Identification nommée du locuteur : exploitation conjointe du signal sonore et de sa transcription

Vincent Jousse

ACADÉMIE DE NANTES
UNIVERSITÉ DU MAINE

THÈSE

présentée à l'Université du Maine
pour obtenir le diplôme de DOCTORAT

SPÉCIALITÉ : Informatique

École Doctorale 503
«Sciences et Technologies de l'Information et Mathématiques»
Laboratoire d'Informatique

Identification nommée du locuteur : exploitation conjointe du signal sonore et de sa transcription

par
Vincent JOUSSE

Soutenue publiquement le 04 mai 2011 devant un jury composé de :

Frédéric Béchet	Professeur, LIF, U. de la Méditerranée	Rapporteur
Frédéric Bimbot	Directeur de Recherche, IRISA, U. de Rennes	Rapporteur
Claude Barras	Maître de Conférences, LIMSI, U. Paris XI	Examinateur
Béatrice Daille	Professeur, LINA, U. de Nantes	Directrice de thèse
Sylvain Meignier	Maître de Conférences, LIUM, U. du Maine	Co-Encadrant de thèse
Christine Jacquin	Maître de Conférences, LINA, U. de Nantes	Co-Encadrante de thèse
Simon Petitrenaud	Maître de Conférences, LIUM, U. du Maine	Invité

Laboratoire d'Informatique de l'Université du Maine
Laboratoire d'Informatique de Nantes Atlantique

Table des matières

Remerciements

Cette thèse a été par bien des aspects atypique, et cette section « Remerciements » ne dérogera pas à la règle. Je remercie donc en premier lieu Jessica, ma femme. Sans son soutien et sa patience cette thèse n'aurait jamais pu débuter, c'est grâce elle que vous pouvez lire ces lignes. Merci pour tout, merci pour notre belle famille.

Je tiens ensuite à remercier les membres de mon jury pour leur participation à la soutenance de cette thèse : Monsieur Claude Barras pour avoir présidé le jury, Monsieur Frédéric Béchet et Monsieur Frédéric Bimbot, rapporteurs de ce travail, pour avoir consacré du temps à la lecture de ce document.

Ensuite je tiens chaleureusement à remercier mon équipe encadrante Nantaise. Tout d'abord Madame Béatrice Daille, directrice de cette thèse. Elle a su me montrer la rigueur que réclamait un travail de thèse et su m'orienter quand il le fallait. Ce travail « invisible » n'en est pas moins important et a eu beaucoup de valeur à mes yeux. Elle a d'autant plus du composer avec un doctorant père de famille, ce qui n'a, je le consens, pas du toujours être facile. Ensuite j'aimerais remercier Madame Christine Jacquin, co-encadrante de cette thèse. J'ai connu Christine lorsque j'étais sur les bancs de l'IUT et pouvoir retravailler en sa compagnie a été un réel plaisir.

Le Mans, une ville pleine de rencontres enrichissantes. Ce fut tout d'abord le cas avec Monsieur Sylvain Meignier, co-encadrant de cette thèse. Sylvain était présent à mes côtés quotidiennement, et sa bonne humeur, son enthousiasme sont inégalables. Si vous voulez savoir ce que collaborer signifie, prenez le temps de connaître Sylvain, il y a des rencontres que l'on n'oublie pas. Ensuite, travailler au Mans a été pour moi l'occasion de rencontrer Monsieur Simon Petitrenaud, un Iron Man des mathématiques. C'est grâce à lui qu'une partie du travail de cette thèse a pu voir le jour, et je lui en suis pleinement reconnaissant.

Ces remerciements ne seraient pas complets si je ne remerciais pas l'équipe du LIUM. Tout d'abord Yannick, un collègue formidable, un guitariste talentueux et maintenant un ami sincère. Ensuite Paul, qui, si l'on ne se fie pas aux

premières impressions, est quelqu'un de très agréable et toujours prêt à rendre service. Pour finir un grand merci à Martine, pour sa bonne humeur certes, mais aussi pour le travail qu'elle accomplit au jour le jour pour faire fonctionner le laboratoire.

Il paraît que l'on garde le meilleur pour la fin. C'est donc ici que je remercierai les acolytes ayant partagé le même bureau que moi : Antoine, Richard, Thierry et Fethi qui nous a rejoint un peu plus tard. Il y a des moments de vie qui restent gravés à jamais dans vos mémoires, ceux que l'on a partagé en font partie. Merci.

Résumé

Le traitement automatique de la parole est un domaine qui englobe un grand nombre de travaux : de la reconnaissance automatique du locuteur à la détection des entités nommées en passant par la transcription en mots du signal audio. Les techniques de traitement automatique de la parole permettent d'extraire nombre d'informations des documents audio (réunions, émissions, etc.) comme la transcription, certaines annotations (le type d'émission, les lieux cités, etc.) ou encore des informations relatives aux locuteurs (changement de locuteur, genre du locuteur). Toutes ces informations peuvent être exploitées par des techniques d'indexation automatique qui vont permettre d'indexer de grandes collections de documents.

Les travaux présentés dans cette thèse s'intéressent à l'indexation automatique de locuteurs dans des documents audio en français. Plus précisément nous cherchons à identifier les différentes interventions d'un locuteur ainsi qu'à les nommer par leur prénom et leur nom. Ce processus est connu sous le nom d'identification nommée du locuteur (INL). La particularité de ces travaux réside dans l'utilisation conjointe du signal audio et de sa transcription en mots pour nommer les locuteurs d'un document. Le prénom et le nom de chacun des locuteurs est extrait du document lui même (de sa transcription enrichie plus exactement), avant d'être affecté à un des locuteurs du document.

Nous commençons par rappeler le contexte et les précédents travaux réalisés sur l'INL avant de présenter *Milesin*, le système développé lors de cette thèse. L'apport de ces travaux réside tout d'abord dans l'utilisation d'un détecteur automatique d'entités nommées (LIA_NE) pour extraire les couples prénom / nom de la transcription. Ensuite, ils s'appuient sur la théorie des fonctions de croyance pour réaliser l'affectation aux locuteurs du document et prennent ainsi en compte les différents conflits qui peuvent apparaître. Pour finir, un algorithme optimal d'affectation est proposé. Ce système obtient un taux d'erreur compris entre 12 et 20 % sur des transcriptions de référence (réalisées manuellement) en fonction du corpus utilisé. Nous présentons ensuite les avancées réalisées et les limites mises en avant par ces travaux. Nous proposons notamment une première étude de l'impact de l'utilisation de transcriptions entièrement automatiques sur *Milesin*.

Mots-clés : Identification nommée du locuteur, reconnaissance du locuteur, transcription enrichie.

Abstract

The automatic processing of speech is an area that encompasses a large number of works : speaker recognition, named entities detection or transcription of the audio signal into words. Automatic speech processing techniques can extract number of information from audio documents (meetings, shows, etc..) such as transcription, some annotations (the type of show, the places listed, etc..) or even information concerning speakers (speaker change, gender of speaker). All this information can be exploited by automatic indexing techniques which will allow indexing of large document collections.

The work presented in this thesis are interested in the automatic indexing of speakers in french audio documents. Specifically we try to identify the various contributions of a speaker and nominate them by their first and last name. This process is known as named identification of the speaker. The particularity of this work lies in the joint use of audio and its transcript to name the speakers of a document. The first and last name of each speaker is extracted from the document itself (from its rich transcription more accurately), before being assigned to one of the speakers of the document.

We begin by describing the context and previous work on the speaker named identification process before submitting *Milesin*, the system developed during this thesis. The contribution of this work lies firstly in the use of an automatic detector of named entities (LIA_NE) to extract the first name / last name of the transcript. Afterwards, they rely on the theory of belief functions to perform the assignment to the speakers of the document and thus take into account the various conflicts that may arise. Finally, an optimal assignment algorithm is proposed.

This system gives an error rate of between 12 and 20 % on reference transcripts (done manually) based on the corpus used. We then present the advances and limitations highlighted by this work. We propose an initial study of the impact of the use of fully automatic transcriptions on *Milesin*.

Keywords : Speaker named identification, speaker recognition, rich transcription.

Chapitre 1

Introduction

Depuis plus de 15 ans nous assistons à l'explosion du nombre de documents numériques accessibles au public à travers des médias multiples comme les réseaux téléphoniques, le câble, le satellite et surtout le *Web*. Ces documents prennent une place croissante dans la vie quotidienne et sont devenus une ressource essentielle.

Les documents disponibles sur le *Web* sont passés en quelques années de la simple page *HTML* ne contenant que du texte, à des pages regroupant divers médias : l'image, le son ou encore la vidéo. *Youtube* ou les *podcast* qui permettent d'écouter les émissions radios en différé sont des exemples de la diversité croissante des documents mis à notre disposition. Sans technique de classification efficace et sans moyen d'accès intuitif, l'information recherchée reste enfouie dans une masse d'information parasite. Seuls des moyens automatiques ou faiblement supervisés d'indexation et de recherche permettent de satisfaire nos besoins.

Dans ces travaux, nous nous intéressons à l'extraction automatique d'informations contenues dans des documents audio, et plus particulièrement aux informations relatives aux locuteurs. Nous cherchons à nommer chaque locuteur d'un document par son prénom et son nom. Ces travaux sont appelés identification nommée du locuteur (INL). L'INL doit permettre de répondre à cette question : **qui a parlé et quand ?**

Certains systèmes d'INL exploitent des connaissances a priori sur les locuteurs à identifier : leur utilisation sur de grandes collections de données est difficile. En effet, obtenir un grand nombre d'informations a priori sur les locuteurs (comme des enregistrements de leur voix) est coûteux en temps. De plus, ces informations peuvent ne pas être disponibles. Un autre type d'approche, ne nécessitant aucune connaissance a priori sur les documents à traiter, a été mis en œuvre pour la première fois en 2005 dans (Canseco-Rodriguez et al., 2005). Cette

approche repose sur l'exploitation de la transcription enrichie d'un document audio pour réaliser l'identification nommée des locuteurs.

L'hypothèse fondatrice de ces systèmes est que les locuteurs sont annoncés par leur prénom et leur nom et qu'il est possible d'exploiter cette information pour nommer les locuteurs du document. Il convient alors de les détecter de manière efficace dans la transcription. Un des premiers objectifs des travaux présentés dans cette thèse, est de disposer d'un système entièrement automatique pouvant être utilisé sans étiquetage manuel, afin de nommer les locuteurs dans de grandes collections de données de manière automatique. Plus précisément, la phase de détection des prénoms et des noms dans la transcription (ce couple prénom /nom sera appelé nom complet par la suite) doit être réalisée de manière automatique.

Ensuite, il faut pouvoir utiliser ces noms complets détectés dans la transcription pour nommer les locuteurs. Toutes les techniques d'INL basées sur la transcription utilisent pour ce faire la même méthode : attribuer le nom complet détecté dans la transcription au locuteur qui vient de parler (le locuteur précédent est remercié pour son intervention par exemple), à celui qui est en train de parler (le locuteur courant s'annonce) ou a celui qui va parler (le locuteur suivant est annoncé). L'un des objectifs de cette thèse est de proposer un cadre permettant d'exploiter de manière efficace les informations fournies par les systèmes statistiques utilisés pendant cette phase d'attribution.

Cette thèse a été menée dans le cadre du projet régional *MILES*. Elle est le fruit d'une collaboration entre le Laboratoire d'Informatique du Maine (LIUM) et le Laboratoire d'Informatique de Nantes Atlantique (LINA). Les travaux menés ont permis la mise en place d'un système d'INL entièrement automatique : *Milesin*. *Milesin* a été développé et testé sur des données en français, mais l'approche est applicable à d'autres langues.

Structure du document

Dans un premier temps, nous présentons dans ce manuscrit les principales composantes du traitement automatique de la parole. Ces composantes permettent de réaliser des transcriptions dites « enrichies ». Ces transcriptions contiennent, en plus de la transcription en mots, des informations supplémentaires sur les locuteurs du document ou les entités nommées. Elles seront ensuite exploitées par les systèmes d'INL pour nommer les différents locuteurs d'un document audio.

Nous étudions ensuite les différents travaux menés sur l'identification nommée du locuteur depuis 2005 (Canseco-Rodriguez et al., 2005). Tout d'abord

en utilisant une approche symbolique puis ensuite à l'aide de différentes approches statistiques, ces travaux, et plus particulièrement ceux réalisés au LIUM (Mauclair et al., 2006), sont à la base du système d'INL développé au cours de cette thèse.

Nous continuons en présentant *Milesin* un système d'INL pouvant fonctionner de manière entièrement automatique. Ce système s'appuie sur un détecteur automatique d'entités nommées (Bechet et Charton, 2010) et sur une théorie mathématique (la théorie des fonctions de croyance) pour nommer les locuteurs d'un document.

Pour finir, nous discutons des avancées et des limites de *Milesin*. Les différentes étapes qui ont mené au développement de *Milesin* sont présentées ainsi que le contexte d'utilisation des systèmes d'INL. Une étude préliminiaire des performances du système sur des transcriptions automatiques est ensuite conduite.

