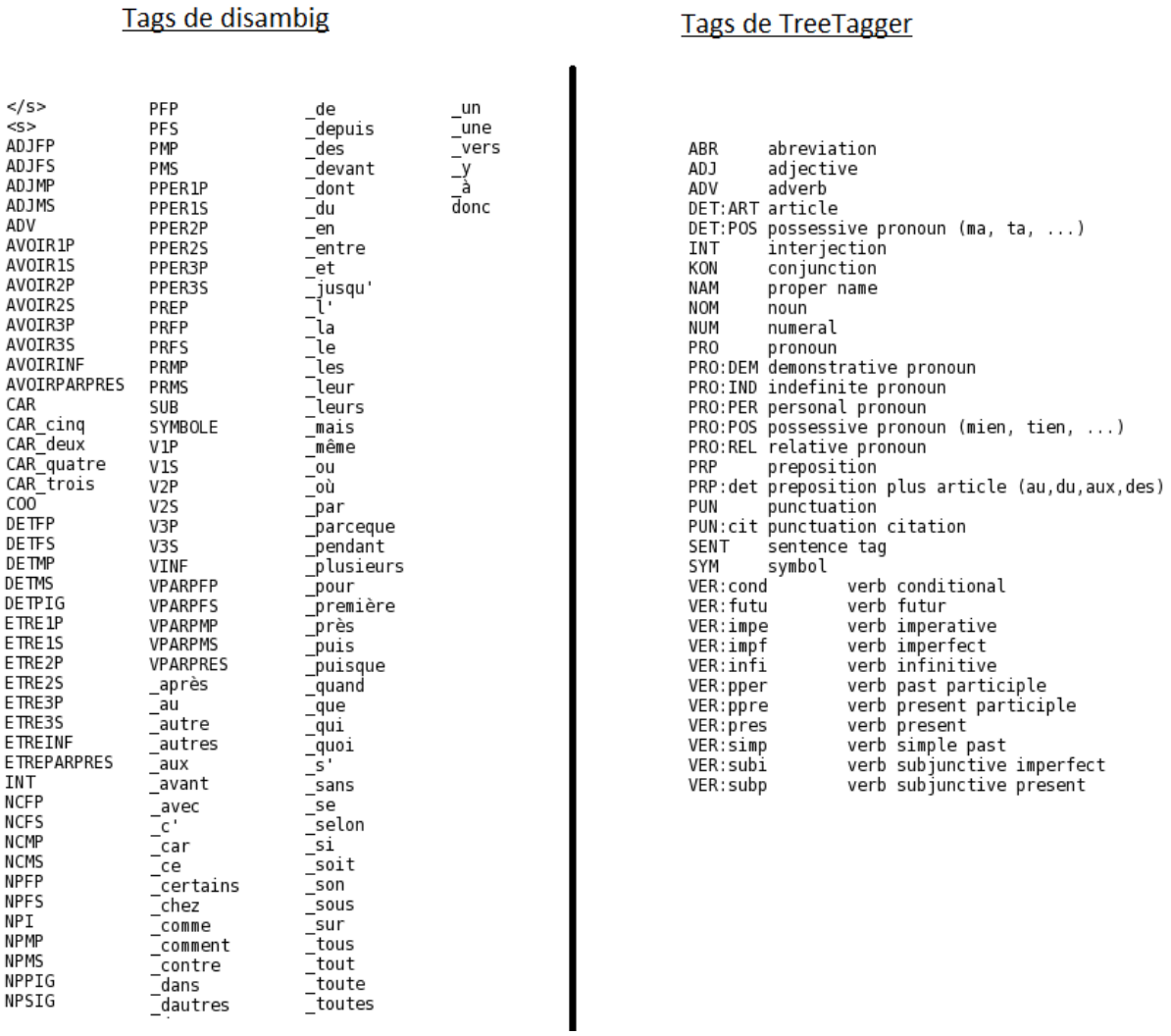
Nous avons réalisé des évaluations afin de déterminer les résultats que nous pourrions obtenir dans le meilleur des cas, c'est-à-dire, en ne sélectionnant que les mots absents d'une transcription. Le tableau suivant contient les mesures de référence de WER et de LER sur les transcriptions dans lesquelles nous avons intégré ces mots. Les colonnes « référence » donnent les scores obtenus avec les transcriptions initiales (sans ajout de mots hors vocabulaire). Les colonnes « avec OOV » donnent les mesures après ajouts des OOV pertinents.

