

Université de la Manouba

ÉCOLE NATIONALE DES SCIENCES DE  
L'INFORMATIQUE



RAPPORT  
de Mémoire de Fin d'Etude

Présenté en vue de l'obtention du titre  
d'INGENIEUR EN INFORMATIQUE  
par

Omar BOURAS

Sujet :

**Extraction automatique d'informations syntaxiques et  
sémantiques en vue  
d'améliorer la reconnaissance automatique de la parole**

Organisme : **LORIA UMR 7503**

Nom du responsable : **Karl TOMBRE**, Directeur du LORIA

Encadré par : **Christophe CERISARA**, Chargé de recherche au CNRS

Supervisé par : **Faouzi Ghorbel**, Professeur à l'ENSI

Adresse : INRIA LORIA, rue du Jardin Botanique, BP 615, 54600, Villers-lès-Nancy, France

Tél : +33 3 83 59 30 00

Fax : +33 3 83 27 83 19



## Remerciements

Je remercie tous les membres de l'équipe parole du Loria pour leur chaleureux accueil et leur complicité. En particulier, je tiens à exprimer ma gratitude envers Christophe *Cerisara*, mon encadreur qui a toujours répondu présent à mes sollicitudes, m'a enrichi par sa vision critique des choses, son savoir faire et son savoir être.

Je ne peux passer cette occasion sans rendre hommage à mes enseignants ainsi qu'à tous ceux qui contribuent à la réussite de la formation et la joie de vivre à l'ENSI.

Merci aux membres du jury d'avoir accepté d'évaluer ce travail.

Enfin, je tiens à saluer tous ceux qui m'ont aidé du près ou de loin dans ce stage et que j'ai oublié de mentionner. A tous un très grand merci.

*Je dédié ce travail,  
A ma famille élargie.  
A mes parents qui ont sacrifié des années de leurs vies pour que je sois là  
aujourd'hui.  
A mes frères, soeurs, beaux frères et belles soeurs avec qui je partage les  
moments de joies et de peines.  
A mes petits indiens : Nouha, Ahmed, Aziz, Amine, Nour et Youssef qui font le  
bonheur de tous.  
A mes amis qui continuent à me supporter comme je suis.*

# Table des matières

|  |            |
|--|------------|
| <b>Table des figures</b>   | <b>vii</b> |
| <b>Introduction</b>  | <b>1</b>   |
| <b>Chapitre 1 Présentation du cadre de stage</b>   | <b>2</b>   |
| 1.1 LORIA, unité mixte de recherche . . . . .  | 2          |
| 1.2 Equipe Parole, organisation et travaux de recherches . . . . .   | 3          |
| 1.2.1 Domaines d'activités . . . . .   | 3          |
| 1.2.2 Travaux récents . . . . .  | 4          |
| 1.2.3 LE SYSTÈME ANTS . . . . .  | 4          |
| <b>Chapitre 2 État de l'art</b>  | <b>7</b>   |
| 2.1 Approche sémantique dans l'aide à la reconnaissance automatique de la parole . . . . .   | 7          |
| 2.2 Recherche de l'information sémantique . . . . .  | 8          |
| 2.2.1 LSA, Latent Semantic Analyse . . . . .   | 8          |
| 2.2.2 Random Indexing . . . . .  | 9          |
| 2.3 Utilisation de l'information sémantique comme mesure de confiance . . . . .  | 9          |
| 2.4 Travaux d' Inkpen sur la détection des erreurs de reconnaissance grâce à des liens sémantiques. . . . .                                  | 12         |
| <b>Chapitre 3 Exploitation de la technique Random Indexing en extraction d'une information sémantique à partir du corpus <i>Le Monde</i></b> | <b>16</b>  |
| 3.1 Selection d'une approche sémantique . . . . .  | 17         |
| 3.1.1 Présentation du <i>Monde</i> . . . . .   | 17         |
| 3.1.2 Choix d'une approche sémantique . . . . .  | 17         |
| 3.2 Préparation des données . . . . .  | 17         |
| 3.3 Présentation de la JAVASDM . . . . .   | 18         |

---

|   |   |           |
|---|---|-----------|
| 3.4   | Tests préliminaires sur l'influence des mots outils . . . . . | 18        |
| 3.5   | Construction d'une mesure de confiance sémantique . . . . .   | 20        |
| 3.5.1   | Exploitation des sorties de la reconnaissance . . . . .       | 20        |
| 3.5.2   | Évaluation du score sémantique . . . . .                      | 21        |
| <b>Chapitre 4 Intégration dans un système de reconnaissance</b> |   | <b>28</b> |
| 4.1   | Présentation de la mesure de Ney . . . . .                    | 28        |
| 4.2   | Apport du score sémantique . . . . .                          | 29        |
| 4.3   | Combinaison de la LLS et de la mesure de Ney . . . . .        | 31        |
| 4.3.1   | Combinaison en cascade . . . . .                              | 31        |
| 4.3.2   | combinaisons linéaires . . . . .                              | 32        |
| 4.4   | Exploitation des <i>N-best</i> . . . . .                      | 33        |
| 4.4.1   | Présentation de l'expérience . . . . .                        | 33        |
| 4.4.2   | Résultats . . . . .   | 35        |
| <b>Conclusion</b>   |   | <b>37</b> |

---



# Table des figures

|     |   |    |
|-----|---|----|
| 1.1 | Image prise à l'intérieur des locaux de LORIA et symbolise l'ouverture de cette entité à l'extérieur . . . . .  | 3  |
| 1.2 | Image de la tête parlante de l'équipe Parole . . . . .  | 4  |
| 1.3 | Architecture du système ANTS. (Extrait de[1]). . . . .  | 5  |
| 2.1 | Distributions des valeurs de <i>PSS</i> pour les bons et mauvais termes.(Extrait de [2]) . . . . .  | 11 |
| 2.2 | Recall/precision curves for PSS, NB, and NBPSS.(Extrait de [2]) . . . . .   | 12 |
| 2.3 | P-R curves of PMI vs. Roget (with All and AVG) on the BBN dataset. Each P-R point corresponds to a different value of the threshold K (high Recall for low values of K, high Precision for high values of K).(Extrait de [3]) . . . . . | 14 |
| 2.4 | Content Words Error Rate (cWER), %Lost good keywords (%Lost) and F-measure as a function of the filtering level K for the Window-PMI-3MAXconfiguration on the BBN dataset.(Extrait de [3]) . . . . .                                    | 15 |
| 3.1 | Diagramme de classe de la JavaSDM modifié . . . . .   | 23 |
| 3.2 | Répartition des cosinus moyens des termes du corpus. . . . .  | 24 |
| 3.3 | Distribution des cosinus des angles entre "et" et tous les autres termes du vocabulaire. . . . .  | 24 |
| 3.4 | Distribution des cosinus des angles entre "sadi" et tous les autres termes du vocabulaire. . . . .  | 25 |
| 3.5 | Histogrammes respectifs des termes mal et correctement reconnus par le système de reconnaissance, 2000 mots outils sont supprimés du vocabulaire. . . . .   | 26 |
| 3.6 | Courbes DET paramétrisées par le nombre de mots outils supprimés . . . . .  | 27 |
| 4.1 | PR de Ney et PR sémantique (500 mots outils supprimés du vocabulaire) à partir d'un fichier d'ESTER2. Nous retrouvons les rappels au niveau des abscisses et les précisions au niveau des ordonnées. . . . .                            | 30 |
| 4.2 | Améliorations potentielles apportés par la LLS à la mesure de Ney . . . . .   | 31 |
| 4.3 | Combinaisons en cascades de la mesure de Ney avec la LLS. . . . .   | 32 |
| 4.4 | Courbes PR des combinaisons linéaires des mesures de Ney et LLS paramétrées par le taux de contribution sémantique <i>alpha</i> . . . . .   | 33 |

---

|     |  |    |
|-----|--|----|
| 4.5 | Courbes PR des combinaisons linéaires des mesures de Ney et LLS paramétrées par le taux de la contribution sémantique <i>alpha</i> (Partie hauts rappels). . . . . | 34 |
|-----|--|----|

# Introduction

La reconnaissance vocale, ou “reconnaissance automatique” de la parole (Automatic Speech Recognition, ASR), est une technologie informatique qui permet d’analyser un mot ou une phrase, enregistrés au moyen d’un microphone, pour les transcrire sous une forme textuelle. Cette technologie exploite des modèles acoustiques et de modèles de langage. Les modèles acoustiques permettent de prendre en compte des contraintes acoustiques et phonétiques au niveau d’un son ou d’un groupe de sons. Alors que les modèles de langages définissent les contraintes syntaxiques et sémantiques au sein d’un groupe de mots ou d’une phrase. [4]

Les modèles acoustiques sont basés sur des approches stochastiques. Ces techniques se sont avérées les plus efficaces en reconnaissance de la parole, dans un laboratoire ou un environnement non bruité. Les modèles acoustiques développés sont principalement formés par des HMM (Modèle de Markov caché) ou des réseaux bayésiens (une distribution jointe d’un ensemble de variables aléatoires donné est associée à un graphique orienté acyclique).

Les modèles de langages sont presque toujours des modèles stochastiques ayant une portée locale ou à court terme. Afin d’améliorer leurs performances, ces modèles ont besoin de continuelles améliorations pour s’adapter à la complexité de la langue. Ainsi l’idée de modéliser quelques phénomènes sémantiques de la langue d’une manière statistique est venue afin de lever certaines ambiguïtés et améliorer en conséquent les taux de reconnaissance.

Ce dernier volet a été le thème de mon stage de PFE effectué à l’UMR 7305. Nous avons essayé d’extraire des informations sémantiques et de les traduire en des entrées exploitables par le système de reconnaissance de la parole ANTS de l’équipe Parole. Dans ce rapport, nous parlerons en première partie du cadre général du stage. Ensuite, dans la deuxième partie, nous présenterons certaines approches et études sur l’introduction de l’information sémantique dans la reconnaissance de la parole. En troisième partie, nous expliciterons notre approche pour l’exploitation de la technique de Random Indexing dans l’extraction d’information sémantique de notre corpus *Le Monde*. Enfin, nous présenterons les résultats et les performances de ce travail.

# Chapitre 1

## Présentation du cadre de stage

### Sommaire

---

|            |   |          |
|------------|---|----------|
| <b>1.1</b> | <b>LORIA, unité mixte de recherche . . . . .</b>                      | <b>2</b> |
| <b>1.2</b> | <b>Equipe Parole, organisation et travaux de recherches . . . . .</b> | <b>3</b> |
| 1.2.1      | Domaines d'activités . . . . .  | 3        |
| 1.2.2      | Travaux récents . . . . .   | 4        |
| 1.2.3      | LE SYSTÈME ANTS . . . . .   | 4        |

---

### Introduction

Dans cette partie, nous allons présenter en premier lieu l'organisation du LORIA (UMR 7503). Ensuite, nous allons nous concentrer sur l'équipe Parole en particulier et faire un petit descriptif de ses travaux en cours.

### 1.1 LORIA, unité mixte de recherche

Le LORIA, Laboratoire Lorrain de Recherche en Informatique et ses Applications, est une Unité Mixte de Recherche, (UMR 7503), officialisée depuis le 19 décembre 1997 et commune à plusieurs établissements : le CNRS (Centre National de la Recherche Scientifique), l'INPL (Institut National Polytechnique de Lorraine), l'INRIA (Institut National de Recherche en Informatique et en Automatique), l'UHP (Université Henri Poincaré, Nancy 1) et Nancy 2 (Université Nancy 2). Le LORIA compte plus de 450 personnes. Chercheurs, enseignants-chercheurs, doctorants, post-doctorants, ingénieurs, techniciens et personnels administratifs sont organisés en équipes de recherche et services de soutien à la recherche. Forte de son grand nombre d'équipes de recherches, comprenant plus de trente unités dont Parole, Cortex (réseaux de neurones), Trio (réseaux), Cassis (Sécurité informatique) et Alice (Image 3D), l'UMR 7503 s'est fixé les missions suivantes :

- La recherche fondamentale et appliquée au niveau international dans le domaine des Sciences et Technologies de l'Information et de la Communication.

- La formation par la recherche en partenariat avec les universités lorraines.
- Le transfert technologique par le biais de partenariats industriels et par l'aide à la création d'entreprises. Chaque année le LORIA invite aussi une trentaine de chercheurs étrangers et coopère avec plusieurs unités de recherche dans les cinq continents.



FIG. 1.1 – Image prise à l'intérieur des locaux de LORIA et symbolise l'ouverture de cette entité à l'extérieur

## 1.2 Equipe Parole, organisation et travaux de recherches

L'équipe Parole est un projet de collaboration entre l'INRIA, le CNRS, l'UHP et Nancy 2. Dirigée par monsieur Yves Laprie, HDR et directeur de recherche au CNRS, l'équipe comporte plus de trente personnes, attachés de recherche, universitaires, thésards, assistants techniques, assistante et ingénieur.