Chapitre 2

Traitement automatique de la parole

Sommaire

La reconnaissance automatique de la parole (RAP) a pour but de transcrire sous forme textuelle un signal audio. La RAP est un problème complexe, seuls des sous-problèmes ont été résolus à ce jour. Un système idéal serait capable de transcrire n'importe quel locuteur (qu'il le connaisse ou non), dans n'importe quel environnement sonore (en studio, au téléphone, dans la rue, etc.) en utilisant un registre de langage et une locution spontanée. Malheureusement, un tel système n'existe pas et des contraintes doivent lui être ajoutées pour fonctionner. Ces contraintes peuvent être la connaissance du locuteur pour apprendre sa voix afin de mieux la reconnaître, ou encore la restriction à un domaine particulier pour réduire la taille du vocabulaire.

Les systèmes de RAP peuvent êtres classés selon plusieurs critères (Cerf-Danon et al., 1991) :

- **le mode d'élocution** : cela peut être des syllabes ou des mots isolés, des mots connectés entre eux mais avec des pauses artificielles ou alors une parole quasi naturelle, appelée parole continue.
  À noter que pour la parole continue, la distinction est faite entre la parole dite « préparée » et la parole dite « spontanée » (Bazillon et al., 2008). La parole « préparée » est la parole qu'aura un journaliste lorsqu'il présente les informations : les répétitions, les faux départs et autres hésitations du langage parlé y sont peu présents. La parole « spontanée » concerne quant à elle la parole utilisée lors des conversations entre plusieurs personnes : les répétitions, hésitations et autres irrégularités du langage parlé y sont très présentes.
- **la taille du vocabulaire** : tous les systèmes ont besoin d'un vocabulaire plus ou moins important. La taille d'un vocabulaire peut être inférieure à 1 000 mots pour les systèmes à petit vocabulaire comme les serveurs vocaux mais peut dépasser la centaine de milliers mots pour les systèmes de dictée vocale ou de transcription de journaux d'information. En fonction de la taille de ce vocabulaire, différentes couvertures grammaticales sont utilisées. Les systèmes à petit vocabulaire utilisent des grammaires contraintes. En revanche, les systèmes à grand vocabulaire utilisent des modèles de langage (cf. section 2.2.3) qui permettent une plus grande couverture grammaticale.
- **la dépendance (ou non) vis-à-vis des locuteurs** : les serveurs vocaux doivent par exemple être indépendants de la personne qui va l'utiliser, tandis que les systèmes de dictée vocale peuvent se permettre d'être dépendants du locuteur.
- **l'environnement** protégé ou non : les serveurs vocaux doivent pouvoir fonctionner dans un environnement bruité (on ne peut connaître à l'avance quel sera l'environnement sonore du locuteur qui interrogera le serveur), en revanche les systèmes de dictée vocale fonctionneront en environne-

ment plus protégé.

Les applications de la RAP sont multiples et peuvent aller de la simple commande vocale (reconnaissance de mots isolés) à la transcription complète d'une émission de radio en passant par la transcription de réunions. Nous donnons par la suite un aperçu des quatre grands types de systèmes qui existent en reconnaissance de la parole.

2.1 Différents types de systèmes

Les progrès réalisés par les techniques de RAP ont permis aux systèmes d'évoluer et d'être de plus en plus performants. Des systèmes de commandes vocales ne reconnaissant que de simples mots, aux systèmes de transcriptions de journaux d'information, la complexité des SRAP (Système de Reconnaissance Automatique de la Parole) ne cesse d'augmenter. Nous décrivons ici les principaux systèmes mis en place au fil des années (Cerf-Danon et al., 1991).

2.1.1 Commandes vocales

Les systèmes à commandes vocales sont des systèmes que l'on trouve abondamment dans les systèmes embarqués. Ils vont permettre une interaction entre l'utilisateur et la machine grâce à des commandes vocales. Ces commandes sont de simples mots isolés que l'utilisateur doit prononcer pour interagir avec le système. Ils se caractérisent par une taille de vocabulaire réduite et une indépendance vis-à-vis du locuteur. Les cas d'utilisation sont nombreux et il serait impossible de tous les citer ; en voici un bref aperçu :

- jeux vidéo : la console portable DS de Nintendo embarque un micro pour permettre l'interaction homme/machine grâce à des mots simples.
- téléphones portables : beaucoup de téléphones récents permettent de naviguer dans l'interface grâce à la voix.
- aide aux handicapés : de tels systèmes peuvent pallier à un organe défaillant ou peuvent être utilisés dans des systèmes destinés à la rééducation d'enfants sourds.

Comme tous les systèmes de RAP, ces systèmes se basent sur un vocabulaire (aussi appelé lexique) préalablement constitué. Ce lexique va définir la liste des mots que le système va être capable de reconnaître et seulement ceux-ci. Bien qu'il soit techniquement possible d'avoir un vocabulaire assez important, il est généralement limité à une centaine de mots. En effet, les personnes se servant de ces systèmes ne sont de toute façon pas capables de mémoriser plus d'une

centaine de commandes. À noter que ces mots doivent être contrastés au niveau phonétique de manière à réduire les risques d'ambiguïté lors de la reconnaissance.

2.1.2 Systèmes de compréhension

Les systèmes de compréhension vont permettre un dialogue contraint avec une machine. L'utilisateur va devoir prononcer une suite de mots-clefs que le système est capable de reconnaître. En plus du système de reconnaissance vocale, un dispositif de compréhension des phrases doit être utilisé afin d'interpréter les phrases et de réagir en conséquence. Comme dans les systèmes à commandes vocales, le vocabulaire est restreint à quelques centaines de mots et le système est indépendant du locuteur. Afin de faciliter la gestion du dialogue par le système, les phrases acceptables se limitent à des schémas grammaticaux simplifiés. Le mode d'élocution utilisé est généralement une parole continue et spontanée.

Bien que peu naturelles, ces restrictions syntaxiques et sémantiques sont viables si l'application est bien ciblée sur un domaine particulier. Les applications choisies pour ces systèmes le sont d'ailleurs en conséquence. Elles se limitent généralement à l'interrogation d'une base de données, de standards téléphoniques automatisés pour des renseignements météo ou des réservations de places.

Ces systèmes ont connu un important essor avec le lancement de projets financés par la DARPA (*Defense Advanced Research Projects Agency*) aux États-Unis dans les années 1970 (Klatt, 1977). Plusieurs laboratoires participent aux projets DARPA comme CMU, MIT, BBN, SRI ou encore Bell Labs.

Les différents systèmes développés ainsi que les approches proposées sont passés en revue dans (Kuhn et Demori, 1998).

Depuis, de nombreux projets se sont intéressés aux systèmes de compréhension. C'est le cas de la campagne d'évaluation française MEDIA 2006 (Méthodologie d'Evaluation automatique de la compréhension hors et en contexte du DIAlogue) et des travaux qui en ont découlé au sein du projet PORT-MEDIA (Lehuen et Lemeunier, 2010; Favre et al., 2010). Un projet européen (LUNA) est actuellement en cours sur ces problématiques [1].

1. `http://www.ist-luna.eu/`

2.1.3 Systèmes de dictée automatique

Les systèmes de dictée automatique ont pour but de transcrire un texte dicté par un locuteur de la façon la plus fidèle possible. Comme pourrait le faire une assistante, le texte transcrit doit respecter les règles orthographiques et grammaticales propres à la langue concernée. La compréhension du texte à transcrire n'est pas requise, et c'est d'ailleurs ce qui amène les erreurs les plus perturbantes pour l'utilisateur. En effet, les systèmes de dictée automatique ne sont pas capables de distinguer les différents sens d'un même mot.

Ce type de système se trouve à la charnière entre l'oral et l'écrit. En effet, même si l'interaction avec le système se fait grâce à la voix, le registre de langue utilisé sera plus proche du registre écrit. Ces systèmes sont souvent utilisés pour transcrire des rapports ou des compte-rendus. Quoiqu'il en soit, la personne est consciente qu'elle s'adresse à un ordinateur et adaptera donc sa locution en conséquence. De plus, les tournures agrammaticales ou les hésitations seront beaucoup moins nombreuses que dans un dialogue spontané entre deux personnes. La complexité est donc moindre que s'il fallait retranscrire un dialogue tel quel.

En revanche, l'exercice même de la dictée sous-entend l'utilisation de plusieurs dizaines voire plusieurs centaines de milliers de formes fléchies. Les mots sont pris en contexte, et régis par une syntaxe aussi libre que la grammaire de la langue naturelle le permet. Le vocabulaire doit lui aussi être conséquent et est constitué de plusieurs milliers de mots (de 5 000 à plus de 60 000 mots). Avec ce type de système, une séance de dictée ressemblera plus à un médecin qui dicte un compte-rendu d'opération à son assistante qu'à une maîtresse d'école qui fait faire un exercice de français à ses élèves. L'utilisateur n'est donc pas comme l'enseignant, même s'il a bien en tête ce qu'il doit dire à la machine, il improvise quand même un minimum. Il doit donc pouvoir se tromper, corriger des mots ou remanier une tournure.

Pour obtenir de bonnes performances en temps réel, ces systèmes à grand vocabulaire sont fortement dépendants du locuteur. Une phase d'apprentissage est requise afin de permettre au logiciel d'apprendre des modèles spécifiques de la voix de la personne qui l'utilise. Cette phase d'apprentissage peut durer jusqu'à une heure.

Historiquement, c'est l'équipe de recherche d'IBM dirigée par F. Jelinek qui est la première à avoir développé un système grand vocabulaire (*Tangora* 5 000 mots en 1985, 20 000 en 1987) pour la dictée vocale. Par la suite, l'ensemble des grands systèmes développés se sont inspirés du système *Tangora*. Avec un taux de réussite supérieur à 95 % pour un vocabulaire de 20 000 mots, *Tangora* est un système multilingue existant notamment pour l'anglais (Averbuch et al,

1987), l'italien (Alto et al., 1987), le français (Cerf-Danon et al., 1990) ou encore l'allemand (Wothke et al., 1989).

Au milieu des années 90 sont apparues des campagnes d'évaluation ayant pour but de tester les performances des systèmes sur des corpus lus. Par exemple, les données disponibles pour la campagne américaine Hub-3 (Pallett, 1996) sont des extraits de journaux comme le *Wall Street Journal* ou le *New York Times*. Ces extraits sont lus et enregistrés dans différentes conditions de manière à constituer des corpus audio. Le langage utilisé est donc celui utilisé pour de l'écrit. De ce fait, ce type de campagne évalue les performances de systèmes s'apparentant à la dictée vocale. Les performances deviennent rapidement excellentes, le système d'IBM affiche des taux d'erreur de l'ordre de 7 % (Bahl et al., 1995). En France, la première campagne lancée par l'AUPELF-UREF consistait à transcrire des textes lus du journal *Le Monde* en utilisant le corpus BREF (Lamel et al., 1991). Comme pour la campagne Hub-3, les résultats sont très bons puisque le système du LIMSI affiche des taux d'erreurs de l'ordre de 11 % (Gauvain et al., 1994). Suite à ces succès, de nouveaux systèmes permettant de transcrire des enregistrements non préparés ont vu le jour.

2.1.4 Systèmes de transcription grand vocabulaire

Les systèmes de transcription grand vocabulaire ont pour but de transcrire des documents audio non préparés en extrayant le maximum d'informations de l'enregistrement. Le signal audio est très riche en informations susceptibles de venir compléter la transcription en mots. Ces informations peuvent être de différentes natures comme les frontières des phrases, les informations sur les locuteurs, les zones de musique, de publicité ou encore les hésitations des locuteurs.

Un système de transcription grand vocabulaire est composé d'un module qui va permettre d'obtenir la transcription en mots du signal ainsi que d'autres modules qui vont extraire les informations additionnelles disponibles dans le signal audio. Les modules permettant d'extraire ces informations sont diverses, comme les modules issus de la reconnaissance automatique du locuteur (modules de segmentation et de classification en locuteurs).

Les documents que les systèmes de transcription grand vocabulaire sont capables de traiter sont multiples. Ce sont notamment les enregistrements de journaux radiophoniques, de compte-rendus de réunions ou d'émissions de télévision. Le mode d'élocution n'est pas contraint et la parole peut être préparée (journaliste qui présente les informations) ou spontanée (interview d'un invité). Le vocabulaire utilisé est très grand, il dépasse généralement les 65 000 mots. Ce type de système est complètement indépendant du locuteur et doit

pouvoir s'adapter à l'environnement utilisé lors de l'enregistrement (environnement bruité, studio, téléphone, etc.).

Aux États-Unis, le DARPA puis le NIST ont organisé des campagnes d'évaluations des systèmes de transcription enrichie (*Rich Transcription, RT*) (Graff, 1996) à partir de 1996. En France, c'est la campagne ESTER (Gravier et al., 2004; Galliano et al., 2006) organisée par l'AFCP, la DGA et ELRA qui a permis de réaliser les premières évaluations des systèmes de reconnaissance français sur des journaux d'informations radiophoniques. Une deuxième campagne d'évaluation française nommée ESTER II (Galliano et al., 2009) a ensuite été menée de 2008 à 2009 de manière à mesurer les progrès des différents systèmes.