	WER		LER	
Segments	référence	avec OOV	référence	avec OOV
1	12.4	12.4	11.5	10.8
2	10.5	10.3	9.0	9.0
3	16.0	16.0	17.0	17.0
4	8.0	8.0	6.3	6.3
5	14.3	13.8	15.1	15.1
6	21.9	20.3	18.2	15.9
7	22.4	20.0	20.0	15.7
8	14.0	14.0	10.5	10.5
9	27.0	27.0	24.9	24.9
10	21.8	20.0	17.3	15.7
11	35.8	35.8	29.6	29.6
12	11.5	10.6	32.4	32.7
13	37.9	37.9	20.3	18.7
14	21.9	21.1	28.0	26.7
15	31.6	31.1	8.9	8.2
16	27.9	27.5	29.1	28.6
17	19.7	19.7	16.4	16.4
18	8.2	8.2	11.5	11.5
19	17.8	16.8	18.5	16.9
20	18.0	18.3	18.4	18.4
21	33.3	33.3	34.2	34.2
22	46.5	46.5	47.8	47.8
23	20.1	20.1	18.9	18.6
24	11.2	11.2	9.6	11.6
25	8.6	8.6	6.2	6.2
26	20.5	21.9	18.8	18.8
27	22.7	22.6	22.9	22.9
28	18.8	18.8	22.2	22.2
29	9.6	9.6	10.3	10.3
30	23.8	23.8	20.5	20.5
31	37.4	37.1	35.5	36.0
32	11.9	11.2	8.6	7.8
33	22.2	22.2	23.7	23.7
34	7.5	7.5	8.7	8.7
35	11.7	11.3	11.0	10.6
36	25.5	25.5	30.2	30.2
37	34.8	34.8	36.6	36.6
38	30.1	29.6	27.5	26.2
39	29.2	28.8	22.1	21.6
40	28.6	28.6	25.5	25.5
41	18.0	18.0	15.8	15.8
42	16.7	16.5	15.9	15.4

TABLE 2 – Tableau récapitulatif des valeurs de WER et LER pour les transcriptions de référence et celles pour lesquelles nous avons rajouté des mots hors vocabulaire

Annexe B

Ici, on présente l'ensemble des étiquettes utilisées par disambig (à gauche) et TreeTagger (à droite).



Tags de disambig

```
</s>          PFP           _de          _un
<s>           PFS           _depuis      _une
ADJFP         PMP           _des         _vers
ADJFS         PMS           _devant      _y
ADJMP         PPER1P        _dont        _à
ADJMS         PPER1S        _du          donc
ADV           PPER2P        _en
AVOIR1P       PPER2S        _entre
AVOIR1S       PPER3P        _et
AVOIR2P       PPER3S        _jusqu'
AVOIR2S       PREP          _l'
AVOIR3P       PRFP          _la
AVOIR3S       PRFS          _le
AVOIRINF      PRMP          _les
AVOIRPARPRES  PRMS          _leur
CAR           SUB           _leurs
CAR_cinq      SYMBOLE       _mais
CAR_deux      V1P           _même
CAR_quatre    V1S           _ou
CAR_trois     V2P           _où
COO           V2S           _par
DETFP         V3P           _parceque
DETFS         V3S           _pendant
DETMP         VINF          _plusieurs
DETMS         VPARPFP       _pour
DETPIG        VPARPFS       _première
ETRE1P        VPARPMP       _près
ETRE1S        VPARPMS       _puis
ETRE2P        VPARPRES      _puisque
ETRE2S        _après        _quand
ETRE3P        _au           _que
ETRE3S        _autre        _qui
ETREINF       _autres       _quoi
ETREPARPRES   _aux          _s'
INT           _avant        _sans
NCFP          _avec         _se
NCFS          _c'           _selon
NCMP          _car          _si
NCMS          _ce           _soit
NPFP          _certains     _son
NPFS          _chez         _sous
NPI           _comme        _sur
NPMP          _comment      _tous
NPMS          _contre       _tout
NPPIG         _dans         _toute
NPSIG         _dautres      _toutes
```