### 1.2.1 Domaines d'activités

Les activités de recherche au sein de Parole sont concentrées sur deux principaux thèmes, à savoir l'analyse de la parole et sa modélisation pour la reconnaissance automatique. L'analyse en premier lieu traite l'aspect " extraction et définition de repères acoustiques " et " l'inversion acoustico-articulatoire ". Ensuite, la modélisation se fait par le biais de constructions de modèles stochastiques et de l'adaptation de systèmes de reconnaissance aux nouveaux locuteurs et au bruit. Ces centres d'intérêts aboutissent à des applications qui peuvent s'étendre de la reconnaissance automatique de la parole jusqu'aux domaines paramédicaux. L'analyse de la parole, par exemple, contribue au développement de nouvelles technologies pour l'apprentissage des langues (pour les malentendants et pour l'enseignement des langues étrangères), ainsi que pour les appareils auditifs.[5]

### 1.2.2 Travaux récents

En analyse de la parole, l'équipe s'intéresse à l'inversion acoustico-articulatoire. La recherche est en cours pour l'exploitation de données 3D en vue de synthétiser une tête parlante (FIG. 1.2). Ainsi, la synthèse de la parole à partir de texte, et la correction de la prosodie pour les apprenants de la langue anglaise, sont des sujets d'actualité au sein de l'équipe.



FIG. 1.2 – Image de la tête parlante de l'équipe Parole

En ce qui concerne la reconnaissance automatique de la parole, la robustesse de la reconnaissance aux bruits et la variabilité au niveau de l'acquisition du signal motive la recherche pour l'amélioration des technologies de reconnaissances. Dans ce sens, les travaux sont en cours pour le développement d'une approche de débruitage bayésien, la reconnaissance avec des données manquantes et l'adaptation des modèles de langages aux locuteurs non natifs (variabilité temporelle au niveau de la prononciation) de la langue. Ainsi, un système de transcription de journaux radiophoniques Français ANTS a été développé. Toujours dans le thème de la reconnaissance de corpus, les recherches se poursuivent sur la segmentation du signal en parole, musique et annonces publicitaires.

### 1.2.3 LE SYSTÈME ANTS

L'équipe Parole du LORIA a développé un système de transcription automatique d'émissions radiophoniques : ANTS <sup>1</sup> dans le cadre du projet ESTER. Ce projet a pour objectif l'évaluation des systèmes automatiques de transcription de données radiophoniques francophones.

Le système ANTS se compose de plusieurs parties : des modules de segmentation, un moteur de reconnaissance, des modèles acoustiques, un lexique et un modèle de langage(FIG. 1.3).

#### Les modules de segmentation

Le but de l'étape de segmentation du signal audio est double : d'une part, découper le signal audio en segments homogènes de taille acceptable par le moteur de reconnaissance, d'autre part, permettre d'utiliser des modèles ou des algorithmes spécifiques suivant la

---

<sup>1</sup>Automatic News Transcription System

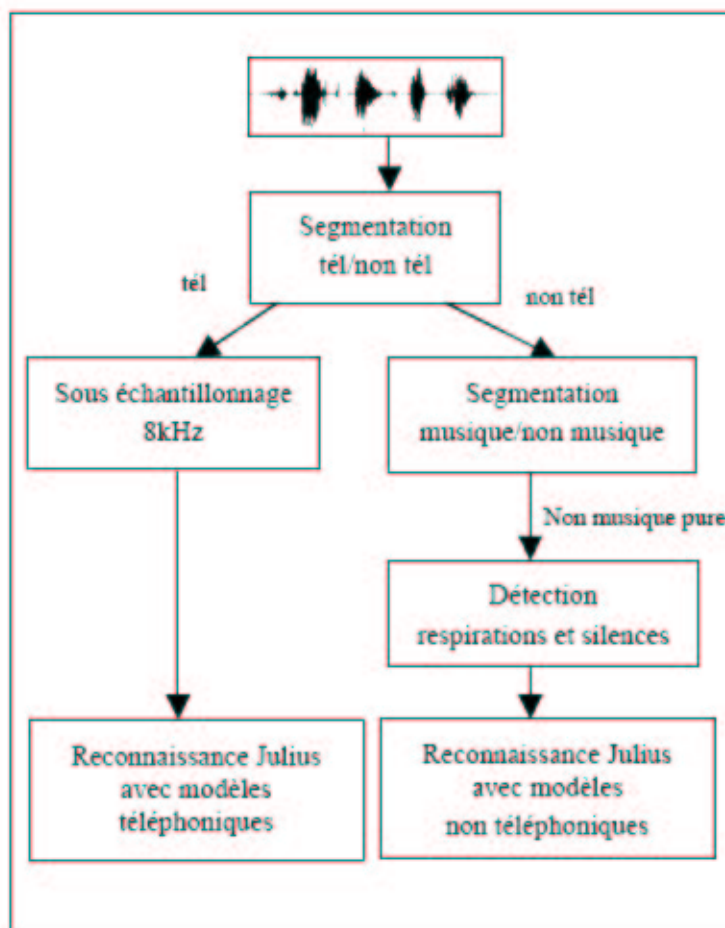


FIG. 1.3 – Architecture du système ANTS. (Extrait de[1]).

nature du segment.

la segmentation parole large bande / parole téléphonique a été séparée de la détection des parties musicales. Ceci a permis de mettre en oeuvre une méthode spécifique pour la segmentation téléphone/non-téléphone. L'étape de segmentation est donc composée de trois modules qui s'enchaînent dans l'ordre suivant :

- La segmentation parole large bande / parole téléphonique repose sur la différence d'énergie entre les basses [0 - 4kHz] et les hautes fréquences [4 kHz - 8kHz]. Cette différence doit être importante si le locuteur est au téléphone. Dans une première étape, le module affecte une valeur  $e(t)$  à chaque trame  $t$  de signal,  $e(t)$  vaut  $1$  si la différence d'énergie est supérieure à un seuil et  $-1$  sinon. Puis, la fonction  $q(t)$  est calculée à chaque trame  $t$  :

$$q(t) = \left| \sum_{t-1}^{t-L} e(t) - \sum_{t+1}^{t+L} e(t) \right|$$

La courbe  $q(t)$ , plus ou moins lissée en fonction du facteur  $L$ , doit présenter un maximum lors du passage téléphone/non téléphone. Enfin, un algorithme de recherche de pics permet de trouver les points de rupture et de segmenter le flux audio.

- La segmentation Parole/Musique est fondée sur la mise en compétition de cinq modèles constitués de mélanges de gaussiennes (GMM) : Parole Téléphonique (PT), Parole Non Téléphonique (PNT), Musique Instrumentale (MI), Chansons (C) et Parole- Musique (P&M), modélisant la superposition de la parole et de la musique. La segmentation est donc réalisée par une phase de reconnaissance dans laquelle une durée minimale de 0,5 seconde est imposée pour chaque segment.
- La détection des respirations et des silences a deux buts : premièrement, découper la parole en morceaux de plus petite taille ; deuxièmement trouver les groupes de souffle qui correspondent souvent à des entités syntaxiques ou sémantiques. Pour réaliser une telle segmentation, nous procédons à une reconnaissance au niveau phonétique en utilisant des modèles de phonèmes, un modèle de silence et un modèle de respiration ou de souffle. La grammaire utilisée pendant cette reconnaissance attribue la même probabilité à toutes les transitions entre les modèles. Toute portion de signal comprise entre deux respirations et/ou silences est alors extraite.

A la fin de cette étape, nous obtenons des segments de taille raisonnable utilisables par le moteur de reconnaissance.

### **Le moteur de reconnaissance, le lexique et le modèle de langage**

Le moteur de reconnaissance Julius développé par Akinobu Lee est utilisé. Ce logiciel effectue la reconnaissance en deux passes : la première passe est trame-synchrone, elle utilise un bigramme et fournit un treillis (graphe) de mots. La deuxième passe est fondée sur un algorithme à pile et utilise un trigramme. Le système Julius offre la possibilité de définir plusieurs prononciations pour chaque mot du lexique.[1]

## **Conclusion**

Tout au long de cette première partie, nous avons cherché à présenter le statut dans lequel est organisé le LORIA. Ensuite, nous nous sommes concentrés sur les domaines de recherches de l'équipe Parole dans l'analyse et la reconnaissance de la parole. Dans la partie suivante, nous nous focaliserons sur l'état de l'art de l'introduction de la sémantique dans la reconnaissance de la parole.

# Chapitre 2

## État de l'art

### Sommaire

---

|            |  |           |
|------------|--|-----------|
| <b>2.1</b> | <b>Approche sémantique dans l'aide à la reconnaissance automatique de la parole . . . . .</b>                  | <b>7</b>  |
| <b>2.2</b> | <b>Recherche de l'information sémantique . . . . .</b>   | <b>8</b>  |
| 2.2.1      | LSA, Latent Semantic Analyse . . . . .   | 8         |
| 2.2.2      | Random Indexing . . . . .  | 9         |
| <b>2.3</b> | <b>Utilisation de l'information sémantique comme mesure de confiance . . . . .</b>                             | <b>9</b>  |
| <b>2.4</b> | <b>Travaux d' Inkpen sur la détection des erreurs de reconnaissance grâce à des liens sémantiques. . . . .</b> | <b>12</b> |

---

### Introduction

Après avoir évoqué dans le chapitre précédent le thème de la reconnaissance automatique de la parole, nous mettrons l'accent sur le rapport qui peut exister entre ce domaine de recherche et les relations sémantiques découvertes dans un texte. Ensuite, nous exposerons les travaux de Cox et ceux d'Inkpen.

## 2.1 Approche sémantique dans l'aide à la reconnaissance automatique de la parole

Le concept d'une approche syntaxique ou sémantique dans le traitement de la parole est assimilé au fait qu'un Homme saurait reconnaître si un mot a un sens ou pas sachant le contexte dans lequel il est présent. Ainsi, cette information peut contribuer à la détection et la correction de certaines erreurs figurant dans le texte sorti du système automatique de la reconnaissance, et non décelé par celui-ci.

La reconnaissance automatique de la parole ( une définition encyclopédique) est une technologie qui permet d'analyser un mot ou une phrase enregistrée au moyen d'un microphone

pour la transcrire sous la forme d'un texte exploitable par une machine. La reconnaissance vocale, ainsi que la synthèse vocale, l'identification du locuteur ou la vérification du locuteur, font partie des technologies de traitement de la parole. Un système de reconnaissance est construit à partir d'un ensemble de modèles statistiques, acoustiques (et/ou) de langages [4]. Pour intégrer une information sémantique, il est envisageable de l'introduire dans une mesure de confiance agissant directement dans le processus ou apportant des corrections à la sortie (phase complémentaire après reconnaissance automatique faite par le système) de la reconnaissance du système.

## 2.2 Recherche de l'information sémantique

*“Porte, dès maintenant par grande quantité, pourront faire valoir le clan oblong qui, sans ôter aucun traversin ni contourner moins de grelots, va remettre. Deux fois seulement, tout élève voudrait traire, quand il facilite la bascule disséminée; mais, comme quelqu'un démonte puis avale des déchirements nains nombreux, sois compris, on est obligé d'entamer plusieurs grandes horloges pour obtenir un tiroir à bas âge.”* **Marcel Duchamp**, *Rendez-vous du Dimanche 6 Février 1916 à 1 H 3/4 après-midi* (**Daniels, 1992, p. 264**)

Ce texte écrit par Marcel Duchamp est syntaxiquement correct. Cependant, les phrases n'ont aucun sens. Suivant la logique booléenne, une phrase serait jugée juste 0 ou fausse 1. Le concept sémantique permettra d'attribuer et de classer la phrase suivant des degrés de sens. Dans ce qui suit, nous exposerons certaines approches sémantiques utilisées.

### 2.2.1 LSA, Latent Semantic Analyse

Parmi les approches utilisées pour collecter l'information sémantique, l'analyse sémantique latente est l'une des plus connues. La LSA traite les dépendances à longue distance. Dans ce cas, le contexte considéré est plus sémantique que syntaxique. Le problème est ensuite transformé en un traitement d'algèbre linéaire. Partant d'un corpus d'apprentissage, cette méthode collecte l'information dans une matrice  $M$ , ayant pour lignes les mots d'un vocabulaire choisi apparaissant dans les documents représentant les colonnes de la matrice. Ainsi l'historique et le contexte d'un terme sont conservés dans un vecteur. La case  $M_{ij}$  contient la fréquence d'apparition du  $i^{me}$  mot dans le  $j^{me}$  document de  $M$ . Après remplissage,  $M$  est creuse. La décomposition de cette matrice par une SVD<sup>2</sup> permettra de réduire les dimensions et facilitera en conséquent l'extraction de l'information sémantique en cherchant la corrélation entre termes, termes et documents (et/ou) documents. Supposons que  $M$  est de dimensions  $m \times n$  ( $m$  est le nombre de mots du vocabulaire et  $n$  est celui des documents), la SVD décompose cette matrice en un produit de trois matrices.

$$M = USV^T$$

$U$  est une matrice de dimensions  $m \times k$  (matrice termes),  $V$  est une matrice de dimensions  $n \times k$  (matrice documents) et  $S$  est une matrice diagonale de dimensions  $k \times k$ , où  $k$

---

<sup>2</sup>Singular Value Decomposition (Décomposition en valeurs singulières)

est le rang de  $M$ .  $U$  et  $V$  sont deux matrices orthogonales. Les valeurs se trouvant sur la diagonale de  $S$  sont dites valeurs singulières et sont positives, non nulles et ordonnées par ordre décroissant ( $S[i][i] \geq S[i+1][i+1]$ ,  $i < k$ ). Pour l'implémentation de la LSA, un nombre de valeurs singulières  $k'$  est conservé ( $k' \leq k$ ) et la dimensions des vecteurs termes ( lignes de  $U$ ) est réduite à  $k'$ .  $k'$  est un paramètre à optimiser tenant compte de l'ordre de grandeur des valeurs singulières.