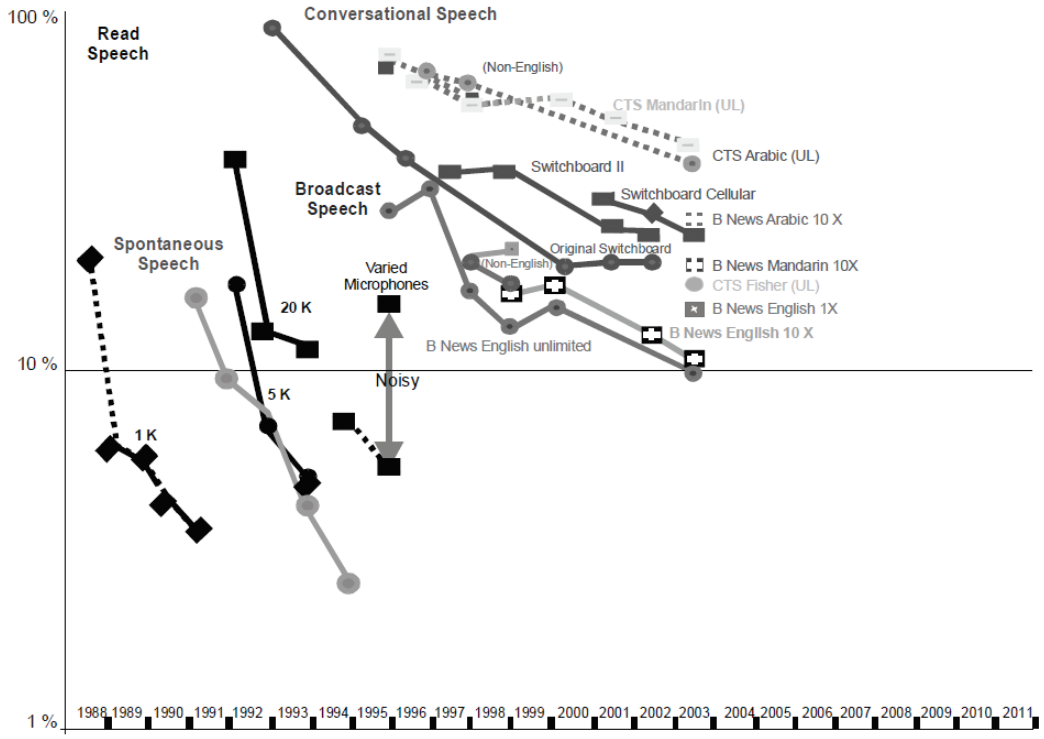

FIGURE 2.1 – *Évolution des taux d'erreur en reconnaissance de la parole*

La figure 2.1 (Pallett, 2003) présente l'évolution des performances des systèmes de reconnaissance de la parole de la fin des années 90 à 2003. Les performances des systèmes de transcription sur des émissions radiophoniques anglaises sont maintenant très bonnes (10 % de taux d'erreur). On observe aussi des taux similaires sur les campagnes d'évaluation ESTER I et ESTER II. Les défis futurs de tels systèmes portent sur l'adaptation à d'autres langues et à d'autres domaines, mais aussi sur leur capacité à *passer à l'échelle*. En effet, les données à traiter évoluent de plus en plus (multiplicité des capteurs, vidéo et son, etc.) et la taille de ces données ne cesse de croître. Un bon exemple est la

Smart Room NIST (Standford et al., 2003) qui génère environ 1 Go de données par minute, issues de multiples capteurs. Le passage à l'échelle devient donc une priorité.

2.2 Transcription automatique de la parole

Les techniques de transcription automatique de la parole sont au cœur du processus de production d'une transcription : ce sont elles qui vont permettre de fournir le texte correspondant aux paroles prononcées.

Les systèmes de transcription automatique de la parole utilisés actuellement sont des systèmes fondés sur des méthodes statistiques. Ces fondements ont été élaborés par Jelinek et ses collègues chez IBM (Jelinek, 1976), au milieu des années 1970. La figure 2.2 donne un aperçu du fonctionnement actuel d'un système de transcription automatique de la parole.

2.2.1 Principes généraux

L'objectif d'un système de RAP probabiliste est d'associer une séquence de mots $\hat{W} = w_1 w_2 ... wk$ (avec w_i qui est un mot de cette séquence) à une séquence d'observations acoustiques X. Le système recherche la séquence de mots qui maximise la probabilité *a posteriori* $P(W|X)$, où $P(W|X)$ est la probabilité d'émission de W sachant X . On obtient, après application de la règle de Bayes :

$$\hat{W} = \underset{W}{arg\,max}\ P(W|X) \ = \ \underset{W}{arg\,max}\ \frac{P(W)P(X|W)}{P(X)} \tag{2.1}$$

Comme la séquence d'observations acoustiques X est fixée, $P(X)$ peut être considérée comme une valeur constante inutile dans l'équation 2.1. On a donc :

$$\hat{W} = \ \underset{W}{arg\,max}\ P(W)P(X|W) \tag{2.2}$$

Deux types de modèles probabilistes sont utilisés pour la recherche de la séquence de mots la plus probable : des modèles acoustiques qui fournissent la valeur de $P(X|W)$, et un modèle de langage qui fournit la valeur de $P(W)$. $P(X|W)$ peut se concevoir comme la probabilité d'observer X lorsque W est prononcée, alors que $P(W)$ se réfère à la probabilité que W soit prononcée dans un langage donné. La difficulté pour obtenir un système de RAP performant

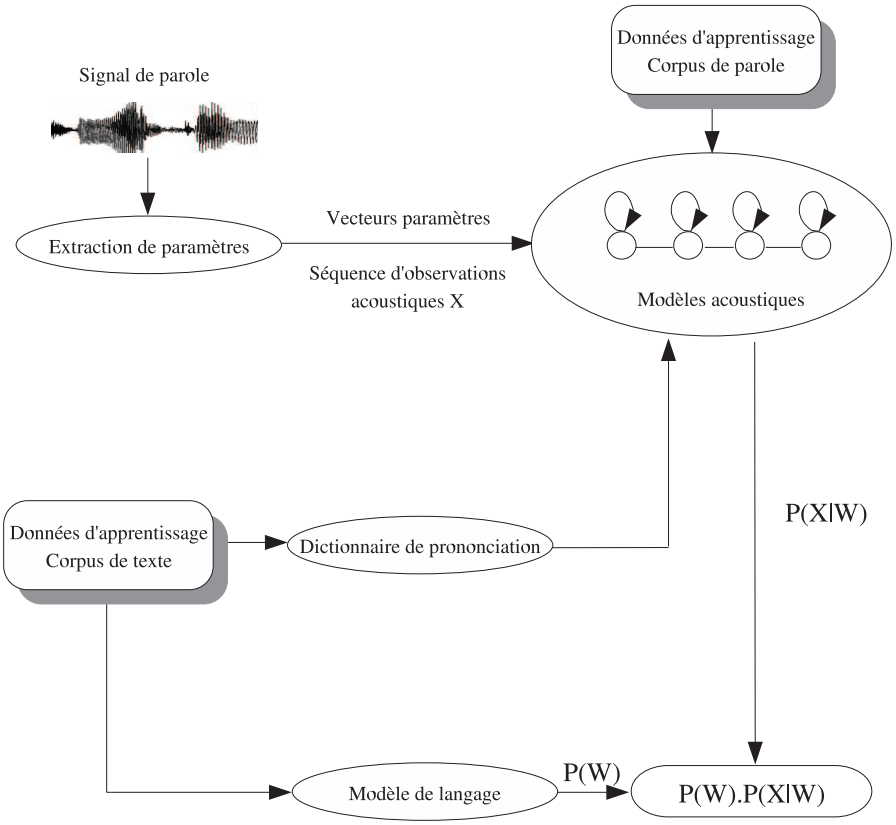

FIGURE 2.2 – *Principes généraux d'un système de reconnaissance automatique de la parole (SRAP)*

est de définir les modèles les plus pertinents possibles pour le calcul de $P(W)$ et $P(X|W)$ (voir figure 2.2).

2.2.2 Modèles acoustiques

Il est impossible pour un système de transcription d'exploiter le signal de la parole directement. En effet, il contient bon nombre d'informations qui ne sont pas utiles au décodage de la parole (mais qui le seront pour la reconnaissance du locuteur, cf 2.3.6) et qui peuvent le parasiter. Le signal contient, en plus des informations permettant de décoder les mots, des informations sur les locuteurs, sur les conditions d'enregistrement, etc. Le signal de la parole étant naturellement très variable et redondant, il nécessite de toute façon des traitements spécifiques avant de pouvoir être exploité. Ces traitements (appelés paramétrisation) vont consister à extraire du signal les paramètres qui sont dépendants

de la parole prononcée.

Plusieurs techniques de paramétrisation existent, les deux plus utilisées sont :
- PLP (*Perceptual Linear Prediction*) : domaine spectral (Hermansky et Cox, 1991)
- MFCC (*Mel Frequency Cepstral Coefficients*) : domaine cepstral (Bridle et Brown, 1974; Mermelstein, 1976)

Le signal de la parole est modélisable par un ensemble d'unités acoustiques, qui peuvent être considérées comme des sons élémentaires de la langue. Classiquement, l'unité choisie est le phonème. Le phonème est la plus petite unité discrète ou distinctive d'un mot, auquel elle peut faire perdre son sens par substitution. Par exemple, *"chou"* et *"pou"* se distinguent par leurs phonèmes initiaux. Dans le cadre des systèmes de transcription de la parole actuels, ces unités acoustiques sont modélisées par des Modèles de Markov Cachés (MMC) (Rabiner, 1989).

L'apprentissage de ces modèles acoustiques s'effectue grâce à des données *a priori* annotées. Cet apprentissage permet d'obtenir une modélisation du message de la parole. Différentes techniques d'apprentissage et d'adaptation sont utilisées, parmi les plus connues figurent notamment :

- l'apprentissage par maximum de vraisemblance (*Maximum Likelihood : ML*) avec l'algorithme EM (Dempster et al., 1977) (*Expectation/Maximisation*),
- l'estimation par information mutuelle maximale (Valtchev et al., 1997) (*Maximum Mutual Information Estimation : MMIE*),
- l'adaptation par maximum *a posteriori* (Jean-luc Gauvain et Lee, 1994) (*Maximum* a posteriori *probability : MAP*),
- l'adaptation par régression linéaire (Gales, 1998) (*Maximum Likelihood Linear Regression : MLLR*).

2.2.3 Modèles de langage

Les systèmes de transcription ont besoin de contraintes linguistiques propres à la langue à décoder. Ces contraintes vont êtres utilisées par le système pour guider le décodage dans la sélection d'hypothèses acoustiques concurrentielles. Ces contraintes sont introduites par les modèles de langage dans les systèmes de transcription. Ces modèles sont des modèles probabilistes, et la probabilité d'apparition de la séquence de k mots W_1^k s'exprime de la façon suivante :

$$P(W_1^k) = P(w_1) \prod_{i=2}^{k} P(w_i|h_i) \tag{2.3}$$

Où h_i correspond aux mots précédents w_i. h_i est aussi appelé l'historique du mot w_i. On a donc $h_i = w_1, ..., w_{i-1}$.

Les modèles utilisés en RAP sont des modèles de langage de type *n-gramme* (Jelinek, 1976). Bien qu'ancien, ce type de modèle constitue toujours l'état de l'art.

Ils correspondent à une modélisation stochastique du langage, où l'historique d'un mot est représenté par les $n-1$ mots qui le précèdent. La formule 2.3 peut donc s'écrire :

$$P(W_1^k) = P(w_1) \prod_{i=2}^{n-1} P(w_i|w_1, ..., w_{i-1}) \prod_{i=n}^{k} P(w_i|w_{i-n+1}, ..., w_{i-1}) \tag{2.4}$$

Les modèles couramment utilisés dans les SRAP sont généralement des modèles d'ordre 3 ou 4 ($n = 3$ ou $n = 4$). Nous parlons de modèle *trigramme* ou *quadrigramme* (pour $n = 1$ *unigramme*, pour $n = 2$ *bigramme*, ...). Dans le cas d'un modèle quadrigramme, l'équation 2.4 s'écrit :

$$P(W_1^k) = P(w_1)P(w_2|w_1)P(w_3|w_1, w_2) \prod_{i=4}^{k} P(w_i|w_{i-3}w_{i-2}w_{i-1}) \tag{2.5}$$

Les modèles de langage $n-grammes$ ont la particularité d'être assez souples, puisqu'ils permettent de modéliser des phrases grammaticalement incorrectes. En revanche, ils peuvent produire des phrases dénuées de tout sens. En effet, ces modèles de langage sont capables d'attribuer une probabilité à des mots inconnus. Le système est donc capable de probabiliser des phrases qu'il n'a jamais observé dans son corpus d'apprentissage, tout en privilégiant les séquences de mots les plus fréquemment observées.