Tags de TreeTagger

```
ABR     abreviation
ADJ     adjective
ADV     adverb
DET:ART article
DET:POS possessive pronoun (ma, ta, ...)
INT     interjection
KON     conjunction
NAM     proper name
NOM     noun
NUM     numeral
PRO     pronoun
PRO:DEM demonstrative pronoun
PRO:IND indefinite pronoun
PRO:PER personal pronoun
PRO:POS possessive pronoun (mien, tien, ...)
PRO:REL relative pronoun
PRP     preposition
PRP:det preposition plus article (au,du,aux,des)
PUN     punctuation
PUN:cit punctuation citation
SENT    sentence tag
SYM     symbol
VER:cond        verb conditional
VER:futu        verb futur
VER:impe        verb imperative
VER:impf        verb imperfect
VER:infi        verb infinitive
VER:pper        verb past participle
VER:ppre        verb present participle
VER:pres        verb present
VER:simp        verb simple past
VER:subi        verb subjunctive imperfect
VER:subp        verb subjunctive present
```

FIGURE 11 – Liste des *tags* des étiqueteurs morphosyntaxique disambig et TreeTagger

Références

[Allauzen & Gauvain 2003] A. Allauzen et J.L Gauvain. *Adaptation Automatique du Modèle de Langage dun Système de Transcription de Journaux Parlés.* Traitement Automatiquedes langues, 2003.

[Allauzen & Gauvin 2005a] A. Allauzen et J.L. Gauvin. *Diachronic Vocabulary Adaptation for Broadcast New Transcription.* In Proc. of Intl Conf. on Speech Language Technology (Interspeech), pages 1305–1308, 2005.

[Allauzen & Gauvin 2005b] A. Allauzen et J.L. Gauvin. *Open Vocabulary ASR for Audiovisual Document Indexation.* In Proc. of Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP), pages 1013–1016, 2005.

[Auzanne *et al.* 2000] C. Auzanne, J-S. Garofolo, J-G. Fiscus et W. Fisher. *Automatic Language Model Adaptation for Spoken Document Retrieval.* SDR 2000, TREC 9,, 2000.

[Bazzi *et al.* 1999] I. Bazzi, R. Schwartz et J. Makhoul. *An Omnifont Open-Vocabulary OCR System for English and Arabic.* IEEE Trans. on Pattern Analysis and Machine Intelligence, pages 495–504, 1999.

[Bazzi 2002] I. Bazzi. *Modelling Out-Of-Vocabulary Words for Robust Speech Recognition.* Ph.D. thesis, Massachusetts Institute of Technology, 2002.

[Béchet 2001] F. Béchet. *LIA_PHON un Système Complet de Phonétisation de Texte.* Dans Traitement Automatique Des Langues, volume 42, pages 47–68, 2001.

[Bechet & Yvon 2000] F. Bechet et F Yvon. *Les Noms Propres en Traitement Automatique de la Parole.* Traitement automatique des langue, 2000.

[Bellegarda 2004] J.R. Bellegarda. *Statistical Language Model Adaptation : Review and Perspectives.* Speech Communication, pages 93–108, 2004.

[Brants 2000a] T. Brants. *TnT - a Statistical Part-of-Speech Tagger.* In Proceedings of the Sixth Conference on Applied Natural Language Processing (ANLP-2000), 2000.

[Brants 2000b] T. Brants. *TnT - A Statistical Part-of-Speech Tagger.* In Proceedings of the 6th Applied NLP Conference, ANLP-2000, 2000.

[Bulyko *et al.* 2007] I. Bulyko, M. Ostendorf, M. Siu, T. Ng, A. Stolcke et O. Cetin. *Web Resources for Language Modeling in Conversational Speech Recognition.* ACM Trans. Speech Lang. Process., pages 1–25, 2007.