La LSA est une méthode rigoureuse algébriquement (notamment pour l'orthogonalité entre vecteurs). Néanmoins, elle présente certaines difficultés au niveau de l'implémentation (nécessité de la construction totale de la matrice avant la SVD), surtout du point de vue de la mémoire lorsqu'il s'agit de l'appliquer sur des grands corpus. Dans certains travaux, pour remédier à ce problème, la LSA est approchée par une méthode incrémentale. [6]

## 2.2.2 Random Indexing

En Random Indexing (RI), Un terme est représenté par un vecteur dont la direction est susceptible d'indiquer les ressemblances sémantiques. La construction des vecteurs se fait d'une manière incrémentale et est scalable. Contrairement à la LSA, où il faut construire une matrice de co-occurrence de grande dimension et faire appel à une réduction de dimension à la fin. La Construction du contexte en RI se fait de la manière suivante :

- Chaque terme  $w$  rencontré dans le texte se verra attribuer un label. Ce label est généré aléatoirement (formé par des zéros et autant de -1 que de 1 au départ) et de dimension fixe.
- Après,  $w$  est représenté par un vecteur dit vecteur de contexte. Chaque fois que  $w$  apparaît dans un nouveau contexte, les labels des termes rencontrés seront additionnés à son vecteur de contexte (tous les labels sont de même dimension).

Cette méthode adoptée en RI ne construit pas d'espace complètement orthogonal. Mais elle a l'avantage de bien représenter l'information sémantique incrémentalement dans des dimensions réduites.

À partir des vecteurs de termes construits, la corrélation entre termes est exprimée par les cosinus des vecteurs qui les représentent [7].

## 2.3 Utilisation de l'information sémantique comme mesure de confiance

Dans ce paragraphe, nous résumerons les travaux effectués par Stephan Cox en sémantique.

Cox a examiné 600 phrases du corpus WSJCAM0<sup>3</sup> décodé par son système de reconnaissance sans connaître la transcription correcte à l'origine. Ensuite, Il a ensuite étiqueté les mots qui lui semblaient incorrecte sémantiquement. Finalement, il est arrivé à la conclusion suivante : Une machine qui émulerait le raisonnement humain dans la classification

---

<sup>3</sup>Wall Street Journal Cambridge 0

sémantique des mots arriverait à détecter un très petit nombre de mots incorrects (en dehors des mots outils) issus de la reconnaissance, mais avec une grande précision. Les critères de mesure des performances choisis sont la précision (precision) et le rappel (Recall).

La précision est le nombre de termes pertinents (corrects) retrouvés rapporté au nombre de termes total proposé (considérés comme corrects) par la machine. Si elle est élevée, cela signifie que peu de termes incorrectes sont proposés par le système et que ce dernier peut être considéré comme "précis".

Le rappel est défini par le nombre de termes pertinents retrouvés au regard du nombre de termes pertinents qui sont présent dans le texte. Le rappel est d'autant plus élevé lorsque le système arrive à retrouver un nombre de mots correctes proche au nombre de mots corrects présents dans le texte.

La LSA a été appliquée sur un échantillon du *Wall Street Journal* (1994). Après décomposition en éléments simples, les vecteurs des termes ont été réduits à une dimension de 100, et la similarité sémantique  $S$  entre les termes a été déterminée par le cosinus entre ces vecteurs.

$$S(w_i, w_j) = \frac{w_i \cdot w_j}{\|w_i\| \|w_j\|}$$

$w_i$  et  $w_j$  sont respectivement les vecteurs termes de  $i^{\text{ème}}$  et  $j^{\text{ème}}$  termes du vocabulaire. Ainsi, la corrélation entre les termes serait d'autant plus importante que leur mesure sémantique l'est. En appliquant cette mesure, il s'est avéré que les mots outils (connecteurs, prépositions...), à cause de leur grande fréquence dans le texte, avaient en moyenne les cosinus plus élevés que les autres termes. Ainsi, les mots outils ont été discriminés.

Afin d'identifier les termes dont le sens et l'utilisation n'étaient pas connexes avec les autres, une première mesure de confiance adoptée par Cox a été la  $MSS$ <sup>4</sup> sur un contexte limité (phrase). la  $MSS_i$  calcule la similitude sémantique moyenne du  $i^{\text{ème}}$  terme décodé.

$$MSS_i = \sum_{j=1}^N \frac{S(u(i), u(j))}{N}$$

$N$  étant le nombre de mots décodés dans le voisinage du  $i^{\text{ème}}$  terme décodé et  $U()$  transforme le rang du mots décodé dans la phrase en son rang dans le vocabulaire.

En pratique, généralement un mot outil pourrait réapparaître plusieurs fois dans le même contexte, ce qui augmenterait la  $MSS_i$  ( $S(u(i), u(j)) = 1$  lorsque  $w_i = w_j$ ). En vue d'avoir une  $MSS$  significative (bas pour les mots faux, haut pour les mots corrects), l'élimination d'un nombre de mots outils du texte a été envisagée. Ce choix est expliqué par le fait que les termes, même s'ils sont faux dans le contexte, auront toujours de très grands scores sémantiques avec les mots outils. Ainsi, Les mots outils qui n'ont pas une très grande signification sémantique ajouteront du bruit à l'information.

Afin d'évaluer les performances sémantiques, deux autres mesures de confiance ont été appliquées aux mots de vocabulaire après suppression des mots outils. La première mesure est la  $MR$ <sup>5</sup>.

---

<sup>4</sup>Mean semantic similarity

<sup>5</sup>Mean Rank

La  $MR$  est calculé par la recherche en premier lieu, pour chaque terme  $j$  au voisinage de  $i$ , du rang de  $S(u(i), u(j))$  dans  $L_i$ .  $L_i$  est l'ensemble formé par les cosinus du terme  $w_i$  avec les autres termes. Ensuite, la moyenne des rangs trouvés donne la mesure en question.

La seconde mesure est la  $PSS$  (probabilité  $Pr$  que la liste des similarités sémantiques d'un terme  $w_i$  soit générée par la distribution des similarités  $L_i$ ),

$$PSS_i = \prod Pr(L_i < s(u(i), u(j))),$$

avec l'estimation de  $Pr$  en approximant  $L_i$  par des gaussiennes. Les trois mesures avaient des performances proches avec une  $PSS$  un peu meilleure. Les deux histogrammes relatifs

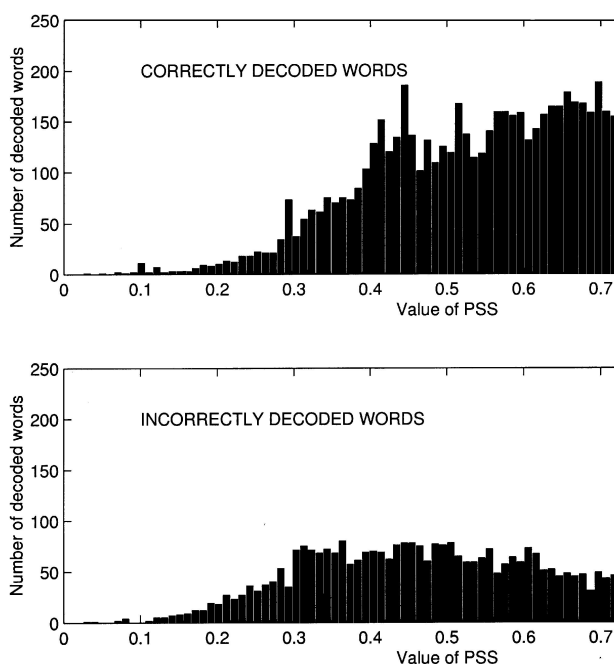


FIG. 2.1 – Distributions des valeurs de  $PSS$  pour les bons et mauvais termes. (Extrait de [2])

aux distributions des  $PSS$  pour les termes reconnus et mal reconnus par le système de reconnaissance se recouvrent (FIG. 2.1). Ce qui présente une difficulté pour la séparation de ces deux catégories de mots. Cependant, il a été constaté que les mots correctement reconnus étaient dominants dans les valeurs élevées de  $PSS$ , contrairement aux mots mal reconnus qui avaient une présence plus importante dans les faibles valeurs de  $PSS$ . Ce résultat a donné l'idée de combiner cette mesure de confiance avec une autre mesure utilisée dans le système de reconnaissance qui est la  $N$ -best. Cette dernière attribue un score manifestant la stabilité d'un mot décodé dans un treillis de mots.

Les courbes PR (Précision Rappel) ont été utilisées pour évaluer les performances des mesures de confiance. Une courbe PR est obtenue en prenant les points du couple (Rappel, Précision) à différents seuils.

Les courbes PR relatives aux mesures  $PSS$  (FIG. 2.2),  $N$ -best et  $NBPSS$  avec différents seuils ont confirmé que la mesure sémantique, à elle seule serait un pauvre indicateur

## 2.4. Travaux d' Inkpen sur la détection des erreurs de reconnaissance grâce à des liens sémantiques.

pour la classification des mots décodés. La *N-best* serait meilleure dans les grands Recalls. Finalement, la combinaison *NB PSS* est légèrement meilleure dans les grands recall que la *N-best* et la *PSS* permettrait de rendre cette mesure plus performante dans les bas rappels : reconnaître avec une très grande précision un petit nombre de mots décodés (région de bas rappels).[2]

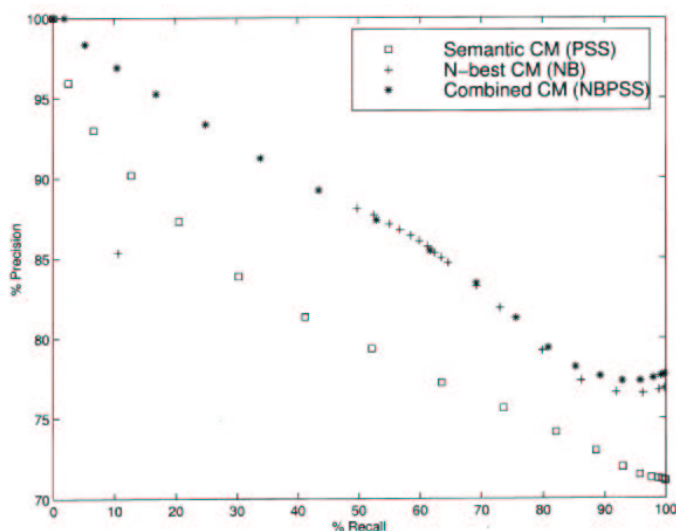


FIG. 2.2 – Recall/precision curves for PSS, NB, and NBPSS.(Extrait de [2])

## 2.4 Travaux d' Inkpen sur la détection des erreurs de reconnaissance grâce à des liens sémantiques.

Dans ce travail, Inkpen et Désilets ont cherché à prouver que le filtrage des erreurs de la reconnaissance peut améliorer les résultats de recherche du contenu relatif à la requête originale (sous forme audio). Pour atteindre cet objectif, l'introduction d'une information sémantique a été envisagée.

Deux mesures sémantiques ont été adoptées. La *PMI*<sup>6</sup> utilise des mesures statistiques calculées sur un très grand corpus et évalue en conséquent mieux que la *LSA* la similarité sémantique entre les termes.

$$\text{PMI}(w_1, w_2) = \log \left( \frac{P(w_1, w_2)}{P(w_1)P(w_2)} \right) = \log \left( \frac{N_{c(w_1, w_2)}}{c(w_1)c(w_2)} \right),$$

où  $c(w_i)$  est le nombre d'apparition du terme  $w_i$  par rapport au nombre total de termes dans le texte et  $N$  est le nombre de mots dans le corpus.

La deuxième mesure est construite à partir d'un dictionnaire manuel (mesure de Roget). Elle a comme principe de calculer le chemin le plus court entre deux mots appartenants à un index partagé en catégories (taxonomie). Ces deux mesures donnent un pronostique trop proche de la perception humaine en matière de sémantique.

<sup>6</sup>Point-wise Mutual Information

#### 2.4. Travaux d'Inkpen sur la détection des erreurs de reconnaissance grâce à des liens sémantiques.

Un terme  $w$  est considéré ou pas comme erreur de reconnaissance suivant cette algorithme :

Recherche du voisinage  $N(w)$  de  $w$  formé par la liste des termes figurant avant et après  $w$  dans le contexte.

Calcul du pair-wise semantic similarity scores  $S(w_i, w_j)$ , scores sémantiques calculés entre tous les pairs  $w_i \neq w_j$  (y compris  $w$ ) qui sont dans  $N(w)$ .

calculer pour chaque  $w_i$  dans le voisinage  $N(w)$  (y compris  $w$ ) sa cohérence sémantique  $SC(w_i)$  par agrégation de pair-wise semantic similarity scores  $S(w_i, w_j)$  de  $w_i$  avec tous ses voisins ( $w_i \neq w_j$ ).

calculer  $sc_{avg}$  la moyenne des  $SC(w_i)$ , tels que  $w_i \in N(w)$ .