2.2.4 Évaluation

La métrique utilisée dans les différentes campagnes (cf. section 2.1.4) pour évaluer les transcriptions est le taux d'erreur mot (McCowan et al., 2004), aussi appelé WER (Word Error Rate). Pour calculer ce taux d'erreur, la meilleure hypothèse de transcription du système est alignée avec les transcriptions réalisées manuellement (transcriptions dites de référence). Trois types d'erreurs sont alors possibles. Lorsqu'un mot apparaît dans l'hypothèse de reconnaissance alors qu'il n'y a aucun mot correspondant dans la référence, il s'agit d'une insertion. Si, au contraire, il y avait un mot dans la référence et que ce mot n'apparaît

pas dans l'hypothèse, nous parlons de suppression. Quand un mot de la référence est remplacé par un autre mot lors du décodage, c'est une substitution. Le taux d'erreur est ensuite déterminé par la formule 2.6.

$$WER = \frac{\text{nb. insertions} + \text{nb. suppressions} + \text{nb. substitutions}}{\text{nombre de mots de la référence}} \tag{2.6}$$

À titre d'exemple, lors de la première campagne ESTER (Galliano et al., 2005), le meilleur système (celui du LIMSI) atteignait un taux d'erreur mot de 11,9 %. L'écart entre les différents systèmes était cependant relativement important, puisque le deuxième système, celui du LIUM, affichait un taux d'erreur mot de 23,6 %. Lors de la campagne ESTER II (Galliano et al., 2009), le système le plus performant était toujours celui du LIMSI avec un taux d'erreur de 12,1 % . En revanche l'écart avec le système du LIUM s'est réduit puisque son taux d'erreur n'était plus que de 17,8 %.

2.3 Reconnaissance automatique du locuteur

La reconnaissance automatique du locuteur (RAL) regroupe un ensemble de méthodes capables d'extraire les caractéristiques vocales propres à chaque individu à partir d'un document audio. La RAL joue un rôle important dans l'enrichissement d'une transcription puisqu'elle va permettre d'apporter des informations sur les différents locuteurs du document.

2.3.1 Caractéristiques et variabilité

Toutes les techniques de RAL se basent sur l'extraction des caractéristiques (*features*) du signal audio (Atal, 1976; Doddington, 1985; Hollien, 1990; O'Shaughnessy, 1986; Sambur, 1975). Ces caractéristiques doivent être robustes par rapport aux conditions d'enregistrement et fournir le plus de renseignements possibles concernant l'identité du locuteur.

Une question importante est celle de la variabilité des caractéristiques mesurées. On parle de variabilité intra-locuteur lorsque l'on s'intéresse à la variation des paramètres mesurés pour un même locuteur et de variabilité inter-locuteur si on considère des locuteurs différents (Doddington, 1985; Hollien, 1990; Rosenberg et Soong, 1992). Pour la reconnaissance du locuteur, on cherche à extraire des caractéristiques du signal de parole qui présentent une forte variabilité inter-locuteur (pour pouvoir différencier les locuteurs entre eux) et une faible variabilité intra-locuteur (pour garantir la robustesse du système).

2.3.2 Applications

Des serveurs vocaux aux terminaux mobiles, en passant par les dispositifs de sécurité, la RAL est utilisée dans de nombreux domaines. Certains serveurs vocaux l'utilisent notamment pour détecter le genre du locuteur qui s'exprime. Par exemple, grâce à la plateforme MISTRAL, la société Calistel (Meignier et al., 2008) propose de rediriger les appels provenant de locuteurs masculins vers des opératrices féminines et inversement, afin d'obtenir un taux de satisfaction des appels plus important. Mais les applications de la RAL les plus connues sont certainement celles qui concernent la biométrie.

Dans les applications biométriques, un système de RAL va essayer de reconnaître, grâce à sa voix, un locuteur qui cherche à s'identifier auprès d'un terminal. Même si ces systèmes donnent de bons résultats, ils ne sont pas parfaits. Les systèmes à l'état de l'art comme MISTRAL ont un taux d'erreur aux alentours de 5 % (Meignier et al., 2008). Ils ne peuvent donc pas être considérés comme aussi sûr que les empreintes digitales et doivent être utilisés avec prudence à des fins judiciaires, par exemple (Boë et al., 1999; Bonastre et al., 2003). En revanche, avec l'arrivée des terminaux mobiles dans la vie de tous les jours, la RAL s'invite avec succès dans les téléphones portables de dernière génération (Larcher et al., 2010).

L'utilisation de la RAL dans les applications que nous venons de citer peut se découper en quatre grandes catégories décrites ci-dessous. Ces catégories sont d'ailleurs sujet à évaluation dans les différentes campagnes comme NIST (Alvin et Martin, 2004) et ESTER (Gravier et al., 2004; Galliano et al., 2009).

2.3.3 Identification automatique du locuteur

L'identification automatique du locuteur (IAL) a été une des premières utilisations de la RAL (Atal, 1976). En IAL, la liste des locuteurs à identifier est connue du système. Le système doit pouvoir décider, à partir d'un échantillon de voix, à quelle identité connue du système correspond l'échantillon. La figure 2.3 décrit le principe général de l'IAL.

L'identification automatique du locuteur se découpe en deux étapes : une étape d'apprentissage et une étape de test. À partir d'un ensemble d'enregistrements de voix de chaque locuteur, le système apprend un modèle pour chaque locuteur lors de l'étape d'apprentissage. Lors de l'étape de test, le système confrontera l'échantillon de voix qu'il recevra aux différents modèles qu'il aura déjà appris afin de déterminer si l'identité du locuteur est déjà connue.

En fonction de l'application, deux types de décisions sont possibles. Prenons

le cas d'un centre d'appel téléphonique qui enregistre les conversations de ses employés. Un système de IAL peut être utilisé pour classer chacun des enregistrements en fonction de l'employé concerné. Dans ce cas, la liste des locuteurs possibles est connue du système (tous les employés), et aucun autre locuteur non employé ne peut être concerné. Dans ce cas, l'application présuppose que l'ensemble des locuteurs possible est fermé et connu du système. Le système de IAL choisira alors, parmi la liste, le modèle de locuteur le plus ressemblant à l'enregistrement de test.

En revanche, dans une application utilisant un ensemble ouvert de locuteurs possibles, le système de IAL ne connait pas tous les locuteurs possibles. Dans ce cas, en plus de déterminer le locuteur le plus vraisemblable, le système a la possibilité de rejeter l'échantillon de test en ne renvoyant aucune identité connue pour cet échantillon.

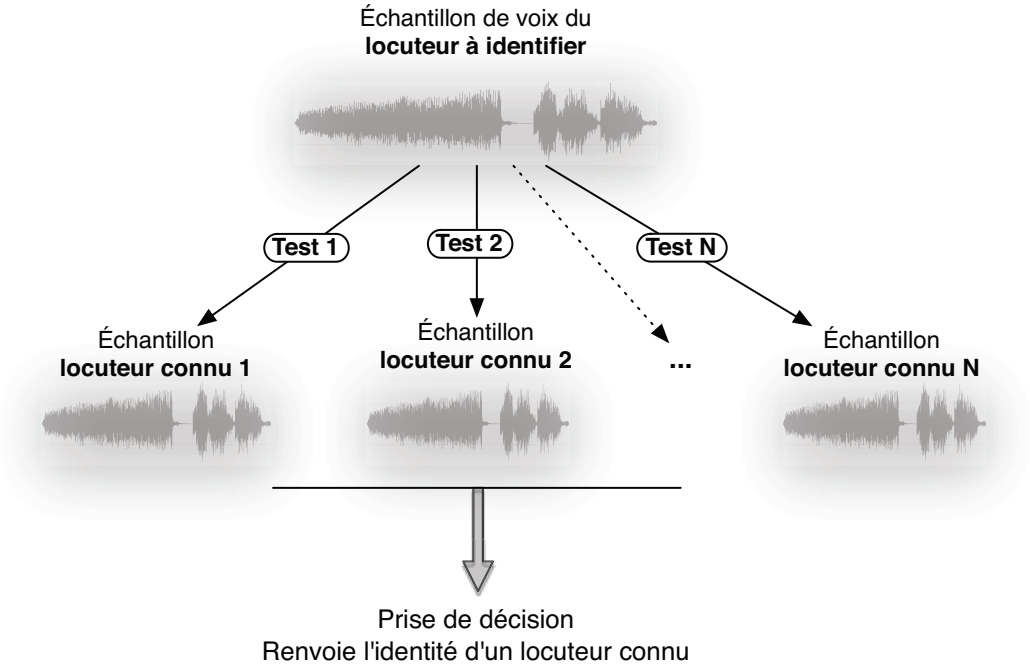

FIGURE 2.3 – *Principe de l'identification automatique du locuteur*

2.3.4 Vérification automatique du locuteur

La vérification automatique du locuteur (VAL) permet de décider si l'identité revendiquée par un locuteur est compatible avec sa voix. Il s'agit donc de trancher entre deux hypothèses : soit le locuteur est bien le locuteur autorisé (on l'appelle aussi locuteur client), c'est-à-dire que son identité correspond à celle

qu'il revendique, soit le locuteur est un imposteur qui cherche à se faire passer pour la personne qu'il n'est pas. À partir d'un échantillon de voix de référence, et d'un échantillon de voix de test, le système va donc devoir dire si oui ou non les deux locuteurs correspondent (Atal, 1976).

Les systèmes de VAL sont très dépendants des différences entre les échantillons de voix de référence et les échantillons de tests (cf. 2.3.1). Accepter un locuteur qui devrait être rejeté peut avoir de lourdes conséquences, en particulier dans les applications où un haut niveau de sécurité est demandé (Bonastre et al., 2003) (contrôle aux frontières, système bancaire, identification judiciaire, etc.).

2.3.5 Suivi de locuteur

Le suivi de locuteur (SL) se base sur la tâche de vérification du locuteur. À partir d'un enregistrement de référence d'un locuteur, le système va devoir déterminer si le locuteur intervient dans le document. Si c'est le cas, il devra être capable de préciser où et quand ce locuteur intervient dans l'enregistrement. À noter qu'à part le locuteur à suivre, aucun autre locuteur du document n'est connu du système. Plusieurs méthodes différentes décrites dans (Bonastre et al., 2000) peuvent être employées pour réaliser cette tâche, la figure 2.4 propose un aperçu général du principe du SL.

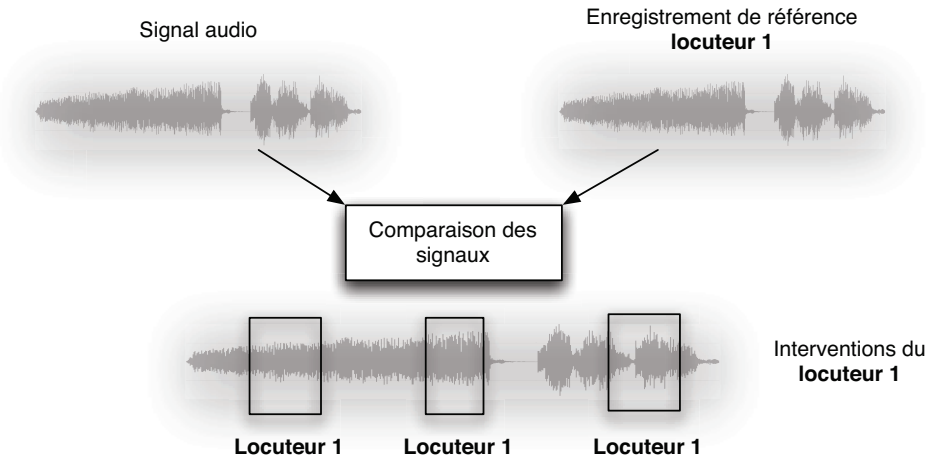

FIGURE 2.4 – *Principe du suivi de locuteur*

2.3.6 Segmentation et classification en locuteur

La tâche de segmentation et de classification en locuteur consiste à délimiter l'intervention de chaque locuteur dans le document audio. À la différence du suivi de locuteur, aucune information *a priori* sur les locuteurs du document n'est disponible pour cette tâche. Plus précisément, aucune information sur le nombre de locuteurs n'est disponible, ni leur identité, ni des modèles de voix qui auraient pu être appris au préalable.

Les travaux fondamentaux de la segmentation et de la classification en locuteur ont été réalisés au début des années 1990 par la société BBN sous la direction de H. Gish (Siu et al., 1991, 1992). Ces travaux consistent à indexer les différents échanges radio entre pilotes et contrôleurs aériens. Ces échanges sont enregistrés puis segmentés automatiquement. Ces enregistrements peuvent contenir plusieurs dialogues contrôleur/pilote. La segmentation est ensuite utilisée pour reconstruire ces dialogues.

Depuis, les techniques ont évoluées (Tranter et Reynolds, 2006) et le champ d'application de la segmentation en locuteurs s'est étendu, il se retrouve intégré dans le cadre plus vaste de l'indexation en locuteurs de bases de données de documents multimédia (Meignier, 2002). Les documents traités sont divers et variés, on peut notamment citer les conversations téléphoniques, les enregistrements de journaux télévisés ou radiophoniques, les films ou encore les enregistrements de réunion.

La figure 2.5 décrit les principes généraux du processus de segmentation et de classification qui s'effectue en plusieurs étapes :

- La **paramétrisation** va extraire les caractéristiques acoustiques du document audio. Ces caractéristiques seront représentées via des vecteurs acoustiques qui pourront ensuite être exploités par les autres phases du processus.
- La **segmentation** va s'attacher à trouver des points de ruptures, des frontières, au sein de l'enregistrement. Ces frontières peuvent être de différentes natures : changement de locuteur, changement de canal de transmission (studio, téléphone), présence d'un long silence, musique, etc. Les portions de signal entre ces différentes frontières sont appelées segments. Ces segments contiennent des données homogènes : même locuteur, même canal de transmission.
- La **classification** groupe les segments locuteur par locuteur jusqu'à découvrir le nombre de classes correspondant au nombre de locuteurs intervenant dans le document. Chaque classe contient à la fin de la classification l'ensemble des segments prononcés par un locuteur. La classe définit alors l'intervention du locuteur dans le document.