[Burget *et al.* 2008] L. Burget, P. Schwarz, P. Matejka, M. Hannemann, A. Rastrow, C. White, S. Khudanpur, H. Hermansky et J. Cernocky. *Combination of Strongly and Weakly Constrained Recognizers for Reliable Detection of OOVS.* In ICASSP, 2008.

[Courtin *et al.* 1994] J. Courtin, D.Dujardin, D. Genthial et I. Kowarski. *Analyse et Génération Morphologique avec le Système PILAF.* T.A.L. 35/2, pages 93–109, 1994.

[de Mareüil *et al.* 2000] P. Boula de Mareüil, F. Yvon, C. dAlessandro, V. Aubergé, J. Vaissière et A. Amelot. *A French Phonetic Lexicon with Variants for Speech and Language Processing.* In 2nd Language Resources Engineering Conference, 2000.

[Dendien & Pierrel 2003] J. Dendien et J-M. Pierrel. *Le Trésor de la Langue Française Informatisé. Un Exemple d'Informatisation d'un Dictionnaire de Langue de référence = The electronic version of the Trésor de la Langue Française (TLF).* TAL. Traitement automatique des langues, 2003.

[Fayolle *et al.* 2010] J. Fayolle, F. Moreau, C. Raymond, G.Gravier et P. Gros. *CRF-based Combination of Contextual Features to Improve A Posteriori Word-level Confidence Measures.* In International Conference on Speech Communication and Technologies, Interspeech'10, pages 1942–1945, 2010.

[Federico & N.Bertoldi 2001] M. Federico et N.Bertoldi. *Broadcast News Adaptation using Contemporary Texts.* Proc. Eurospeech, Aalborg, Denmark, 2001.

[Geutner *et al.* 1998a] P. Geutner, M. Finke et P. Scheytt. *Adaptive Vocabularies for Transcribing Multilingual Broadcast News.* In Proc. of the IEEE Intl Conf. On Acoustics, Speech and Signal Processing (ICASSP), pages 925–928, 1998.

[Geutner *et al.* 1998b] P. Geutner, M. Finke et A. Waibel. *Phonetic-Distance-Based Hypothesis Driven Lexical Adaptation for Transcribing Multilingual Broadcast News.* In Proc. of 5th Conf. On Spoken Language Processing (ICSLP), pages 1297–1313, 1998.

[Gravier *et al.* 2005] G. Gravier, F. Yvon et M. Ben. *IRENE, le Système Commun IRISA - ENST dIndexation d'Émissions Radiophoniques.* In Atelier ESTER Phase II, 2005.

[Guinaudeau 2008] C. Guinaudeau. *Contrôle Automatisé de Contenu Télévisuel.* Rapport de Master 2, IRISA/Université Caen Basse Normandie, 2008.

[Huet 2007] S. Huet. *Informations morpho-syntaxiques et adaptation thématique pour améliorer la reconnaissance de la parole.* Thèse de doctorat, université de Rennes 1, 2007.

[Jelinek 1976] F. Jelinek. *Continuous Speech Recognition by Statistical Methods.* In Proc. of the IEEE, pages 532–556, 1976.

[Jelinek 1998] F. Jelinek. Statistical Methods for Speech Recognition. The MIT Press, 1998.

[Jiang 2004] H. Jiang. *Confidence Measures for Speech Recognition : A Survey.* In Speech Communication, 2004.

[Kemp & Waibel 1998] T. Kemp et A. Waibel. *Reducing the OOV Rate in Broadcast News Speech Recognition.* In Proc. of the 5th Intl Conf. on Spoken Language Processing (ICSLP), pages 1839–1842, 1998.

[Lecorvé 2007] G. Lecorvé. *Adaptation thématique d'un système de transcription automatique de la parole.* Mémoire de D.E.A, IRISA/INSA Rennes, 2007.

[Lecorvé 2010] G. Lecorvé. *Adaptation thématique non supervisée d'un système de reconnaissance automatique de la parole.* Thèse de doctorat, université de Rennes 1, 2010.