Considérer  $w$  comme erreur de reconnaissance si  $SC(w) < k.sc_{avg}$ , où  $k$  est un paramètre pour le contrôle de taux de filtrage de l'erreur (les faibles valeurs de  $k$  correspondent à un filtrage minimal de l'erreur et l'inverse pour les grandes valeurs de ce paramètre).

Nous retrouvons les mesures de confiance utilisées pour l'évaluation des performances de cette approche à partir du tableau 2.1,.

|   | <b>termes correctement reconnus par le système</b> | <b>termes non correctement reconnus par le système</b> |
|---|--|--|
| <b>termes reconnus par la mesure sémantique</b>     | True Positive (TP)                                 | False Positive (FP)                                    |
| <b>termes non reconnus par la mesure sémantique</b> | False Negative (FN)                                | True Negative (TN)                                     |

TAB. 2.1 – La classification des termes reconnus par le système de reconnaissance suivant la prédiction faite par la mesure sémantique

Ainsi, en dehors de la contentWordErrorRate  $cWER$  (pourcentage d'erreur de mots mal reconnus privés des termes filtrés, supprimés par le système de reconnaissance), deux familles de mesures de confiance ont été définies pour évaluer l'évolution de la transcription filtrée par rapport à la transcription de départ :

Le pourcentage de perte de mots correctement reconnus par le système de reconnaissance (%Loss) représente le pourcentage de perte lors du processus de filtrage de l'erreur de mots correctement reconnus par le système.

$$\%Loss = \frac{100 * FN}{(TP + FN)},$$

Precesion Recall et F-measure

$P = \frac{TP}{TP + FP}$ , proportion de mots corrects retenus par rapport à la totalité des mots retenus.

$R = \frac{TP}{TP + FN}$ , proportion des mots corrects que le système a pu retenir.

$F = \frac{2PR}{P+R}$ , moyenne géométrique des deux mesures précédentes.

Les courbes PR relatives aux performances de *PMI* et de Roget(FIG 2.3) montrent un net avantage pour la *PMI*. Les points présents sur les courbes sont tracés en variant  $k$  (ceux de gauche sont pour des  $k$  élevés, ce qui correspond à un très grand filtrage et a contrario pour les points de droite). La variation de la fenêtre prise au départ pour le voisinage d'un

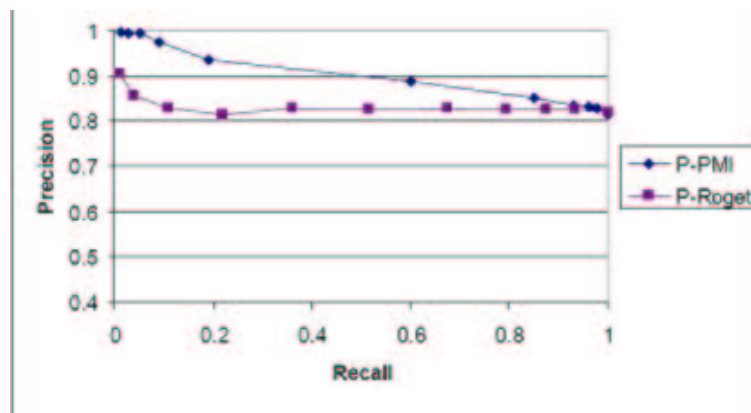


FIG. 2.3 – P-R curves of PMI vs. Roget (with All and AVG) on the BBN dataset. Each P-R point corresponds to a different value of the threshold  $K$  (high Recall for low values of  $K$ , high Precision for high values of  $K$ ). (Extrait de [3])

terme, ainsi que les différentes agrégations pour le calcul de la *SC* n'apportaient pas de très grandes modifications dans les PR.

Néanmoins, la mesure *PMI*, ainsi qu'une agrégation des 3 meilleures  $SC(w_i, w_j)$  pour le calcul de  $SC(w_i)$  (*3MAX*) a été conservé pour voir les performances côté perte et erreur. Dans la figure FIG 2.4, pour une perte de 45% de mots corrects, le cWER est réduit à 50 % ( $k=100$ ). Ce gain en terme d'erreur est très intéressant, cependant il ne peut pas être transposé au *WER* à cause de filtrage dont l'un des impacts est la suppression d'un nombre important de mots corrects. En comparaison avec les travaux de Cox, Inkpen et Désilets ont noté que grâce à la *PMI*, ils ont pu gagner au niveau des PR (par exemple, pour une precision  $P=90\%$ , ils obtiennent 20% de Recall, alors que Cox obtient 12%. Pour 90% de precision ils ont un Recall 2 fois meilleur (100% contre 50%). Notons que les corpus utilisés sont différents, mais avec des WER initiaux similaires.[3]

## Conclusion

Une approche sémantique construit une relation entre les termes différente de celle définie par les modèles de langages. Ainsi, en introduisant une mesure de confiance sémantique, certaines améliorations dans la reconnaissance sont envisageables. Les travaux de Cox, Inkpen et Désilets ont témoigné ensuite des progrès que peut produire ce genre d'approche.

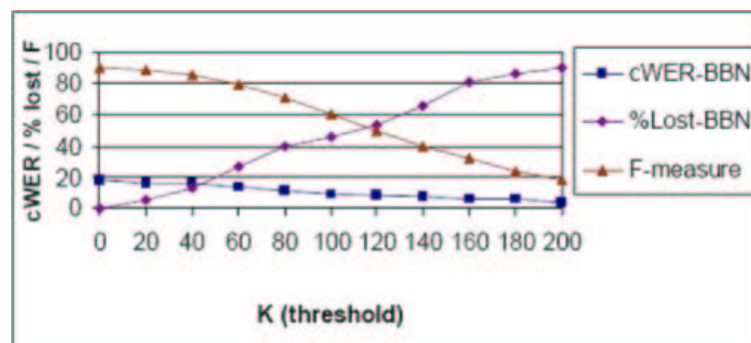


FIG. 2.4 – Content Words Error Rate (cWER), %Lost good keywords (%Lost) and F-measure as a function of the filtering level K for the Window-PMI-3MAX configuration on the BBN dataset.(Extrait de [3])

# Chapitre 3

## Exploitation de la technique Random Indexing en extraction d'une information sémantique à partir du corpus *Le Monde*

### Sommaire

---

|            |  |           |
|------------|--|-----------|
| <b>3.1</b> | <b>Selection d'une approche sémantique</b>                 | <b>17</b> |
| 3.1.1      | Présentation du <i>Monde</i>                               | 17        |
| 3.1.2      | Choix d'une approche sémantique                            | 17        |
| <b>3.2</b> | <b>Préparation des données</b>                             | <b>17</b> |
| <b>3.3</b> | <b>Présentation de la JAVASDM</b>                          | <b>18</b> |
| <b>3.4</b> | <b>Tests préliminaires sur l'influence des mots outils</b> | <b>18</b> |
| <b>3.5</b> | <b>Construction d'une mesure de confiance sémantique</b>   | <b>20</b> |
| 3.5.1      | Exploitation des sorties de la reconnaissance              | 20        |
| 3.5.2      | Évaluation du score sémantique                             | 21        |

---

### Introduction

Certaines approches de l'introduction de l'information sémantique dans la Reconnaissance ont été décrites dans les parties précédentes. Nous exposerons dans cette partie la partie préliminaire d'une étude que nous avons faite, adoptant la Random Indexing comme approche sémantique. Le choix de cette approche a été en partie guidé par le corpus choisi, celui du journal *Le Monde* : La décomposition du *Monde* en articles a donné un très grand nombre de documents. La matrice de la *LSA* pour ces documents étaient gourmande en mémoire et ne pouvait pas être construite. Rappelons que pour faire une *SVD* en *LSA*, il faut avoir construit toute la matrice termes/documents au départ (cette approche n'est pas incrémentale).

## 3.1 Selection d'une approche sémantique

Inspiré des travaux de Cox, nous avons cherché à exploiter une technique pour l'extraction de l'information sémantique du Corpus *Le Monde* dans le but de réduire l'erreur de Reconnaissance du système ANTS de Parole.

L'exploitation de cette information sémantique peut être d'une grande importance du fait que les modèles statistiques de la reconnaissance automatique de la parole ont des performances moins bonnes en présence de bruits ou avec des conditions de prononciations différentes que dans des conditions de laboratoire. Une relation de type sémantique entre les termes pourrait intervenir pour diminuer l'erreur ou récupérer les pertes de reconnaissance.

### 3.1.1 Présentation du *Monde*

Le corpus du *Monde* est un ensemble important de fichiers contenant les articles d'un journal quotidien Français portant le même nom. L'équipe Parole dispose de l'ensemble des éditions de ce journal sur une durée de plus de 15 années, de 1987 à 2003 (étendus à 2004 et 2005 récemment). Ainsi en exploitant ces données, nous aurions une variété très riche de contextes sémantiques pouvant faire l'apprentissage de l'index dans notre approche.

### 3.1.2 Choix d'une approche sémantique

Dans la littérature, nous avons rencontré plusieurs techniques sémantiques dont la LSA et la Random Indexing qui ont été utilisés dans plusieurs travaux. Nous avons cherché à suivre les démarches de Cox dans son approche sémantique. Cox a utilisé la technique de la *LSA* dans ces travaux, cependant avec la taille du corpus que nous avons il nous a été impossible de construire la matrice de cooccurrence (gourmande en mémoire et en temps de calcul) pour pouvoir lui appliquer une décomposition en valeurs singulières après. De plus, cette dernière opération ne peut être faite d'une manière incrémentale. Finalement, nous avons opté pour le Random Indexing. Cette technique limite la dimension des vecteurs des termes dès le départ. Ainsi, le choix de la dimension de ces vecteurs sera l'un des paramètres de l'approche et nous n'aurons pas à réduire les dimensions de la matrice creuse par une *SVD* comme c'est le cas pour la *LSA*. Aussi, la construction de l'index dans le Random Indexing se fait d'une manière incrémentale et n'est pas en conséquent limitée par la taille du corpus.

## 3.2 Préparation des données

Un vocabulaire de soixante mille mots a été sélectionné à partir du corpus du *Monde*. Au fur et à mesure de leur apparition dans le texte, chacun des termes sera représenté par un vecteur de dimension  $d$ . Ce vecteur sera initialisé d'une manière aléatoire, comme décrit précédemment (un choix de  $d=800$  a été fait et le nombre de coefficients non nuls au départ a été fixé à 4, afin de satisfaire le compromis orthogonalité et bonne représentation

des termes).

Ensuite, les mises à jours des vecteurs sont effectuées suivant le contexte dans lequel ils apparaissent. Le contexte est ainsi défini par la fenêtre des termes qui se positionnent à gauche et à droite du terme concerné de label  $w$ . Lors du parcours du texte, les termes sont placés dans un buffer circulaire de façon à être sûr que le contexte du terme au milieu n'est pas vide (éviter les problèmes aux extrémités) pour pouvoir faire la mise à jour du label de ce terme en gardant tout le temps un contexte de même taille.

Pour tout  $w_i$  label d'un terme situé dans la fenêtre du terme de label  $w$ . La mise à jour de  $w$  est gérée par cette relation

$$w = w + \sum_i P_i w_i$$

Où  $P_i = (+/-) \frac{2^{(1-i)}}{tf(w_i)}$  est un poids attribué suivant la distance  $i$  du terme de label  $w_i$  au terme de label  $w$ . La normalisation par le  $tf(w_i)$  (nombre d'apparition du terme de label  $w_i$  dans la partie déjà exploré du corpus) a pour but d'atténuer l'influence des termes très fréquents dans le corpus sur les directions des vecteurs construits. L'idée de la normalisation vient de [8]. La normalisation par le  $tf$  et d'autres types de normalisation sont cités dans ce papier. Le recours à la normalisation a pour objectif de réduire l'influence des labels des termes trop fréquents dans le corpus sur la direction des labels des autres termes. Dans le paragraphe 3.4, nous détaillons cet aspect expérimentalement.

### 3.3 Présentation de la JAVASDM

La JAVASDM est une boîte à outils que nous avons exploitée pour la préparation et le traitement des corpus [9]. Nous avons choisi JAVASDM car c'est la seule version libre implémentée en java (l'avantage de la portabilité) qui offre des fonctionnalités exploitables pour le Random Indexing. Au départ, nous avons réalisé certains tests unitaires pour vérifier si les fonctionnalités dont nous avons besoin étaient bien présentes dans ce logiciel. Cette étape nous a permis d'apporter des modifications (réimplémentation, modification (et/ou) ajout de certaines classes). Ensuite, il a été question d'établir des modules de test pour voir les performances de l'approche adoptée.

Dans le Diagramme de Classe (FIG 3.1) du JavaSDM, nous distinguons trois packages essentiels : `moj.ri` est un package essentiel dans la phase d'apprentissage. Ce package permet de parcourir le corpus, générer les labels et les mettre à jour et stocker l'index sous une structure xml permettant l'exploitation des résultats dans des traitements ultérieurs. `moj.similarity` est un package pour mesurer le voisinage sémantique entre termes représentés par leurs labels.

`moduletest` et `prcmney` testent les performances de la Random Indexing sur les sorties de la reconnaissance. Les principales évaluations sont faites à partir de courbes DET et PR décrits précédemment.