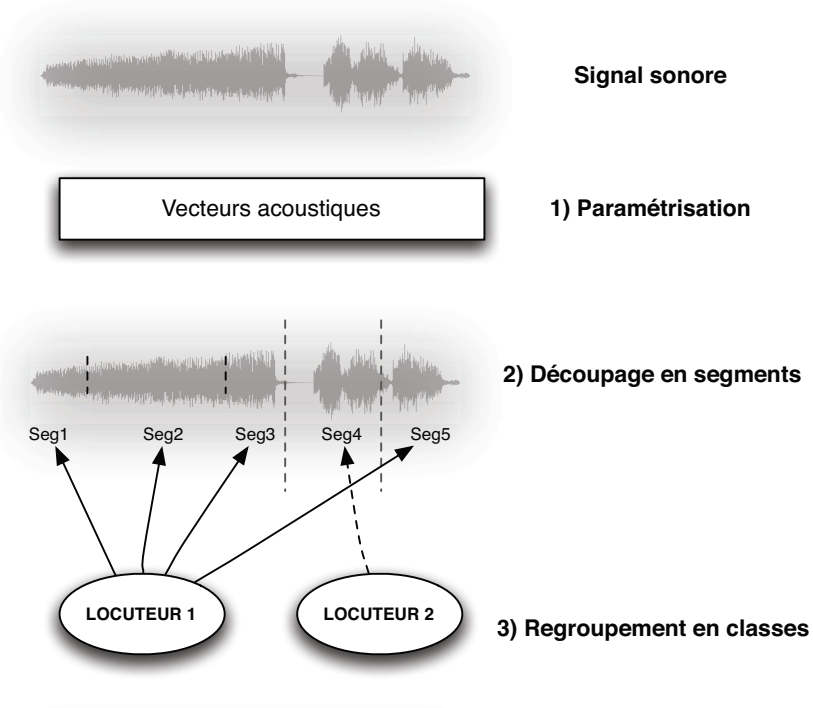

FIGURE 2.5 – *Principe de la segmentation et classification en locuteur*

À noter que les classes de locuteur produites sont anonymes. En effet, ce type de système étiquette chacune des classes avec un nom unique, mais dénué de tout sens. Le processus qui consiste à étiqueter ces classes avec le vrai couple prénom/nom du locuteur auquel elles se rapportent est appelé **identification nommée du locuteur**. C'est à cette tâche que s'intéressent les travaux de cette thèse.

2.4 Transcription enrichie pour la reconnaissance en locuteur

Avec l'évolution des systèmes au fil des années (de la simple reconnaissance de mots isolés à la transcription de journaux d'informations) est apparue la notion de transcription enrichie. En effet, comme décrit dans 2.1.4, nombre d'informations annexes à la transcription en mots peuvent être extraites du signal.

Par rapport aux systèmes de dictée vocale, les documents utilisés pour la production de transcription enrichie sont fréquemment multi-locuteurs. Au sein d'un même enregistrement, des dizaines de locuteurs différents peuvent parler

tout au long du fichier. De plus, ces locuteurs peuvent s'exprimer depuis un téléphone ou en studio. Toutes ces informations permettent d'étayer, d'enrichir la transcription. La reconnaissance du locuteur est utilisée pour segmenter le document en locuteurs ainsi que pour détecter leur genre ou l'environnement à partir duquel ils s'expriment (studio ou téléphone). La détection des entités nommées est quant à elle utilisée pour étiqueter des groupes de mots particuliers comme les noms de personnes ou encore les lieux.

La figure 2.6 décrit les différentes composantes d'une transcription enrichie. Ces composantes constituent la base de tous les systèmes de transcription actuels. Il est possible de distinguer deux grandes parties que nous détaillons ci-dessous.

2.4.1 Segmentation et classification

Cette première partie consiste à découper le fichier audio de manière à obtenir des portions homogènes, c'est à dire concernant le même locuteur et les mêmes conditions d'enregistrement. Ce premier niveau de découpage est communément appelé découpage en segments. Contrairement à ce que l'on pourrait penser, les segments ne représentent généralement pas des phrases, mais plutôt des groupes de souffle : ils commencent et s'arrêtent lorsqu'un silence significatif est détecté. Ce premier découpage est effectué pour des raisons techniques, et notamment pour faciliter la tâche des systèmes de transcription automatique (cf. 2.2), qui doivent décoder segment par segment pour éviter des problèmes d'explosion combinatoire. Cette étape utilise la méthode décrite dans la section 2.3.6 à laquelle un module de détection de zones de parole ou de non-parole (silence, musique, bruit, etc.) a été ajouté.

Tous les segments contigus appartenant au même locuteur définissent un tour de parole de ce locuteur. Un même locuteur qui parle à plusieurs endroits d'un enregistrement aura donc plusieurs tours de parole au sein de cet enregistrement. En reconnaissance du locuteur (cf. 2.3.6), on dit que tous les segments (et donc les tours de parole) d'un même locuteur sont regroupés au sein d'une même classe. À chaque locuteur du document correspond donc une classe de locuteur.

Au niveau purement acoustique, il est aussi possible de détecter le genre (masculin ou féminin) du locuteur qui s'exprime (Meignier et Merlin, 2010). Cette information peut être ajoutée à la transcription enrichie, comme c'est le cas dans la figure 2.6.

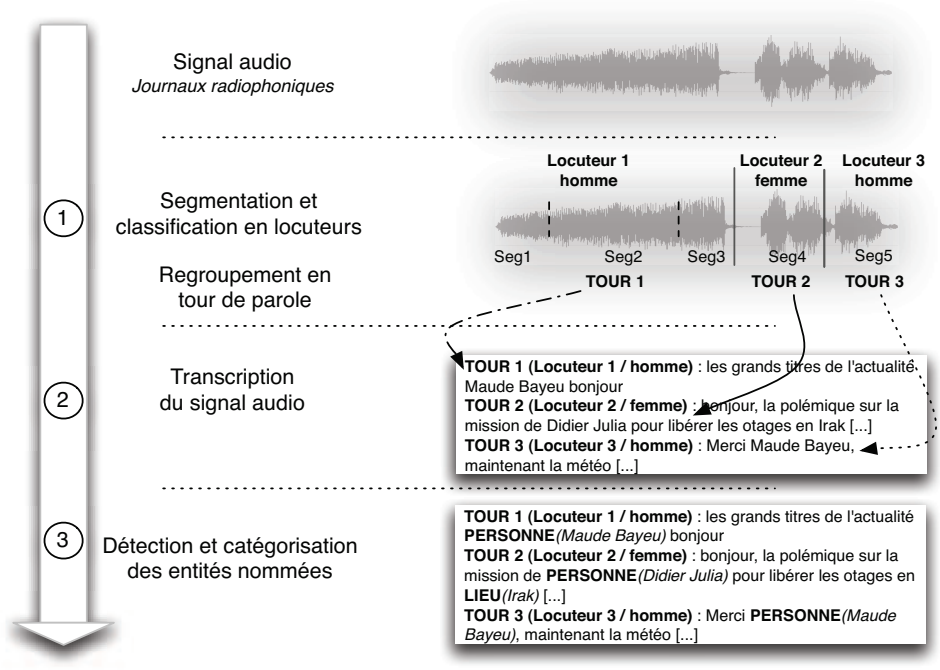

FIGURE 2.6 – *Les étapes de réalisation d'une transcription enrichie*

2.4.2 Transcription et entités nommées

La deuxième partie consiste tout simplement à écrire les paroles prononcées par les différents locuteurs afin d'obtenir la transcription en mots du signal. Lorsque cette transcription est réalisée automatiquement, la ponctuation ainsi que les majuscules ne sont pas requises. En revanche, des techniques existent pour rajouter la ponctuation et les majuscules aux transcriptions afin de se rapprocher, tant que faire se peut, d'une transcription réalisée manuellement (Beeferman et al., 1998; Huang et Zweig, 2002).

Les transcriptions, comme celles de journaux d'informations par exemple, contiennent un grand nombre d'entités nommées (noms de personnes, lieux, radios, etc). Ce sont des informations supplémentaires qui vont pouvoir être utilisées en recherche d'informations, ou plus précisément dans le cadre des présents travaux, pour identifier les locuteurs d'un document. Des systèmes permettent de détecter ces entités nommées, ils sont décrits dans la section suivante.

2.5 Détection des entités nommées

Dans le but d'enrichir une transcription, une des étapes est généralement de détecter les entités nommées. Ces entités nommées vont permettre d'extraire un certain nombre d'informations des transcriptions, comme les noms de personnes, les lieux, les noms d'organisations, etc.

2.5.1 Catégorisation

Par la richesse des informations qu'elles contiennent, les entités nommées (EN) sont des éléments très importants pour les systèmes d'extraction d'information. Dans les années 1980, les campagnes d'évaluation MUC ont permis de définir ce qu'était une tâche de reconnaissance des entités nommées. Pour MUC-6 (Grishman et Sundheim, 1996; Sundheim, 1996), les EN sont les noms propres, les acronymes et éventuellement d'autres mots qui rentrent dans les catégories suivantes :

- *Organisation* : regroupe les entreprises, les institutions gouvernementales et les autres organisations ;
- *Person* : regroupe les noms de personnes ou de familles ;
- *Location* : regroupe les noms de lieux politiquement ou géographiquement définis (villes, pays, régions, etc.) ;
- *Time* : regroupe les dates et données temporelles ;
- *Number* : regroupe les données numériques comme les sommes d'argent ou les pourcentages.

Il existe des typologies des EN beaucoup plus complètes que celle des campagnes MUC comme celle de Paik et al. (Paik et al., 1996) ou de Coates-Stephens (Coates-Stephens, 1992). Par exemple, celle de Coates-Stephens définit 8 catégories d'EN :

- les noms de personnes ;
- les noms de lieux ;
- les noms d'organisations ;
- les noms d'origines (noms d'habitants de pays, de villes, de régions, etc.) ;
- les noms de législations (*loi Évin*), d'indices boursiers (*Nikkei, CAC 40, Dow Jones*) ;
- les noms d'événements (guerres, révolutions, catastrophes, salons, JO, etc.) ;
- les noms d'objets.

Elles couvrent donc les noms propres et des formes linguistiques spécifiques.

2.5.2 Les différents types de systèmes

Selon Poibeau (Poibeau, 2002) et Sekine et Eriguchi (Sekine et Eriguchi, 2000), les systèmes de détection des EN peuvent êtres classés en trois grands types :

Les systèmes fondés sur des règles écrites « à la main »
Dans ces systèmes, le concepteur doit élaborer un ensemble de règles qui seront ensuite utilisées pour détecter les entités nommées. Historiquement, cette technique fût la première utilisée dans le domaine. Bien que depuis la campagne d'évaluation MUC-6, l'apprentissage automatique ait fait son apparition dans ce domaine, les systèmes à base de règles écrites « à la main » restent encore très utilisés aujourd'hui.

Les systèmes à base d'apprentissage automatique
Ces systèmes utilisent des techniques d'apprentissage automatique pour apprendre un modèle capable d'étiqueter des entités nommées à partir d'un corpus d'apprentissage. Ces travaux sont largement inspirés des techniques utilisées en reconnaissance automatique de la parole, et utilisent diverses techniques d'apprentissage : modèles de Markov cachés, modèle d'entropie maximum, arbres de décision, etc. Le système (LIA_NE) développé par Béchet (Bechet et Charton, 2010) au Laboratoire d'Informatique d'Avignon (LIA) pour ESTER2 (Galliano et al., 2009) en est un exemple.

Les systèmes mixtes
Ces systèmes utilisent généralement des lexiques initiaux. Parmi ces systèmes, Poibeau (Poibeau, 2002) distingue deux approches. La première consiste à apprendre automatiquement des règles, puis à utiliser un expert pour les réviser. Dans la seconde, un ensemble de règles de base est constitué par le concepteur, puis étendu (semi-)automatiquement par inférence, afin d'obtenir une meilleure couverture. Par exemple, Nemesis, développé par Fourour (Fourour, 2004) au Laboratoire d'Informatique de Nantes Atlantique (LINA), ou plus récemment les travaux présenté dans (Brun et Ehrmann, 2010) pour la campagne d'évaluation ESTER2, font partie de ces systèmes.

2.5.3 Reconnaissance et découverte des Entités Nommées

Lors de la reconnaissance des entités nommées, plusieurs types de problèmes doivent être résolus (Fourour, 2004). Pour exprimer ces problèmes, la terminologie introduite par (Daille et Morin, 2000) fondée sur des critères graphiques

est utilisée :

- **Entités nommées pures** : les entités nommées d'une seule forme commençant par une majuscule comme *France*, *FFF*, *États-Unis*.
- **Entités nommées composées** : les entités nommées constituées de plusieurs formes pleines comportant toutes une majuscule comme *Quai Branly*, *Grand Palais*.
- **Entités nommées mixtes** : les entités nommées constituées de plusieurs formes qui peuvent ou non commencer par une majuscule, mais dont au moins une est entièrement en minuscules comme *Parti socialiste*, *château de Versailles*.