[Marin *et al.* 2009] M.A. Marin, S. Feldman, M. Ostendorf et M. Gupta. *Filtering Web Text to Match Target Genres.* Proc. of the IEEE Intl. Conf. on Acoustics, Speech and Signal Processing (ICASSP), pages 3705–3708, 2009.

[Martins *et al.* 2006] C. Martins, A. Texeira et J. Neto. *Dynamic Vocabulary Adaptation for a Daily and Real-Time Broadcast News Transcription System.* In Proc. of the Spoken Language Technology Workshop, pages 146-149, pages 146–149, 2006.

[Muscariello *et al.* 2009] A. Muscariello, G. Gravier et F. Bimbot. *Variability Tolerant Audio Motif Discovery.* In Intl. Multimedia Model Conference, 2009.

[Oger *et al.* 2008] S. Oger, G. Linarès, F. Béchet et P. Nocera. *On-Demand New Word Learning Using the World Wide Web.* In Proc. of the IEEE Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP), pages 4305–4308, 2008.

[Ohtsuki *et al.* 2005] K. Ohtsuki, N. Hiroshima, M. Oku et A. Imamura. *Unsupervided Vocabulary Expansion for Automatic Transcription of Broadcast News.* In Proc. of the IEEE Intl Conf. on Acoustics, Speech and Signal Processing (ICASSP), pages 1021–1024, 2005.

[Palmer & Ostendorf 2005] D.D. Palmer et M. Ostendorf. *Improving Out-Of-Vocabulary Name Resolution.* Computer Speech & Language, pages 107–128, 2005.

[Rabiner 1989] L.R. Rabiner. *A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition.* In Proc. of the IEEE, pages 257–286, 1989.

[Rastrow *et al.* 2009] A. Rastrow, A. Sethy et B. Ramabhadran. *A New Method for OOV Detection using Hybrid Word/Fragment System.* Proceedings of ICASSP, 2009.

[Rosenfeld 1995] R. Rosenfeld. *Optimizing Lexical and N-gram Coverage Via Judicious Use of Linguistic Data.* In Proc. European Conf. on Speech Technology, pages 1763–1766, 1995.

[Salton 1989] G. Salton. Automatic text processing : the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc., 1989.

[Schmid 1994] H. Schmid. *Probabilistic Part-of-Speech Tagging Using Decision Trees.* In Proceedings of the International Conference on New Methods in Language Processing, 1994.

[Schmid 1995] H. Schmid. *Improvements in Part-of-Speech Tagging With an Application To German.* In Proceedings of the ACL SIGDAT-Workshop, 1995.

[Schwarm *et al.* 2004] S. Schwarm, I. Bulyko et M. Ostendorf. *Adaptive Language Modeling with Varied Sources to Cover New Vocabulary Item.* IEEE Trans. on Speech and Audio Processing, pages 334–342, 2004.

[Seymore & Rosenfeld 1997] K. Seymore et R. Rosenfeld. *Using Story Topics for Language Model Adaptation.* In Proc. of the 5th European Conf. on Speech Communication and Technology (Eurospeech), pages 1987–1990, 1997.

[Vaufreydaz *et al.* 1999] D. Vaufreydaz, M. Akbar et J. Rouillard. *Internet Documents : A Rich Source for Spoken Language Modeling.* In Proc. of the IEEE Workshop Automatic Speech Recognition and Understanding (ASRU), pages 277–280, 1999.

[Wang 2009] S. Wang. *Using Graphone Models in Automatic Speech Recognition.* Masters thesis, Massachusetts Institute of Technology, 2009.

[Wechsler *et al.* 1998] M. Wechsler, E. Munteanu et P. Schäuble. *New Techniques for Openvocabulary Spoken Document Retrieval.* In 21st annual international ACM SIGIR Conference on Research and Development in Information Retrieval, 1998.

[Wessel *et al.* 2001] F. Wessel, R. Schlüter, K. Macherey et H. Ney. *Confidence measures for large vocabulary continuous speech recognition.* IEEE Transactions on Speech and Audio Processing, 2001.