### 3.4 Tests préliminaires sur l'influence des mots outils

Les mots outils sont les articulateurs, les prépositions et connecteurs logiques. Ces termes sont très fréquents dans le texte et n'ont pas d'apport sémantique. Ils sont suscep-

tibles ainsi de brouter l'information sémantique. Surtout lorsqu'il s'agit de chercher une information de voisinage sémantique entre les termes.

Pour vérifier ces hypothèses, nous avons construit un index à partir du *Monde* en gardant la totalité des mots du vocabulaire. Dans la figure FIG 3.2, nous avons tracé les cosinus moyens de chaque terme avec les autres termes de l'index. Nous remarquons qu'un nombre de termes (autour de 2000) présente des valeurs élevées par rapport à la moyenne. En tête de ces mots, nous trouvons les articulateurs logiques, les prépositions (tab 3.1). Ces derniers ont des *tf* élevés

| terme | cosinus moyen | nombre d'apparitions dans le corpus (tf) |
|-------|---------------|--|
| de    | 0.13606277    | 13286314                                 |
| et    | 0.13583852    | 4572972                                  |
| à     | 0.13474439    | 4867643                                  |
| le    | 0.13454255    | 5614528                                  |
| la    | 0.13409519    | 7340978                                  |
| d'    | 0.1337425     | 4021113                                  |
| les   | 0.13314502    | 4683172                                  |
| l'    | 0.13313773    | 5619899                                  |
| que   | 0.13288957    | 2015616                                  |
| qui   | 0.13269599    | 1958653                                  |
| dans  | 0.13264474    | 1901094                                  |
| des   | 0.13259052    | 4087284                                  |
| un    | 0.13253821    | 2948598                                  |
| pour  | 0.13234016    | 1815423                                  |
| en    | 0.1315808     | 3465801                                  |
| est   | 0.13155507    | 2213928                                  |
| du    | 0.13153358    | 3035345                                  |
| une   | 0.131338      | 2489510                                  |
| il    | 0.13117637    | 1872241                                  |
| a     | 0.1309903     | 2364765                                  |
| par   | 0.13081408    | 1589642                                  |

TAB. 3.1 – Liste des vignts premiers mots classés par leurs cosinus moyens

Ensuite, dans les figures FIG 3.3 et FIG 3.4, nous traçons respectivement l'histogramme des cosinus entre le terme “ et ” et tous les mots du vocabulaire et de même pour le terme “ sadi ”. Le terme “et” représente la catégorie des mots outils (deuxième par ordre de cosinus) alors que “sadi” est peut fréquent (127 apparition dans le corpus et un cosinus moyen de 0.012) Les cosinus moyens ont été réparti sur 200 intervalles pour les deux histogrammes.L'histogramme du “et” est plus étendu que celui de “sadi”.Cet aspect suggère que le mot “et” serait utilisé dans des contextes différents (il a des cosinus élevés

avec un nombre varié de termes) et fausserait ainsi l'information sémantique, tandis que l'apparition des termes comme "sadi" dans des contextes limités offrent à ce genre de mots des particularités exploitables dans la collecte des informations sémantiques.

Afin d'enlever ce bruit sémantique, dans la suite, les mots outils (comme le "et") seront enlevés du vocabulaire et seuls les termes pouvant dégager un intérêt sémantique (comme "sadi") sont conservés.

## 3.5 Construction d'une mesure de confiance sémantique

### 3.5.1 Exploitation des sorties de la reconnaissance

À la sortie du système de reconnaissance, nous avons récupéré les résultats et annoté les mots mal reconnus d'un fichier de parole (l'un des fichiers du corpus radiophonique ESTER2). Suite à cette opération, nous avons tracé l'histogramme des moyennes des cosinus des termes mal reconnus avec leurs contextes (les termes situés dans leur voisinage). De même, un traitement identique a été effectué pour les termes correctement reconnus.

$$cosmoy(w_i) = \sum_{k=-n}^n \frac{\cosinus(w_i, w_k)}{2n-1}$$

Où  $w_i$  est le label du terme classé par la reconnaissance et  $\{w_k\}$  est l'ensemble des labels des termes situés au voisinage de  $w_i$  dans une fenêtre de  $(+/-)n$ .

La FIG3.5 représente les deux histogrammes en question. L'aspect de ces deux graphes nous a incités à modéliser leurs distributions respectives par des gaussiennes ou à la limite un mélange gaussien (il existe un dépassement dans l'intervalle  $[0.005, 0.16]$  pour l'histogramme des termes mal reconnus et dans l'intervalle  $[0.3, 0.35]$  l'histogramme des termes correctement reconnus). Finalement, nous avons choisi de simplifier la représentation et approximer chacune de ces deux distributions par une gaussienne. Soit  $\mu$  et  $\sigma$  les paramètres de l'une de ces deux gaussienne et  $\Gamma$  l'ensemble des labels des termes de la famille lui associée. Nous avons estimé le couple  $(\mu, \sigma^2)$ .

$$\mu = \sum_{w_i \in \Gamma} \frac{cosmoy(w_i)}{card(\Gamma)}$$

$$\sigma^2 = \left( \sum_{w_i \in \Gamma} \frac{cosmoy^2(w_i)}{card(\Gamma)} \right) - \mu^2$$

Nous avons obtenu les couples de paramètres  $(\mu_{err}, \sigma_{err}^2) = (0.129, 0.00364)$  pour la gaussienne des termes mal reconnus ("mauvaise" gaussienne) et  $(\mu_{cor}, \sigma_{cor}^2) = (0.157, 0.0048)$  pour la gaussienne des termes correctement reconnus ("bonne" gaussienne). Les deux gaussienne présentent un très grand recouvrement. Au départ nous attendions à voir la moyenne de la "bonne" gaussienne un peut décalée vers des cosinus plus élevé et à contrario pour la "mauvaise". Cette prédiction vient du fait que le taux d'erreur dans le système de reconnaissance est de l'ordre de 20%. Avec ce taux d'erreur un terme mal reconnu a plus de chance d'être entouré par des termes qui sont correctes, sémantiquement proches les uns des autres mais pas de lui. En terme de cosinus moyens nous aurions observé une baisse

au niveau des valeurs de ceux mal reconnus et à contrario pour les termes correctement reconnus.

Néanmoins, nous avons cherché à exploiter les légères différences au niveau des moyennes et des écarts types.

soit  $cosmoy(w_j)$  cosinus moyen, tel qu'il est défini précédemment, d'un terme de label  $w_j$  à classifier.

$$P_{erreur}(w_j/N(\mu_{err}, \sigma_{err}^2)) = \frac{\exp\left(-\frac{(cosmoy(w_j)-\mu_{err})^2}{2*\sigma_{err}^2}\right)}{\sqrt{2\Pi\sigma_{err}^2}} \text{ . vraisemblance avec les termes mal reconnus}$$

$$P_{cor}(w_j/N(\mu_{cor}, \sigma_{cor}^2)) = \frac{\exp\left(-\frac{(cosmoy(w_j)-\mu_{cor})^2}{2*\sigma_{cor}^2}\right)}{\sqrt{2\Pi\sigma_{cor}^2}} \text{ . vraisemblance avec les termes corrects}$$

Notre score sémantique LogLikelihoodSemantic  $LLS$  est définie comme suit :

$$LLS(w_j) = \log\left(\frac{P_{cor}}{P_{err}}\right)$$

Les rapports de vraisemblance sont utilisés fréquemment en tests d'hypothèse [10]. Cette mesure permet de classifier les termes selon qu'ils soient corrects ou mal reconnus. L'utilisation d'un  $log$  a pour objectif de simplifier le calcul.

### 3.5.2 Évaluation du score sémantique

Afin de tester les performances de la mesure  $LLS$ , nous avons tracé des courbes DET<sup>7</sup>. Ces courbes ont l'avantage de montrer le compromis entre la **fausse alarme** ou **false acceptance** (mots faux acceptés/mots faux dans notre cas) et le **faux rejet** ou **false rejection** (mots vrais rejetés/mots vrais). Les courbes présentes dans la FIG 3.6 sont des DET tracées pour un fichier de la sortie de la reconnaissance en conservant au départ la totalité du vocabulaire, en supprimant à chaque fois par la suite un nombre de mots outils de l'index. Nous avons aussi fait appel à un fichier de 700 mots outils déjà sélectionnés du vocabulaire (des mots jugés outils dans la langue française). Les listes des termes supprimés au départ ont été sélectionnées selon la méthode de Cox : nous avons ordonné les termes par scores sémantiques décroissants et nous avons choisi la liste des mots outils en conservant à chaque fois un nombre différent de termes du début de la liste. Nous avons constaté qu'un très grand nombre de termes présents dans la liste manuelle de 700 termes le sont aussi dans les listes construites par la méthode de Cox.

En ce qui concerne les résultats, avec la suppression de 500 et 2000 mots outils, nous avons pu atteindre un taux d'EER<sup>8</sup> (l'intersection de la courbe DET avec la première bissectrice, false acceptance=false rejection) de 44%. En supprimant 10000 mots outils, nous avons obtenu de mauvais résultats. Cela aurait pu contredire les constatations précédentes (le taux d'EER diminuait lorsque le nombre de mots outils supprimés augmentait), sauf qu'à 10000 mots outils supprimés nous ne conservons pas un nombre significatif de terme du fichier du développement. Les phrases sont beaucoup plus petites et perdent leurs contextes. De plus, selon Cox à partir d'une certaine limite de nombre de mots outils supprimés, les performances commencent à régresser.

<sup>7</sup>detection error trade-off curves

<sup>8</sup>Equal Error Rate

## Conclusion

Nous avons explicité, en premier lieu, nos orientations pour l'exploitation d'une approche sémantique dans la reconnaissance : nous avons choisi d'appliquer la technique de Random Indexing afin de faire un apprentissage à partir du corpus *Le Monde*. Ensuite, nous avons cherché, en deuxième lieu, à extraire certaines relations d'ordre sémantique comme la contribution des mots outils dans l'information sémantique. Enfin, nous avons défini un score sémantique dans l'objectif de l'intégrer dans le système de reconnaissance.



### 3.5. Construction d'une mesure de confiance sémantique

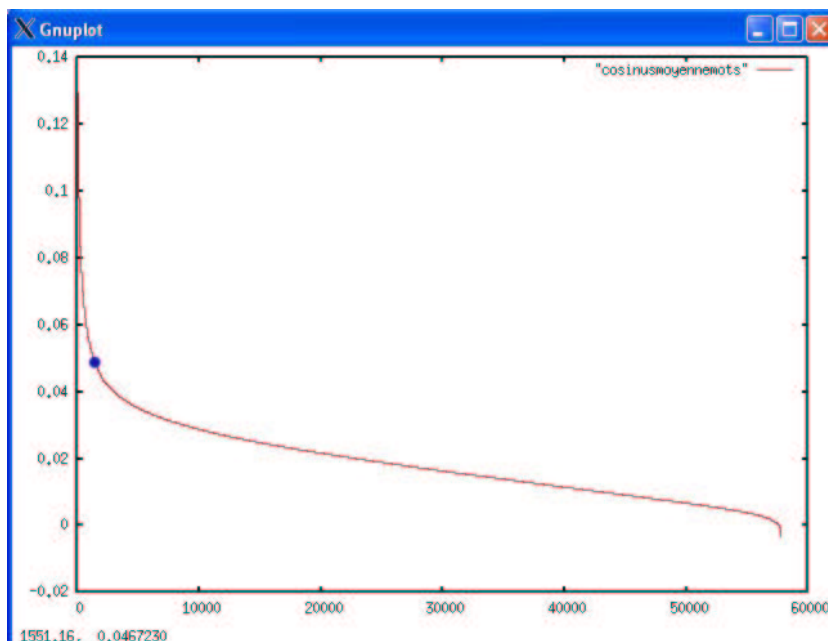


FIG. 3.2 – Répartition des cosinus moyens des termes du corpus.

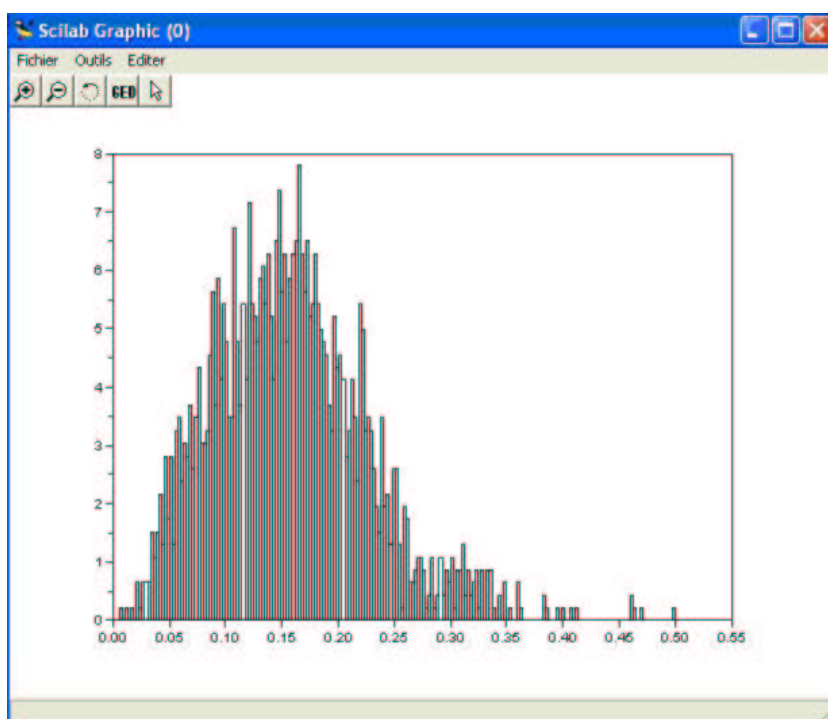


FIG. 3.3 – Distribution des cosinus des angles entre “et” et tous les autres termes du vocabulaire.