Tout d'abord, la reconnaissance des entités nommées connues est réalisée grâce à la projection de bases lexicales sur le texte à étiqueter. Les systèmes sont alors confrontés au problème de l'ambiguïté des entités nommées pures : *Paris* est associé à un nom de ville mais peut apparaître au sein d'entités nommées composées ou mixtes. Il peut alors désigner une personne comme le *Comte de Paris*, un nom de rue *rue de Paris* ou encore une société *Paris International*. Même pour ces entités nommées pures, une analyse du contexte est nécessaire pour obtenir une catégorisation correcte.

À ce problème d'ambiguïté viennent s'ajouter les variations que peuvent subir les entités nommées : les variations graphiques (*Parti Socialiste*, *Parti socialiste*), les sigles, acronymes ou abréviations, (*Société Nationale des Chemins de fer Français* $\rightarrow$ *SNCF*), les métaphores (*l'Everest* $\rightarrow$ *le toit du monde*), etc.

Outre la gestion de ces différents problèmes, les systèmes doivent être capable de découvrir de nouvelles entités nommées qui ne sont pas présentes dans leurs bases lexicales.

Par exemple, les entités nommées mixtes sont par définition constituées d'une ou plusieurs formes contenant une majuscule et de noms communs. Pour ce type d'entités nommées, la difficulté réside dans la délimitation de l'entité nommée : où se trouvent les bornes gauche et droite de l'entité nommée ?

Pour la délimitation à gauche, le problème est d'inclure ou non les noms communs situés avant la forme comportant une majuscule au sein de l'entité nommée mixte. Ces noms communs peuvent préciser un rôle social (*le premier ministre François Fillon*), des statuts particuliers (*l'exilé Musil*) ou encore une catégorisation géographique.

La délimitation à droite se heurte aux problèmes de la modification, de la résolution de l'attachement prépositionnel et de la portée de la coordination (Wacholder et al., 1997). Par exemple, l'adjectif modifiant une entité nommée comme *fédéral* dans *Allemagne fédérale* constitue une entité nommée mixte, ce qui n'est pas le cas avec *lointain* dans *Chine lointaine*. En plus de cette ambiguïté

sur la limite droite des entités se pose le problème de la sur-composition : une entité nommée mixte peut contenir une entité nommée d'une autre catégorie (par exemple *Université du Mans*, *Guerre du Vietnam*).

Les systèmes adoptent différentes techniques pour résoudre ces problèmes. Ces techniques font appel à des solutions linguistiques, à des solutions à base d'apprentissage ou à un mixte des deux. Quoiqu'il en soit, la délimitation des entités nommées peut varier en fonction du type d'application auxquelles elles sont destinées. Par exemple, une application cherchant à extraire les prénoms et patronymes d'un document aura besoin d'entités nommées très courtes, ne contenant que les informations recherchées. Dans l'entité nommée mixte *le président Jacques Chirac*, seuls le prénom et le nom lui seront utiles. La recherche de la borne gauche ne devra donc pas englober le nom commun *président*.

Chapitre 3

L'identification nommée du locuteur

Sommaire

L'identification nommée du locuteur (INL) consiste à nommer les locuteurs d'un document audio en leur attribuant un prénom et un patronyme. Ces informations viennent enrichir la transcription et peuvent être utilisées dans plusieurs domaines d'applications.

3.1 Applications

Disposer des noms complets des intervenants d'un document audio peut être utilisé à des fins de recherche d'informations. En effet, les bases de données audio et vidéo ne cessent de croître avec la numérisation massive des documents audiovisuels. Ainsi, les besoins en indexation et recherche automatiques croissent de manière identique. Par exemple en France, l'Institut National de l'Audiovisuel (INA) a lancé un plan de sauvegarde et de numérisation de ses différents documents en 1999. Ce plan a pour but de numériser plusieurs millions d'heures d'enregistrements de radio et de télévision. Suite au lancement de ce plan de sauvegarde, de nombreux travaux ont été menés à l'INA sur l'indexation automatique des documents (Veneau, 2001; Allauzen, 2003; Joly, 2004; Poli, 2007). L'utilisation de l'INL pour ce type de bases de données pourrait apporter un réel plus pour l'indexation. Elle permettrait par exemple de faciliter la recherche des différents intervenants dans les documents indexés ou encore de remplir automatiquement les grilles de programmes avec le nom des intervenants.

Les récentes campagnes d'évaluation des systèmes de reconnaissance automatique de la parole (cf. 2.1.4) s'intéressent à l'enrichissement de la transcription (cf. 2.4) par diverses informations comme les entités nommées ou encore les informations sur les locuteurs du document. Cette transcription contient notamment une segmentation et une classification en locuteurs (cf. 2.3.6) qui permettent de délimiter les différentes interventions dans l'enregistrement. En revanche, les informations fournies par le processus de segmentation et de classification en locuteur ne permettent pas d'identifier les locuteurs du document par leur prénom et patronyme. Les locuteurs sont uniquement identifiés par des labels anonymes du type *Locuteur1*, *Locuteur2*. L'INL s'inscrit dans l'enrichissement de la transcription en permettant de remplacer ces labels anonymes par le prénom et le nom des locuteurs. Ce couple prénom/patronyme sera par la suite désigné comme étant le **nom complet** du locuteur.

Afin d'évaluer les différents systèmes d'INL, une même métrique est utilisée. Nous commençons par décrire cette métrique commune avant de présenter les différentes approches.

3.2 Métrique d'évaluation

L'évaluation généralement utilisée est celle proposée dans (Tranter, 2006) et consiste à comparer les noms complets des locuteurs des transcriptions de référence aux locuteurs proposés par le système d'INL. Ces comparaisons sont fondées sur les opérations élémentaires permettant de comparer deux chaînes de caractères. Plusieurs cas possibles sont étudiés lors de la comparaison entre la référence et les hypothèses fournies par le système d'INL :

- Correct (C) : le système propose une identité correspondant à celle indiquée dans la référence ;
- Substitution (S) : le système propose une identité différente de l'identité présente dans la référence ;
- Insertion (I) : le système propose une identité alors que le locuteur n'est pas identifié dans la référence ;
- Suppression (D) : le système ne propose pas d'identité alors que le locuteur est identifié dans la référence ;
- Inconnu (U) : le système ne propose pas d'identité et la référence ne contient pas d'identité.

Une mesure de Précision et de Rappel est définie à partir des 5 cas d'erreur :

$$P = \frac{C}{C+S+I} \;\; ; \;\; R = \frac{C}{C+S+D} \tag{3.1}$$

Afin de pouvoir attribuer aux locuteurs d'un document leurs noms complets, il faut obtenir leurs prénoms et patronymes. Deux principales approches sont possibles : disposer d'informations a priori sur les locuteurs comme dans le cas du suivi de locuteur (cf. 2.3.5) ou utiliser les informations fournies par le document lui même pour déterminer l'identité des locuteurs. Nous commençons par présenter la première approche avant de nous intéresser plus longuement à la deuxième approche qui constitue le cœur de cette thèse.

3.3 Utilisation de connaissances a priori

Obtenir l'identité d'un locuteur a d'abord été réalisé avec des méthodes purement acoustiques (Bimbot et al., 2004). Le suivi de locuteur (décrit dans le chapitre 2.3) peut être étendu à n locuteurs afin d'identifier tous les locuteurs d'un document. Bien qu'utilisées depuis des années, ces méthodes présentent plusieurs inconvénients.

Tout d'abord, il est nécessaire de connaître les personnes que l'on cherche à identifier. Les systèmes doivent commencer par apprendre des modèles acoustiques correspondant à chaque locuteur du document pour ensuite les associer aux différents intervenants de l'enregistrement traité. Ces systèmes ne peuvent identifier que des personnes dont ils possèdent déjà les modèles. De plus, pour apprendre ces modèles acoustiques, une quantité de données suffisante est nécessaire : plusieurs minutes sont un minimum.

Enfin, les conditions acoustiques des données d'apprentissage doivent être similaires à celles du document à traiter : des données trop éloignées dans le temps (et donc avec une voix qui peut avoir changé) ou des conditions d'enregistrement différentes (studio ou téléphone par exemple) dégraderont les performances d'identification (cf 2.3.1).

Dans l'hypothèse d'une utilisation sur des enregistrements de journaux radiophoniques par exemple, il faudrait disposer des enregistrements de tous les locuteurs (présentateurs, invités, intervenants, etc.) en quantité suffisante. S'il semble plausible de pouvoir les obtenir pour les journalistes qui sont souvent présents dans les émissions de radio, il semble beaucoup plus compliqué de les obtenir pour certains invités pas encore célèbres ou pour les inconnus interrogés par téléphone.

Ces différentes limitations ont motivé les récents travaux ne nécessitant pas de connaissances a priori sur les locuteurs pour réaliser l'identification. Ces travaux se basent sur la transcription du signal audio pour identifier les locuteurs d'un enregistrement.

3.4 Utilisation des informations de la transcription

Sans connaissance a priori, ces méthodes doivent extraire les noms complets à partir des informations disponibles dans le signal audio. La transcription en mots du signal est une source d'information de choix pour trouver les prénoms et les patronymes des locuteurs : dans beaucoup d'enregistrements (et notamment ceux provenant des journaux radiophoniques) les noms complets des intervenants font partie de la transcription. L'identification nommée consiste donc à détecter les noms complets présents dans la transcription, pour ensuite les attribuer aux différents locuteurs du document.

La figure 3.1 donne un aperçu de ce que l'identification nommée à partir de la transcription cherche à réaliser.

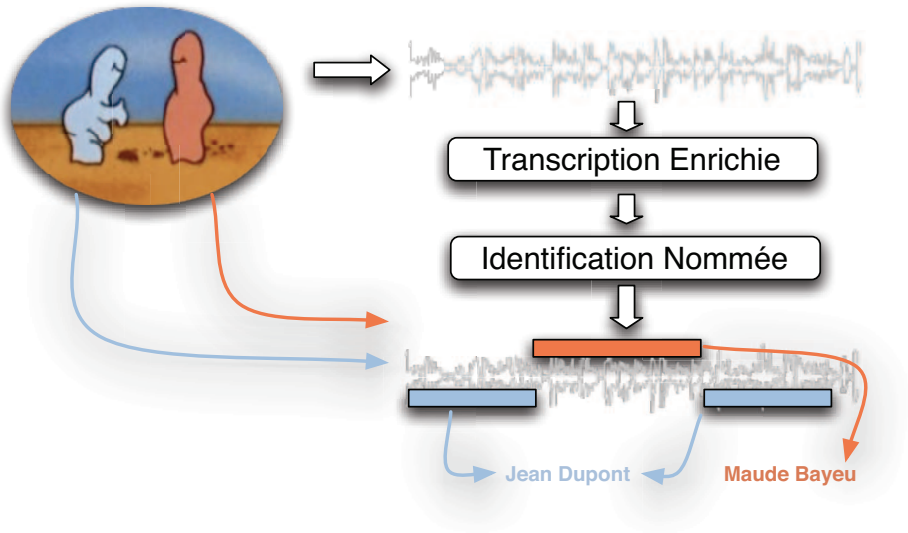

FIGURE 3.1 – *Aperçu global de l'identification nommée du locuteur utilisant la transcription du signal audio*

3.4.1 Hypothèses

Identifier les locuteurs en se basant sur la transcription d'un document audio nécessite que plusieurs hypothèses soient vérifiées :

- Les locuteurs sont **annoncés par leurs noms complets** dans le document.
  À partir du moment où les noms complets des locuteurs sont présents dans la transcription d'un document audio, il semble possible de les exploiter pour identifier les différents intervenants.
  C'est une hypothèse forte, qui couvre un grand nombre de documents. Les journaux d'informations (qu'ils soient radiophoniques ou télévisuels) répondent tout à fait à cette hypothèse. Dans ce type de document, la majorité des locuteurs s'annoncent ou sont annoncés. C'est systématiquement le cas des présentateurs et des invités par exemple. Dans les enregistrements de réunions, les locuteurs sont présentés et se passent la parole : il est donc possible de récupérer leurs noms complets. De manière plus générale, tout type de document où les locuteurs s'annoncent ou sont annoncés vont pouvoir être exploités avec les techniques d'identification nommée basées sur la transcription du signal audio.
- Les noms complets sont exploitables dans la transcription ; ils ont été **correctement transcrits**.
  Les systèmes d'INL doivent être capable de transcrire correctement les noms complets présents dans l'enregistrement si ils veulent pouvoir les

exploiter. Lorsqu'ils sont bien transcrits, c'est ensuite le travail des systèmes de détection des entités nommées (présentés dans la section 2.5) d'identifier les différents noms complets.
Dans le cadre de la mise en place de systèmes automatiques d'INL, cette hypothèse est importante : la transcription des noms propres est un réel problème pour les systèmes automatiques de reconnaissance de la parole puisque, généralement, ils ne font pas partie du vocabulaire du SRAP (Béchet et al., 2000).

– Le **contexte lexical** au voisinage d'un nom complet permet de déterminer si ce nom complet est un **locuteur du document**.
Cette dernière hypothèse rend possible l'attribution d'un nom complet présent dans la transcription à un des locuteurs du document. Pour ce faire, il faut que le contexte lexical de chaque nom complet permette de déterminer si ce nom complet se rapporte à un locuteur du document et à quel locuteur.

3.4.2 Attribution locale

La technique utilisée pour attribuer un nom complet détecté à un des locuteurs du document est commune à tous les systèmes d'INL à partir de la transcription. Cette technique consiste à analyser le contexte lexical d'un nom complet de manière à déterminer si ce nom complet se rapporte à un locuteur qui est en train de parler (tour de parole courant), au locuteur qui parle ensuite (tour de parole suivant) ou au locuteur qui vient de parler (tour de parole précédent). La figure 3.2 résume le principe de cette attribution, que l'on doit aux travaux de Leonardo-Canseco (Canseco-Rodriguez et al., 2005).

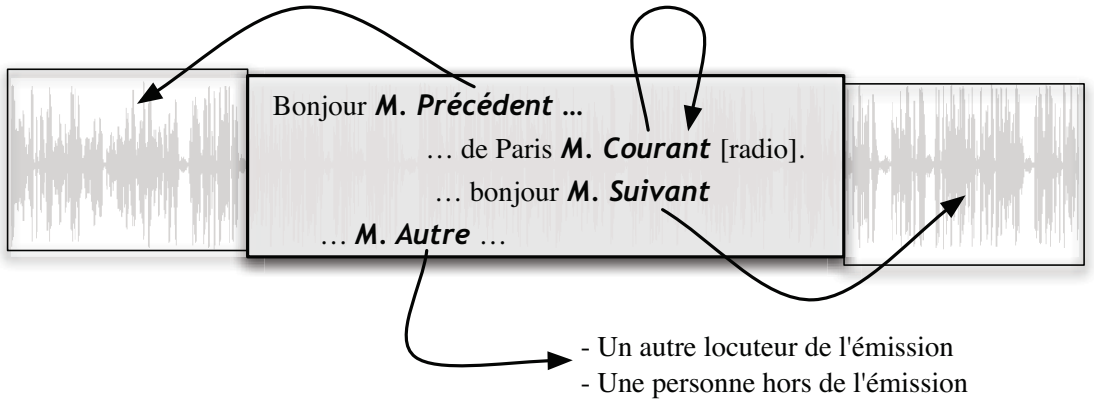

FIGURE 3.2 – *Principe de base des systèmes d'identification nommée basés sur la transcription*

Typiquement, un nom complet peut être attribué au locuteur courant lorsque

celui-ci s'annonce ou se présente. Par exemple les journalistes ont pour habitude de finir leurs interventions en rappelant leur identité, comme par exemple « *c'était Paul Dupond en direct de ...* ». Ici, le nom complet *Paul Dupond* devrait être attribué au locuteur qui a prononcé cette phrase (le locuteur courant). À l'inverse, un locuteur prononçant cette phrase « *Nous écoutons maintenant, Jean Durand.* » annonce une personne qui va parler ensuite. Le nom complet *Jean Durand* devra donc être attribué au locuteur suivant. Pour finir, une phrase du type « *Merci Maude Bayeu, nous passons maintenant à la météo* » rappelle le locuteur qui vient de s'exprimer juste avant. Le nom complet *Maude Bayeu* sera donc attribué au locuteur précédent.

Certains noms complets ne désignent pas des locuteurs du document. Par exemple, *Jean-Jacques Rousseau* aura peu de chance d'être un des intervenants du document. En revanche, il sera surement un des sujets des interventions. D'autres noms complets font bien référence à des locuteurs du document, mais ces locuteurs ne s'expriment pas dans un des tours de parole contigus. Ces cas sont appelés « autre » sur la figure 3.2.

Ce type d'attribution sera par la suite appelée « attribution locale ». On attribue ici un nom complet à un des tours de parole contigu au tour de parole où le nom complet a été détecté. Aucune attribution au niveau du document entier n'est pour l'instant effectuée, seuls les tours de paroles contigus sont concernés. La propagation de ces attributions locales au sein de l'ensemble des tours de parole est une étape appelée « attribution globale ».

3.4.3 Attribution globale

Afin de bien comprendre la problématique de l'attribution globale des noms complets, il faut placer les attributions locales dans le contexte d'une transcription enrichie (cf. 2.4).

Chaque nom complet détecté dans un tour de parole est attribué au tour de parole courant, précédent ou suivant. Chacun de ces tours de parole appartient à un locuteur, que l'INL cherche à nommer en lui attribuant un et un seul nom complet. Or, plusieurs noms complets peuvent être candidats pour le même locuteur. Il est tout à fait possible que plusieurs noms différents soient proposés pour un même locuteur, suite à une erreur du processus d'attribution locale. Par exemple, dans la figure 3.3, deux noms complets sont possibles pour le LOCUTEUR 2 : *Maude Bayeu* et *Pierre Moscovici*.

L'attribution du nom complet au locuteur consiste à choisir un nom complet parmi les noms complet obtenus à partir des décisions locales. Cette phase sera nommée « décision globale ».

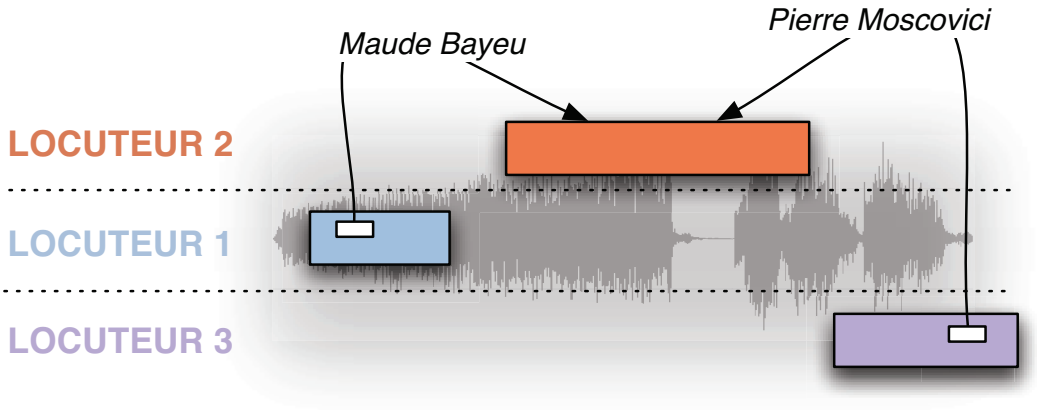

FIGURE 3.3 – *Exemple de décisions locales conflictuelles*

3.4.4 Processus d'attribution

Le processus d'INL se résume en quatre principales étapes décrites dans la figure 3.4 :

- préparation d'une transcription enrichie,
- attributions locales,
- attributions globales,
- production d'une transcription avec les locuteurs nommés.

Dans l'étape 1, la transcription en mots est enrichie d'une segmentation et d'une classification en locuteurs ainsi que d'un étiquetage des entités nommées. Chacune des entités nommées de type *PERSONNE* est ensuite attribuée localement au tour de parole correspondant dans l'étape 2.

Par exemple, le nom complet *C* est attribué aux tours de parole 5 et 11, qui correspondent respectivement à deux locuteurs différents (locuteur 3 et locuteur 1). Un nom complet ne pouvant être attribué qu'à un locuteur, *C* se retrouve donc en conflit entre le locuteur 3 et le locuteur 1. Le nom *B* quant à lui se retrouve aussi en conflit, puisqu'il est attribué aux tours de parole 1 et 8 qui appartiennent respectivement au locuteur 1 et au locuteur 2.

À l'issue de cette seconde étape, plusieurs noms complets sont donc candidats pour un même locuteur : le locuteur 1 peut être nommé avec *A*, *B* ou *C*, le locuteur 2 avec *A* ou *B* et le locuteur 3 avec *C*. En revanche les locuteurs 4 et 5 n'ont pas de candidats. Les méthodes d'INL utilisent un processus de décision (lors de l'étape 3) permettant de régler les conflits entre ces attributions globales afin de nommer chaque locuteur avec un et un seul nom complet.

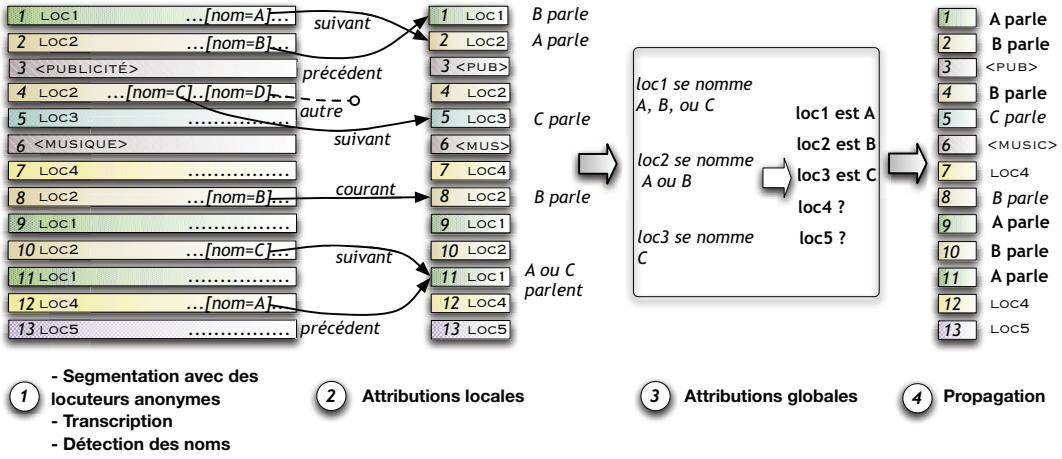

FIGURE 3.4 – *Vue globale du processus attribution d'un nom complet à une classe de locuteur anonyme*

Une fois les conflits de la phase d'attribution globale résolus, la dernière étape consiste à produire une transcription enrichie avec les identités des locuteurs choisies.

Même si les façons de réaliser les étapes diffèrent entre les méthodes d'identification nommée, les principes qui viennent d'être décrits sont communs à toutes. Le traitement d'un enregistrement audio avec un système d'identification nommée se résume aux étapes qui sont décrites dans la figure 3.5.

Deux principales approches sont proposées dans la littérature pour l'attribution locale des noms complets, à savoir une approche symbolique avec l'utilisation de règles écrites manuellement et plusieurs approches à base de méthodes statistiques. La section suivante décrit l'approche symbolique. Les méthodes à base de systèmes statistiques sont décrites dans la section 3.6.

3.5 Approche symbolique

Les travaux réalisés par Canseco-Rodriguez (Canseco-Rodriguez et al., 2005; Canseco-Rodriguez, 2006) ont été les premiers à montrer que le couple prénom/patronyme d'un locuteur apparaissant dans un contexte lexical donné permettait d'identifier de manière précise l'identité des locuteurs s'exprimant dans les tours de parole contigus. Ces travaux sont à la base de toutes les techniques d'INL, ils ont en particulier défini le principe de l'attribution locale décrit en section 3.4.2.