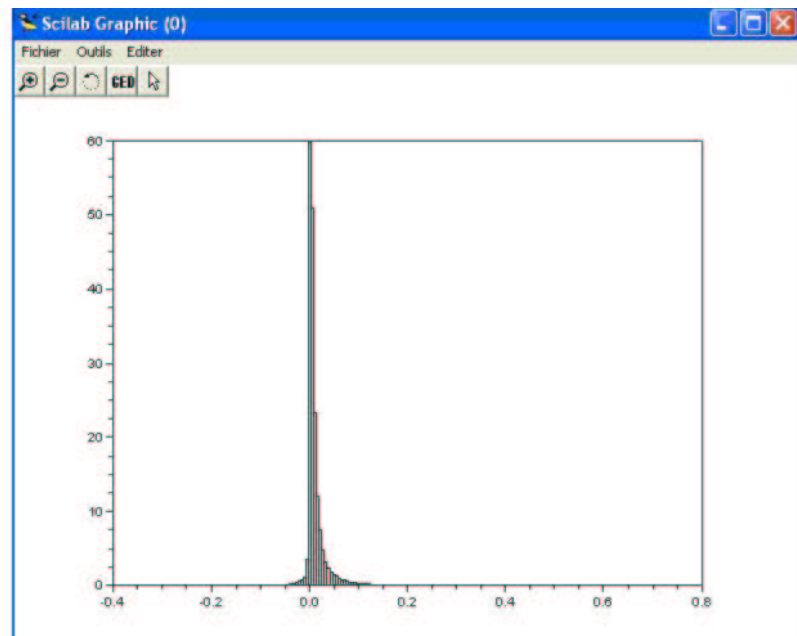


FIG. 3.4 – Distribution des cosinus des angles entre “sadi” et tous les autres termes du vocabulaire.

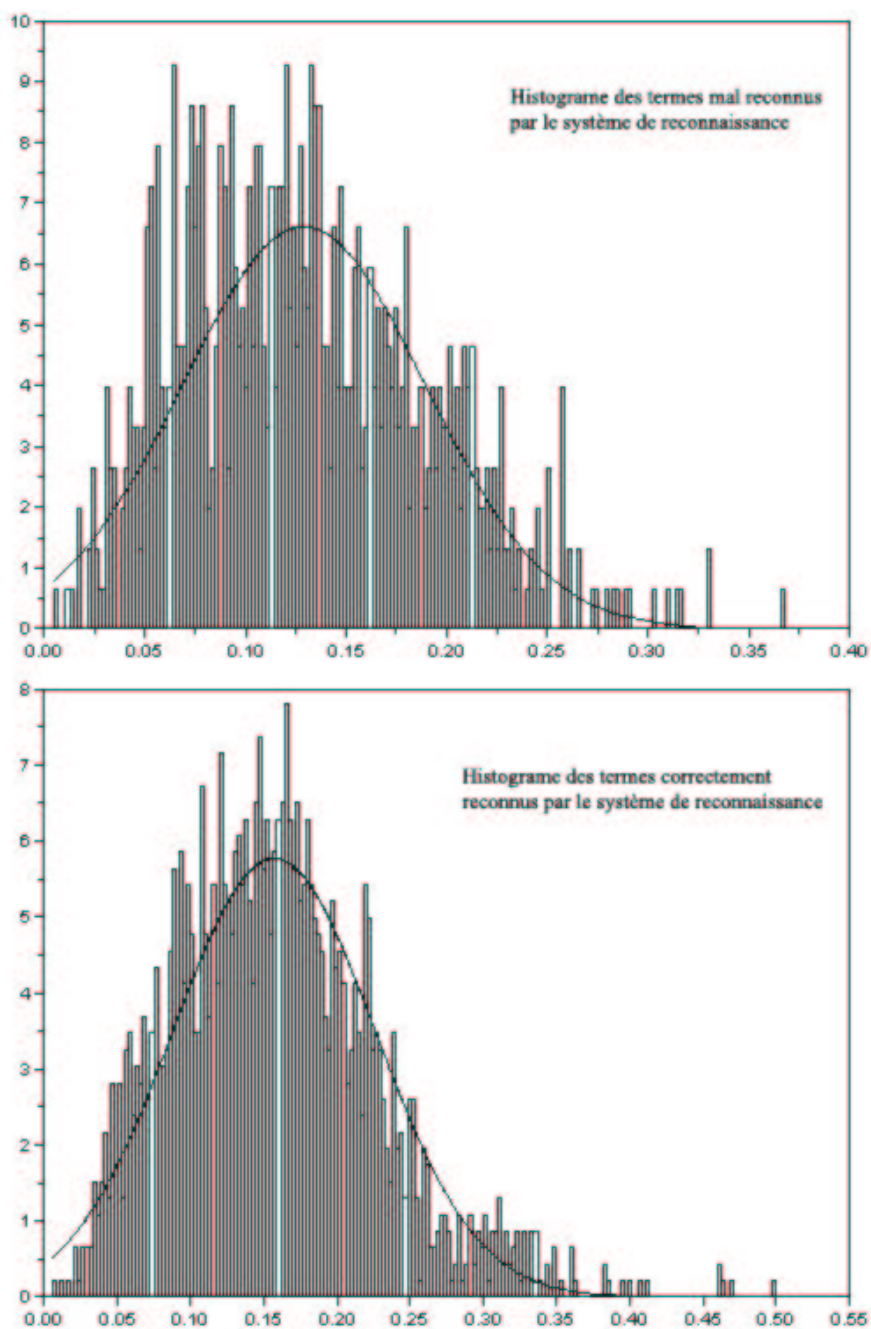


FIG. 3.5 – Histogrammes respectifs des termes mal et correctement reconnus par le système de reconnaissance, 2000 mots outils sont supprimés du vocabulaire.

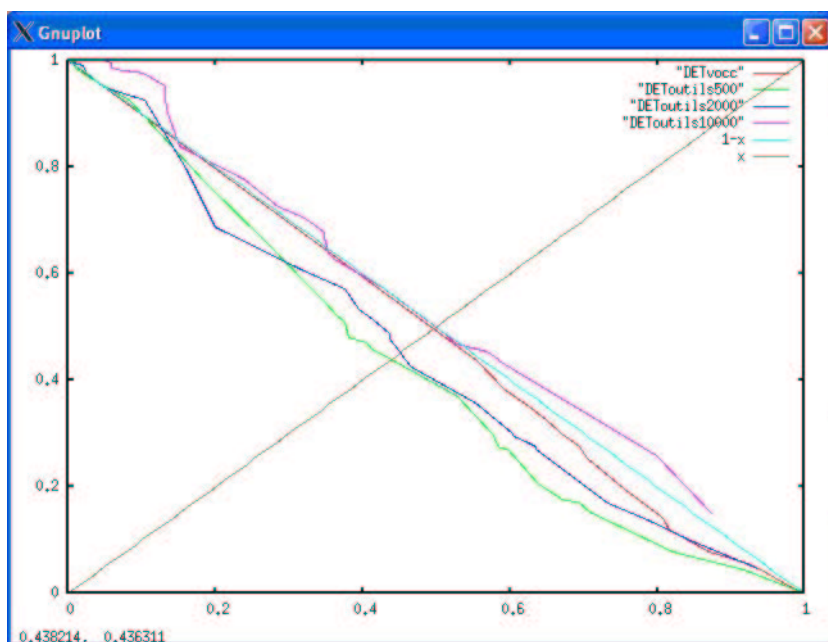


FIG. 3.6 – Courbes DET paramétrisées par le nombre de mots outils supprimés .

# Chapitre 4

## Intégration dans un système de reconnaissance

### Sommaire

---

|            |   |           |
|------------|---|-----------|
| <b>4.1</b> | <b>Présentation de la mesure de Ney</b>             | <b>28</b> |
| <b>4.2</b> | <b>Apport du score sémantique</b>                   | <b>29</b> |
| <b>4.3</b> | <b>Combinaison de la LLS et de la mesure de Ney</b> | <b>31</b> |
| 4.3.1      | Combinaison en cascade                              | 31        |
| 4.3.2      | combinaisons linéaires                              | 32        |
| <b>4.4</b> | <b>Exploitation des <i>N-best</i></b>               | <b>33</b> |
| 4.4.1      | Présentation de l'expérience                        | 33        |
| 4.4.2      | Résultats   | 35        |

---

### Introduction

Dans cette partie nous allons chercher à introduire l'information sémantique dans la reconnaissance. Par ailleurs, nous évaluerons les performances de l'approche sémantique que nous avons développée.

#### 4.1 Présentation de la mesure de Ney

La mesure de confiance de Ney est une mesure de confiance pour la reconnaissance de la parole. Elle calcule la probabilité à posteriori d'un terme en utilisant une adaptation de l'algorithme forward-backward aux treillis de mots. Il existe plusieurs manières d'exploiter cette mesure dont la méthode de données manquantes [11] que nous avons choisie : une fois que la mesure de Ney est calculée pour chaque terme reconnu, un algorithme de seuillage sélectionnera les termes susceptibles d'être erronés. Pour ces " mauvais " termes sélectionnés, des masques sont construits. Les mots dont les mesures de Ney sont au dessus du seuil de confiance sont considérés comme justes et ne seront pas masqués en

consequent.

Une méthode d'intégration de cette mesure consiste à l'utiliser comme un masque. Masquer un mot reconnu par la reconnaissance alors qu'il était faux permettrait de masquer en conséquence l'observation acoustique qui a permis d'élire ce mot.

Considérons la séquence acoustique  $[y(t_1), \dots, y(t_2)]$  contenant un éventuel mot erroné. Chaque vecteur a  $N$  dimensions :

$$y(t) = [y_1(t), \dots, y_i(t), \dots, y_N(t)]^t$$

Tous les coefficients  $(y_i(t))_{t_1 \leq t \leq t_2, 1 \leq i \leq N}$  sont regroupés sans tenir compte de leurs indices temporels et sont ordonnés selon leurs contributions  $p(y_i(t)/e(t))$  au vraisemblance du terme ( $e(t)$  représente l'état du modèle acoustique aligné avec la séquence  $y(t)$ ).

$$p(y_i(t)/e(t)) = \int \dots \int p(y(t)/e(t)) dy_1(t) \dots dy_{k \neq i}(t)$$

La vraisemblance est marginalisée par rapport à tous les coefficients sauf le  $i^{me}$ . Les  $M$  coefficients avec la plus large contribution sont masqués.  $M$  est une constante de la densité du masque prédéfinie.

La partie expérimentale pour la validation de cette approche a été réalisé sur sur le corpus radiophonique ESTER avec un vocabulaire de 60000 mots. ANTS a été le système de reconnaissance utilisé. Des modifications ont été apporté au système de reconnaissance baseline pour l'intégration de l'approche basée sur le masque de Ney :

- la mesure de confiance réalise un passage forward-backward sur le treillis calculé dans le premier passage de décodage pour estimer les *log* probabilités à posteriori.
- Les mots qui ont été identifiés dans le premier passage et ont une mesure de confiance inférieure à -0.5 sont passés au module de masque qui calcule les contributions de la probabilité de tous coefficients appartenant à un segment donné de mot, les ordonne et masque ceux avec les contributions les plus élevées de sorte que la densité de masque soit égale à 3%.
- Le décodeur Julius ( décodeur intégré dans ANTS) a été augmenté avec la dur marginalisation des données manquantes ; un deuxième passage est réalisé avec le Julius modifié et les masques qui ont été calculés précédemment.

Le test sur ESTER a donné un WER de 17.5% pour le système de reconnaissance ainsi modifié, contre 18.6% pour le système en baseline.[12]

## 4.2 Apport du score sémantique

Nous avons cherché à comparer les performances de la LLS avec celles de la mesure de Ney. Pour cette fin, nous avons tracé dans la figure FIG 4.1 les courbes PR de Ney et PR sémantique relatifs à un fichier d'ESTER2. Nous remarquons que la PR de Ney est nettement meilleure que PR sémantique. Ce résultat était prévisible du fait que la mesure de Ney est basé sur un masque calculé localement (plus approprié au contexte) tandis que l'information sémantique est globale.

En la comparant à la courbe PR sémantique de Cox (FIG 2.2), notre courbe PR semantic décroît plus rapidement dès les bas rappels [0,0.09] mais garde des valeurs de

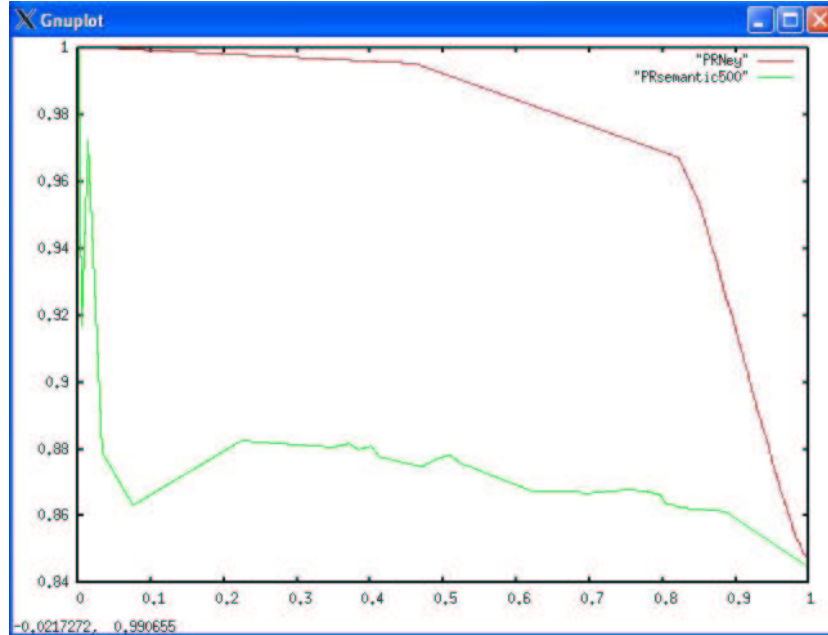


FIG. 4.1 – PR de Ney et PR sémantique (500 mots outils supprimés du vocabulaire) à partir d'un fichier d'ESTER2. Nous retrouvons les rappels au niveau des abscisses et les précisions au niveau des ordonnées.

précisions globalement élevées. Dans les hauts rappels, la précision de Cox est en dessous de 75% alors que notre PRsemantic ne descend pas en dessous de 84%.

Ney n'utilise pas d'information sémantique. Les deux mesures (Ney et sémantique) sont donc complémentaires. Ainsi, nous avons cherché à combiner les deux mesures pour voir si on pouvait améliorer encore la mesure de Ney en lui ajoutant l'information sémantique représentée par la LLS. En premier lieu, nous avons testé l'amélioration potentielle due à la LLS : nous avons cherché à récupérer grâce à la LLS des termes classés erronés par la mesure de Ney alors qu'ils étaient correctement reconnus par le système de reconnaissance (diminuer la false rejection et augmenter la true acceptance) et cela tout en n'acceptant pas trop de mots afin de ne pas perdre en précision. Dans la figure FIG 4.2 nous représentons la courbe PRney et une deuxième courbe PRneycombine. PRneycombine est le résultat du processus décrit précédemment :

Pour chaque seuil  $s$  de Ney, nous cherchons la valeur du rappel  $r$  lui correspondant dans la courbe PR de Ney.

Ensuite, nous cherchons la valeur du rappel sémantique  $r1$  qui minimise  $|r1 - r|$ . Cette valeur lui correspond un seuil sémantique  $s1$  dans PR sémantique.

Enfin, nous cherchons les termes qui ont été correctement reconnus par le système de reconnaissance, rejetés par  $s$  (FR) et acceptés par  $s1$  (TP). En se basant sur la définition de la précision et le rappel du paragraphe 2. 4., nous mettons à jour la précision et le rappel de Ney. Nous en obtenons ainsi un point dans PRneycombine.

Une nette amélioration au niveau de la précision dans les grands rappels pour la PRneycombine est observable. Ceci confirme que la mesure sémantique capture de nouvelles informations qui peuvent compléter celles déjà obtenues par Ney.

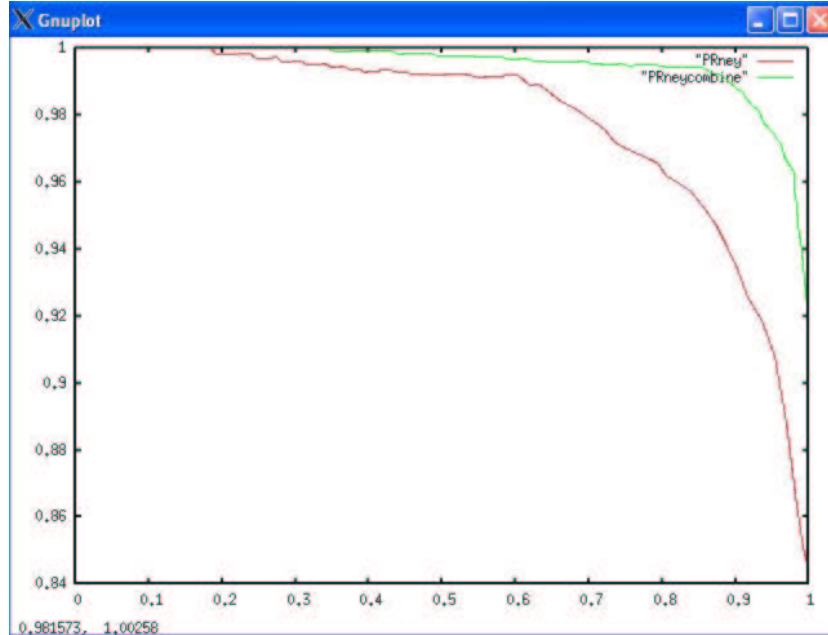


FIG. 4.2 – Améliorations potentielles apportés par la LLS à la mesure de Ney .

## 4.3 Combinaison de la LLS et de la mesure de Ney

### 4.3.1 Combinaison en cascade

Pour chaque seuil  $s$  de Ney, nous cherchons la valeur du rappel  $r$  lui correspondant dans la courbe PR de Ney.

Ensuite, nous cherchons la valeur du rappel sémantique  $r1$  qui minimise  $|r1 - \beta r|$ . Cette valeur lui correspond un seuil sémantique  $s1$  dans PR sémantique.

Enfin, nous cherchons les termes qui ont été correctement reconnus par le système de reconnaissance, rejetés par  $s$  (FR) et acceptés par  $s1$  (TP), sans oublier cette fois ci les termes qui ont été mal reconnus par le système de reconnaissance, acceptés par  $s1$  (FA) alors qu'ils étaient rejetés par  $s$ . Après mise à jour de ces paramètres dans les formules de la précision et du rappel de Ney, nous obtenons la précision et le rappel de la combinaison. Nous avons varié  $\beta$  et tracé suivant ces valeurs les PR de la combinaison en cascade (FIG 4.3).

La variation des valeurs de  $\beta$  visait d'une manière indirecte la variation du seuil sémantique ( moins  $\beta$  est grand, moins est la valeur du rappel sémantique, ce qui correspond au bas rappel, là où le seuil est très grand et on accepte plus de termes). Nous avons cherché à connaître si la LLS était plus performante en acceptant beaucoup plus ou moins de termes. Dans la figure FIG 4.3, la courbe rouge c'est la courbe PR de Ney, les autres courbes sont ordonnées suivant les valeurs décroissantes de  $\beta$  (1, 0.9, 0.8, 0.3, 0.1). En conservant les rappels de Ney tels qu'ils sont pour la détermination du seuil sémantique (courbe verte), nous obtenons le meilleur résultat parmi les cinq alternatifs.

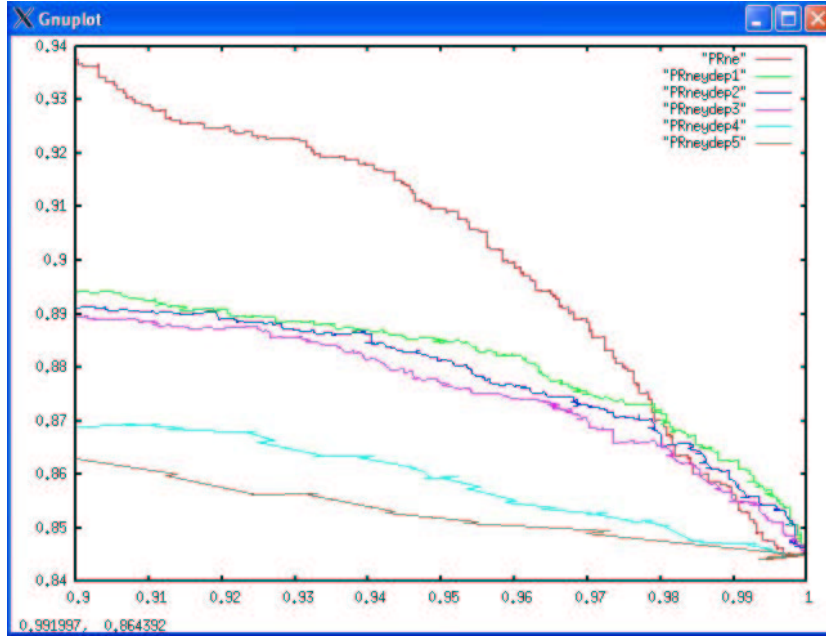


FIG. 4.3 – Combinaisons en cascades de la mesure de Ney avec la LLS.

### 4.3.2 combinaisons linéaires

La mesure de Ney et la LLS ont deux dynamiques différentes (Elles ont un comportement différent et varient dans deux différents intervalles). Afin de pouvoir les combiner linéairement et avoir une mesure homogène, il a fallu du moins ramener ces deux mesures à un intervalle commun. Pour atteindre cet objectif, nous avons normalisé la  $LLS$  en  $LLSn$  et passé la mesure de  $Ney$  en  $\exp(Ney)$ , exploitant le fait que cette mesure soit une  $\log$  de probabilité. Les deux nouvelles mesures obtenues varient dans  $[0,1]$ .

$$LLSn = \frac{LLS - LLS_{min}}{LLS_{max} - LLS_{min}},$$

$LLS_{min}, LLS_{max}$  sont respectivement le minimum et le maximum absolus des mesures LLS des termes du fichier traité (un fichier d'ESETR2).

Ainsi, nous avons pu combiner linéairement les deux mesures et obtenir une nouvelle mesure  $M$ .

$$M = \alpha LLSn + (1 - \alpha) \exp(Ney)$$

$\alpha$  définit le taux de la contribution de la mesure sémantique et celle de Ney dans la combinaison  $M$ . Nous avons tracé les courbes PR en variant à chaque fois  $\alpha$  (FIG 4.4, FIG 4.5). Dans la FIG 4.4, les PR des différentes combinaisons ne dépassent pas la PR de Ney dans la partie de bas rappels. Néanmoins, nous avons pu observer un léger dépassement dans les grands rappels (FIG 4.5) en valorisant plus la mesure sémantique (pour des valeurs 0.7 et 0.5 d' $\alpha$ ). Ce dépassement est contraint par des pertes au niveau des faibles rappels. Le choix optimal serait de conserver une valeur d' $\alpha$  égale à 0.5. C'est avec cette valeur que nous pouvons observer un dépassement au niveau des hauts rappels tout en perdant moins au niveau des bas rappels.

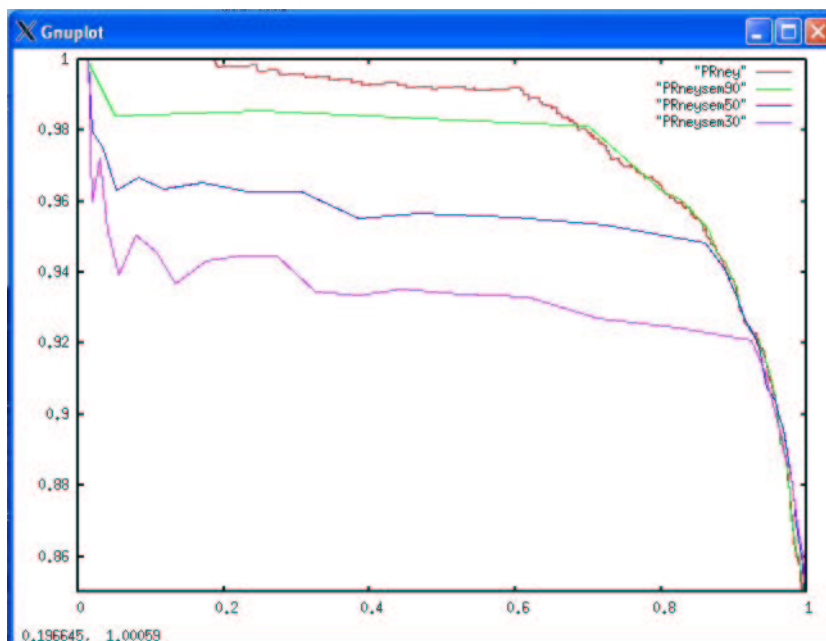


FIG. 4.4 – Courbes PR des combinaisons linéaires des mesures de Ney et LLS paramétrées par le taux de contribution sémantique  $\alpha$

Dans les bas rappels, Nous n'avons pas obtenu un dépassement similaire à celui observé dans les résultats de Cox (FIG 2.2) en combinant la mesure N-best avec sa mesure sémantique. Il est à signaler que cette dernière était meilleure que la mesure N-best dans les bas rappels alors que notre mesure sémantique est moins performante que la mesure de Ney dans cette même région. La différence des performances peut ainsi être traduite par la différence des comportements des deux mesures choisies au départ (N-best vs PR Ney). Dans la région des hauts rappels nos résultats sont en adéquation avec ceux de Cox.

## 4.4 Exploitation des *N-best*

### 4.4.1 Présentation de l'expérience

Dans ce paragraphe, nous présentons une autre approche que nous avons envisagée dans l'objectif d'introduire l'information sémantique dans la reconnaissance :

A la sortie du système de la reconnaissance, pour chaque phrase reconnue il nous a été possible de récupérer les  $N$  meilleures transcriptions (*N-best*)[13]. Nous avons cherché à combiner le score acoustique pour chaque transcription obtenue avec un score sémantique afin d'améliorer le taux de reconnaissance. Cette combinaison établit un nouvel ordre pour les *N-best* et apporte en conséquent l'amélioration souhaitée par l'introduction de la sémantique dans la reconnaissance (Sachant que pour chaque transcription, le taux de reconnaissance est évalué à partir de la meilleure combinaison obtenue parmi les *N-best*). Afin d'obtenir un score sémantique caractérisant chaque alignement, nous avons procédé ainsi :

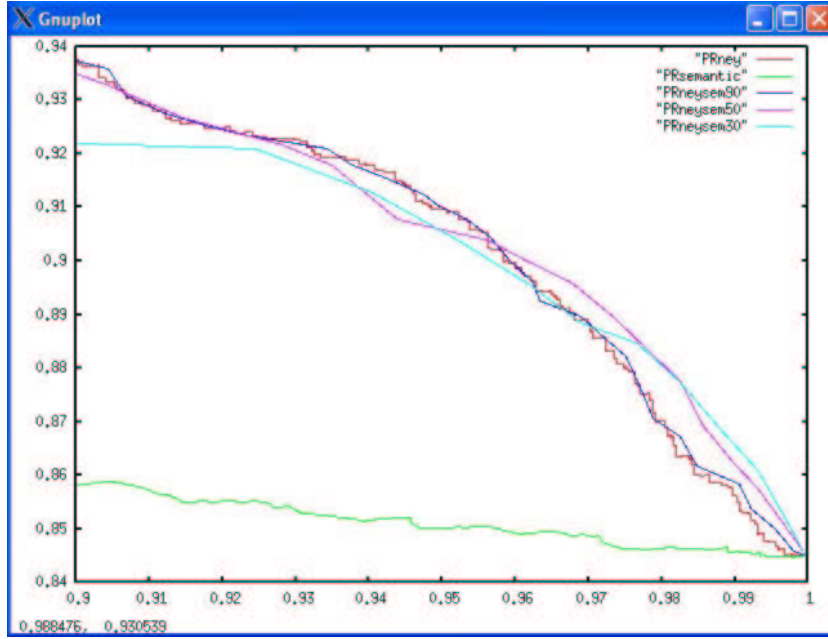


FIG. 4.5 – Courbes PR des combinaisons linéaires des mesures de Ney et LLS paramétrées par le taux de la contribution sémantique  $\alpha$  (Partie hauts rappels).

Soit  $W$  l'ensemble des labels de termes présents dans l'une des transcriptions. Pour chaque  $w_i \in W$  nous avons calculé un cosinus moyen  $\text{cosmoy}(w_i)$ .

$$\text{cosmoy}(w_i) = \sum_{w_j \in W, i \neq j} \frac{w_i \cdot w_j}{\|w_i\| \|w_j\|}.$$

Ensuite, comme dans la partie 3.5, nous avons pu obtenir la vraisemblance de chaque terme avec l'erreur  $E$   $P_{\text{erreur}}(w_i/E)$  et avec les termes justes  $P_{\text{cor}}(w_i/C)$ . Finalement, nous avons défini pour chaque terme  $i$  une probabilité  $P_{\text{semi}}$ ,

$$P_{\text{semi}} = \frac{P_{\text{cor}}(w_i/C)P(C)}{P_{\text{cor}}(w_i/C)P(C) + P_{\text{cor}}(w_i/E)P(E)}$$

avec  $P(C)$  (respectivement  $P(E)$ ) la probabilité que le mot soit correctement reconnu (respectivement mal reconnu). Ces probas sont estimées par le pourcentage de chacune des deux catégories de mots dans le fichier d'apprentissage.

En conservant une hypothèse d'indépendance entre les  $w_i$ , nous obtenons  $P_{\text{sem}}$  caractérisant toute la phrase  $P_{\text{sem}}$

$$P_{\text{sem}} = (\prod_i P_{\text{semi}})^\gamma$$

$\gamma$  est un poids qui pondère l'influence de la contribution sémantique par rapport à la vraisemblance acoustique et que nous chercherons à optimiser par la suite. Nous avons pu définir ainsi un score sémantique qui est une *Log* vraisemblance,

$$\log P_{sem} = \gamma \sum_i \log (P_{semi}).$$

Les *N-best* ont été déjà ordonné selon un score acoustique défini comme suit

$$\log Lac = \sum_i \log (P(O w_i)P(w_i)^\lambda),$$

avec  $O$  l'observation acoustique et  $\lambda$  le poids du modèle de langage.

En combinant les deux scores (sémantique et acoustique), nous obtenons un autre score : la Log vraisemblance globale LLG,

$$LLG = \log(P_{sem}) + \log(Lac)$$

Les *N-best* sont ensuite réordonnés à partir de cette nouvelle mesure.

#### 4.4.2 Résultats

En variant nos paramètres  $N$  (le nombre des *N-best* conservés) et  $\gamma$  (le poids sémantique), nous avons cherché à améliorer les résultats de la reconnaissance. Notre indicateur de performance est le taux de reconnaissance,

$$\%ACC = \frac{H-I}{T} \times 100\%,$$

$T$  est le nombre de termes à reconnaître,  $H$  est le nombre de termes correctement reconnus et  $I$  est le nombre d'insertions [13].

Dans le tableau 4.1, nous présentons les taux de reconnaissances obtenus en variant  $N$  et  $\gamma$  d'un fichier qui initialement à la sortie de reconnaissance présentait un taux de reconnaissance de 77.85%. Les cases vides dans le tableau coïncident avec des configurations que nous n'avons pas testées

| $N$        | <b>40</b> | <b>20</b> | <b>10</b> | <b>5</b> | <b>3</b> |
|------------|-----------|-----------|-----------|----------|----------|
| $\gamma$   |           |           |           |          |          |
| <b>10</b>  | 76.85     |           |           |          |          |
| <b>0.5</b> | 76.91     | 77.18     | 77.26     | 77.46    | 77.57    |
| <b>0.2</b> | 76.97     |           | 77.30     | 77.49    |          |
| <b>0.1</b> |           |           |           | 77.49    | 77.63    |

TAB. 4.1 – taux de reconnaissance d'un fichier de la sortie de reconnaissance obtenus en variant le poids sémantique *gamma* et le nombre de *N-best*  $N$

Au départ, les améliorations que nous pouvions espérer en intervertissant l'ordre des *N-best* étaient de l'ordre de 2%. Ce taux a été calculé en comparant, pour chaque phrase reconnue, les taux de reconnaissance de ses *N-best* et voir si l'ordre dans lequel ils sont

déjà disposés est en accord avec celui des taux de reconnaissance.

Nous remarquons une augmentation des valeurs des taux de reconnaissance dans le sens décroissant des valeurs de  $\gamma$  et des *N-best*. La croissance au niveau des taux tend vers la valeur de référence (77.85%) sans la dépasser et cela en minimisant la contribution sémantique ( petites valeurs de  $\gamma$ ) et en réduisant *N*.

## Conclusion

Nous avons essayé d'exploiter l'information sémantique en l'introduisant en premier lieu sous la forme d'une mesure de confiance. Combinée avec la mesure de Ney, la mesure sémantique a apporté une petite amélioration au PR au niveau des hauts rappels. En deuxième lieu, nous avons cherché à utiliser l'information sémantique sous forme d'un score combiné avec un score déjà exploité dans la reconnaissance pour le réordonnement des *N-best*. En changeant l'ordre de ces derniers suivant la combinaison réalisée des deux scores, nous n'avons pas pu obtenir une amélioration dans les taux de reconnaissance .

# Conclusion

Nous avons essayé de décrire l'environnement et le travail effectué durant ce projet de fin d'étude tout au long des chapitres de ce rapport.

Nous avons présenté dans le premier chapitre l'équipe parole et ses principaux travaux de recherches dans les domaines de l'analyse et la reconnaissance automatique de la parole. Ensuite, il a été question de faire l'état de l'art de certaines techniques de l'extraction de l'information sémantique et son utilisation dans le domaine du traitement de la parole. Les travaux de Cox et de Inkpen ont été résumé dans ce volet. Puis, dans le troisième chapitre, nous avons exposé notre démarche pour l'exploitation de la technique du Random Indexing dans l'extraction de l'information sémantique à partir du corpus *Le Monde*. Ainsi, nous avons pu définir et tester un score sémantique exploitable dans la reconnaissance. Enfin, le quatrième chapitre a traité la partie intégration du score sémantique dans le système de reconnaissance ANTS de l'équipe parole sous la forme d'une mesure de confiance combinée avec la mesure de Ney et sous la forme d'un score exploité pour rescorer les N-best à la sortie de reconnaissance.

L'approche sémantique utilisée n'a pas donné des résultats trop concluants. L'utilisation du voisinage sémantique comme mesure de confiance est une information maigre à elle seule. Il est envisageable d'introduire un aspect syntaxique à cette information surtout que la syntaxe fait partie des erreurs les plus fréquents dans la reconnaissance et que par la suppression des mots outils dans notre approche nous avons négligé cet aspect.

# Bibliographie

- [1] A. Brun, C. Cerisara, D. Fohr, I. Illina, O. Mella et K. Smaïli, “ANTS : le système de transcription automatique du LORIA.,” tech. rep., Equipe Parole, Loria, France, (2006).
- [2] S. Cox, “High-Level Approaches to Confidence Estimation in Speech Recognition.,” *IEEE TRANSACTIONS ON SPEECH AND AUDIO PROCESSING*, VOL. 10, NO. 7, (2007).
- [3] D. Inkpen et A. Désilets, “Semantic Similarity for Detecting Recognition Errors in Automatic Speech Transcripts.,” *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, (2005).
- [4] J.-P. Haton, C. Cerisara, D. Fohr, Y. Laprie et K. Smaïli, *Reconnaissance automatique de la parole Du signal à son interprétation*. DUNOD, (2006).
- [5] E. Parole, “Analysis, Perception and Speech Recognition : Activity Report.,” tech. rep., INRIA, (2006).
- [6] Z. Erlangung, “Semantic Similarity in Automatic Speech Recognition for Meetings.,” Master’s thesis, Université technique de Graz, (2007).
- [7] M. Hassel, “Word Sense Disambiguation Using Co-Occurrence Statistics on Random Labels.,” tech. rep., Royal Institute of Technology, Stockholm Sweden, (2007).
- [8] J. Gorman et J. R. Curran, “Random Indexing using Statistical Weight Functions.,” tech. rep., School of Information Technologies University of Sydney, NSW, (2006).
- [9] M. Hassel, *JavaSDM*, [www.nada.kth.se/~xmartin/java/JavaSDM/](http://www.nada.kth.se/~xmartin/java/JavaSDM/).
- [10] J.-L. FOULLEY, C. DELMAS et C. ROBERT-GRANIE, *Maximum likelihood method in linear mixed model*. Société française de statistique, Paris, FRANCE (Revue), (1998).
- [11] G. B. Durrant et C. Skinner, “Utilisation de méthodes de traitement des données manquantes pour corriger l’erreur de mesure dans une fonction de distribution,” *Statistique Canada*, (2006).
- [12] C. Cerisara, “Exploiting confidence measures for missing data speech recognition.,” *ACOUSTICS*, (2008).
- [13] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev et P. Woodland, *The HTK Book 3.4*. Cambridge University Engineering Department, (2006).

## Résumé

La LSA et la Random Indexing sont deux techniques fréquemment utilisées dans les approches sémantiques du traitement de la parole. Nous avons appliqué le Random Indexing sur le corpus du *Monde*. L'information sémantique récupérée a donné un score que nous avons combiné en premier lieu avec un score acoustique (mesure de Ney) pour obtenir une mesure de confiance exploitable par le système de reconnaissance *ANTS* de l'équipe parole du Loria. En deuxième lieu, nous avons cherché à utiliser ce même score sémantique pour rescorer et réordonner les N-best dans le but d'améliorer le taux de reconnaissance du système. Ces expérimentations ont été effectuées sur le corpus radiophonique *ESTER2*.

**Mots-clés:** LSA, Random Indexing, Sémantique, *Le Monde*, mesure de confiance, *ANTS*, rescorer, N-best, *ESTER2*.

## Abstract

The LSA and the Random Indexing are two frequently used technics in semantic approaches for speech treatment. We experimented the Random Indexing with *Le Monde* corpus. In the beginning, gathered semantic information gave a semantic score that was combined with acoustic one (Ney's measure) in order to obtain a confidence measure for *ANTS*, speech recognition system of Loria speech's team. Then, we used the same semantic score to rescore and reorder the N-best aiming to increase the accuracy of the system. These experiments have been developed on broadcast corpus *ESTER2*.

**Keywords:** LSA, Random Indexing, semantic, *Le Monde*, confidence measure, *ANTS*, rescore, N-best, *ESTER2*.