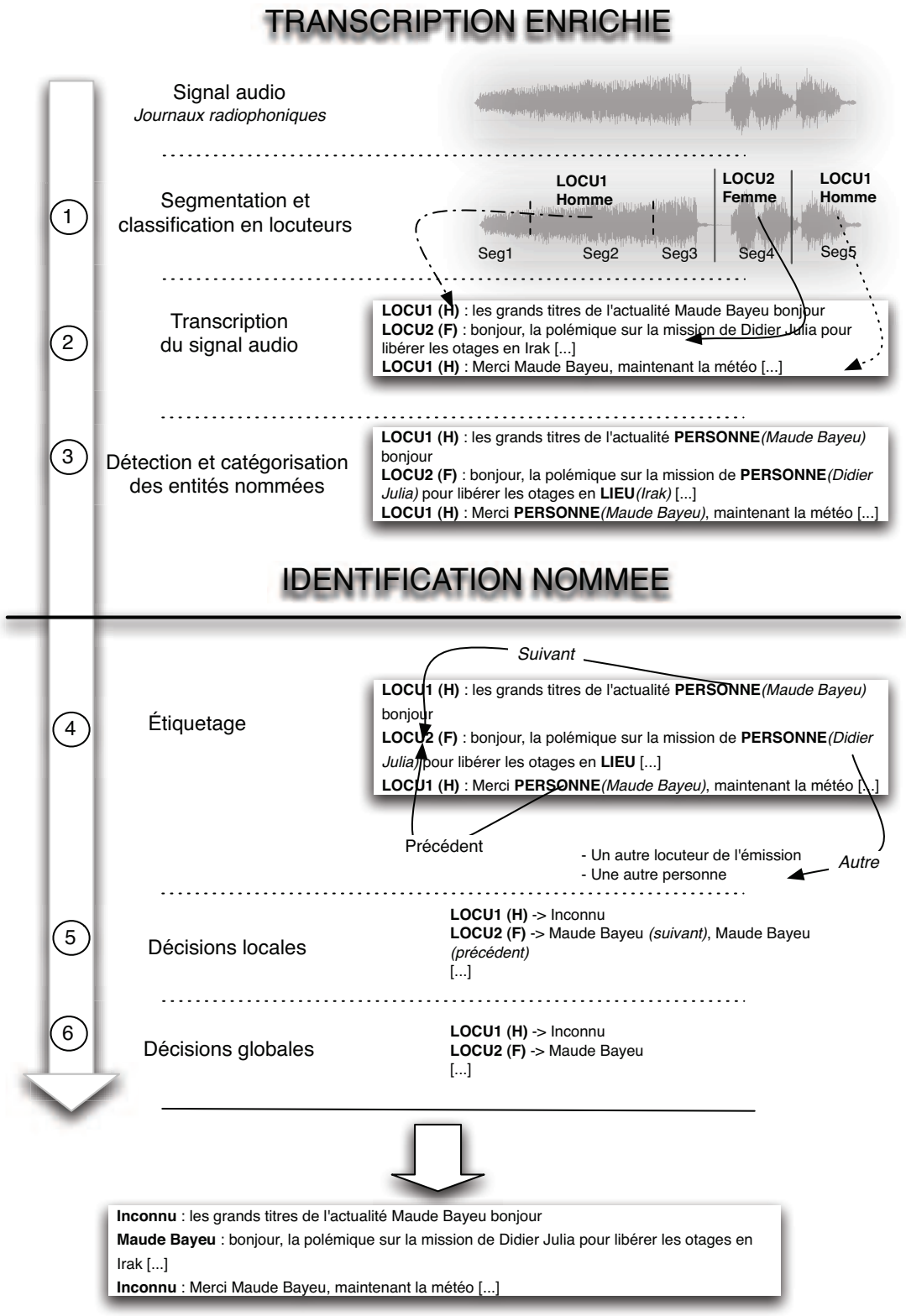

FIGURE 3.5 – *Aperçu global d'un processus d'identification nommée du locuteur*

3.5.1 Règles linguistiques

Ces travaux s'appuient sur l'utilisation de règles linguistiques pour réaliser l'attribution locale des noms complets. L'idée est d'utiliser le contexte linguistique de chaque nom complet pour lui attribuer une des trois étiquettes : « locuteur précédent », « locuteur courant » ou « locuteur suivant ». Ces règles ont été définies manuellement après analyse d'un corpus de langue anglaise (*Hub-4e* [1]). Ce corpus est composé de transcriptions manuelles d'émissions de radio et de télévisions américaines (ABC, CNN, CSPAN et NPR) enregistrées entre 1993 et 1998.

Dans le but de réaliser des règles les plus génériques possibles, certains mots ou groupes de mots ont été regroupés au sein de classes sémantiques. Ces groupes de mots ont ensuite été substitués par leur classe sémantique dans les textes. Par exemple, tous les noms de lieux sont remplacés par le nom de leur classe sémantique [location]. De ce fait, les deux règles « *reporting from Bagdad* » et « *reporting from Paris* » peuvent être fusionnées en une seule règle « *reporting from [location]* ». Plusieurs dictionnaires de classes sont utilisés pour remplacer les mots. Les dictionnaires utilisés, construits à partir d'informations extraites d'internet, sont au nombre de 6 :

- **noms complets**, couple prénom/patronyme ([name]) : ce dictionnaire contient les noms de personnes anglaises les plus communs, y compris les personnes célèbres et les personnalités internationales. Ce dictionnaire a été complété par des informations trouvées dans les textes à traiter, et notamment par l'identité des différents locuteurs (tous corpus confondus).
- **toponymes** [(location]) : ce dictionnaire est constitué des lieux géographiques les plus courants comme les villes, les régions, les états, les monuments...
- **noms d'émissions** ([show]) : ce dictionnaire contient le nom des émissions et des radios. Les informations ont été extraites des transcriptions de référence (tous corpus confondus).
- **titres** ([title]) : ce dictionnaire est composé des titres (Monsieur, Madame, ...), des rangs militaires et des professions.
- **vocabulaire de la communication** ([comm]) : c'est un dictionnaire global incluant une variété de mots et d'expressions utilisés pour gérer la communication. Il contient une sélection des mots les plus fréquemment utilisés pour les remerciements ([thanks]), les salutations ([greet]), les consentements ([agree], [comm-agreement]), les questions ([quest]), les réflexions personnelles ([reflect]) et les pronoms démonstratifs ([dem]).
- **flexions** : ce dictionnaire contient la liste complète des verbes anglais. Ils sont conjugués et classés comme verbe à l'infinitif ([infverb]), au passé

1. Disponible via LDC : `http://www.ldc.upenn.edu/`

([pasverb]), au présent ([preverb]) ou au participe ([parverb]).

À partir du corpus de développement de 150 heures d'enregistrements de radio et de télévision en langue anglaise (corpus Hub-4e), les règles les plus fréquentes pour trouver l'identité des locuteurs ont été extraites. Elles sont décrites dans le tableau 3.1. Au final, douze règles sont utilisées pour désigner le locuteur courant, 34 pour le suivant et 6 pour le précédent.

Nombre d'occurrences	Règle
3162	[title] [name]
848	I am [name]
673	[show]'s [name]
382	[agree] [name]
293	[name] [show] [location]
186	[show]'s [name] reports
176	[thanks] [name]

TABLE *3.1* – *Règles les plus fréquentes pour déterminer l'identité d'un locuteur sur le corpus de développement*

À ces règles viennent s'ajouter d'autres mécanismes comme l'utilisation d'un caractère générique (*) permettant de remplacer n'importe quel mot ou suite de mots, de manière à généraliser les règles. Des exemples de règles avec caractère générique sont données dans le tableau 3.2.

Nombre d'occurrences	Règle
458	with [comm] us * [name]
109	joining * [name]
108	[name] * joins
45	with * [comm] me
24	[comm-agreement] * [name] reporting
24	we are joined * by [name]

TABLE *3.2* – *Exemple de règles utilisant des joker (*) sur le corpus de développement*

3.5.2 Expériences et métriques d'évaluation

L'évaluation des règles d'attribution locale a été réalisée sur un corpus de test de 10 heures de données en langue anglaise (corpus Hub-4e, années 1997, 1998 et 1999). Deux types de transcriptions enrichies ont été utilisées : transcriptions manuelles et automatiques. Le taux d'erreur mot pour les transcriptions automatiques est de 18 %, le système utilisé est celui présenté dans (Lamel et al., 2002).

Les erreurs réalisées par le système de transcription automatique, et notamment les erreurs de découpage en tours de paroles (produits par le système de segmentation et de classification en locuteur), posent problème pour l'évaluation des règles. Lorsqu'un tour de parole produit par le système automatique correspond à plusieurs tours de parole de la référence, il est difficile de décider si le résultat de la règle doit être considéré comme bon ou mauvais. Pour résoudre ce problème, deux types de tours de parole ont été pris en compte dans les transcriptions réalisées automatiquement : les tours de paroles « corrects », c'est à dire les tours de paroles qui correspondent aux tours de parole de la référence et les tours de paroles « bruités », c'est à dire les tours de paroles qui regroupent plusieurs tours de parole de la référence (à cause d'une erreur du système de segmentation et de classification). Les erreurs commises par le système sont ensuite calculées selon différents cas de figure :

- **Identité correcte, tour correct** : l'identité extraite de la transcription est associée avec un tour de parole correct et correspond exactement à l'identité indiquée dans la référence (prénom et patronyme).
- **Identité correcte, tour bruité** : l'identité extraite de la transcription est associée avec un tour de parole bruité et correspond exactement à une des identités indiquées dans la référence (prénom et patronyme).
- **Identité partielle, tour correct** : l'identité extraite de la transcription est associée avec un tour de parole correct et correspond partiellement à l'identité indiquée dans la référence. Par exemple, seul le prénom a été extrait.
- **Identité partielle, tour bruité** : l'identité extraite de la transcription est associée avec un tour de parole bruité et correspond partiellement à une des identités indiquées dans la référence.
- **Indéfini** : la règle correspond à un locuteur inconnu dans la référence (ces cas sont exclus des résultats de test)
- **Mauvaise identification** : Aucun des cas ci-dessus ne s'applique, c'est donc une erreur d'association avec un tour de parole.

3.5.3 Résultats

Les tableaux 3.3 et 3.4 résument les résultats obtenus sur les transcriptions manuelles et automatiques. Sur le corpus de test avec les transcriptions manuelles, les règles identifiant le locuteur courant sont les plus fiables (aucune fausse identification), tandis que les règles identifiant le locuteur précédent (43,1 % de fausse identification) et suivant (20,2 % de fausse identification) présentent des taux d'erreur plus élevés. Cette tendance se confirme avec l'utilisation de transcriptions automatiques, les règles de type « courant » représentent 8 % des erreurs, les règles de type « suivant » 19 % des erreurs et les règles de type « précédent » 50 % des erreurs. Le taux d'erreur global des règles sur le corpus de test

avec les transcriptions manuelles est d'environ 13 %, contre environ 18 % pour les transcriptions automatiques.

	courant	*suivant*	*précédent*
Id. correcte, tour correct	115 (95,0 %)	50 (55,0 %)	7 (16,0 %)
Id. correcte, tour bruité	-	-	-
Id. partielle, tour correct	7 (5,0 %)	22 (24,8 %)	18 (40,9 %)
Id. partielle, tour bruité	-	-	-
#Mauvaise identité	-	16 (20,2 %)	19 (43,1 %)
#Indéfini	-	3	1
Total	122	91	45

TABLE 3.3 – *Taux d'erreur des règles manuelles sur les transcriptions manuelles (corpus d'évaluation 97-98-99 Hub-4e)*

	courant	*suivant*	*précédent*
Id. correcte, tour correct	94 (84,0%)	38 (60,3%)	8 (21,0%)
Id. correcte, tour bruité	2 (1,7%)	3 (4,8%)	-
Id. partielle, tour correct	7 (6,2%)	10 (15,9%)	11 (29,0%)
Id. partielle, tour bruité	-	-	-
#Mauvaise identité	9 (8,0%)	12 (19,0%)	19 (50,0%)
#Indéfini	-	2	1
Total	112	65	39

TABLE 3.4 – *Taux d'erreur des règles manuelles sur les transcriptions automatiques (corpus d'évaluation 97-98-99 Hub-4e)*

Ces travaux pionniers démontrent qu'identifier les locuteurs d'un document à partir de sa transcription est possible. En revanche, ils ne prennent pas en compte les possibilités offertes par les techniques d'apprentissage automatique : les corpus nécessitent des traitements manuels. Le temps de mise en place de telles règles peut être long suivant la quantité de corpus à analyser et demande une expertise du domaine pour pouvoir être réalisée. De plus, le passage d'un corpus à un autre est fastidieux : il faut réécrire le jeu de règles pour l'utiliser sur des documents d'une autre langue ou provenant d'autres types d'émissions.

Ces travaux traitent de l'attribution locale des noms complets. Aucun processus global permettant de répercuter ces décisions au sein du document entier n'a été mis en œuvre. Ce problème d'attribution globale est abordé dans les différentes approches statistiques dans la section qui suit.

3.6 Approche statistique : N-grammes

Suite aux travaux de Canseco-Rodriguez (2005), le laboratoire de l'Université de Cambridge s'est intéressé à l'identification nommée du locuteur (Tranter, 2006). Leurs travaux consistent à automatiser l'apprentissage des règles linguistiques et à mesurer l'impact de leur système d'identification nommée sur des données provenant de systèmes automatiques (classification en locuteur et transcription automatiques).

3.6.1 Attribution locale : utilisation de N-grammes

Cambridge utilise des N-grammes pour modéliser les règles linguistiques. Ils sont obtenus via un corpus d'apprentissage dans lequel la liste des personnes (leurs noms complets) intervenant dans les enregistrements est connue à l'avance. Les occurrences de ces noms complets sont détectées grâce à cette liste.

Le contexte lexical de chaque nom complet locuteur du document est analysé. Une fenêtre glissante est utilisée sur ce contexte lexical : sa taille varie de 2 à 5 mots (en incluant le nom complet). Pour chaque contexte lexical, le système va générer un ensemble de N-grammes (de 2 à 5-grammes) et va leur assigner une probabilité de désigner de manière correcte le locuteur suivant, courant ou précédent. Un N-gramme (avec un minimum de 2 mots et un maximum de 5) est appris pour chaque glissement de la fenêtre. Chaque N-gramme apparaissant plus d'un certain nombre de fois (cinq dans le cas de ces travaux) est considéré comme une règle de prédiction et est donc retenu comme règle linguistique.

Comme dans les précédents travaux de Canseco-Rodriguez (2005), Tranter (2006) utilise des classes sémantiques pour généraliser les règles. Ces classes sont issues des données d'apprentissage et ont été complétées avec d'autres sources (journaux écrits notamment). Ces classes regroupent celles utilisées par Canseco-Rodriguez, avec quelques différences. Là où Tranter différencie HELLO et GOODBYE, Canseco-Rodriguez les regroupait au sein d'un même concept [comm]. De plus, Tranter n'utilise pas la catégorie [quest] et les verbes.

3.6.2 Attribution globale

Deux processus d'attribution globale ont été étudiés pour cette approche basée sur les N-grammes. Tout d'abord, Tranter a utilisé une combinaison de probabilités pour prendre en compte les règles qui fournissent la même hypothèse. Par la suite, Microsoft (Chengyuan et al., 2007) a mis en place un modèle

d'entropie maximum pour combiner les règles et certaines informations supplémentaires comme le genre des locuteurs.

Nous commençons par décrire le premier processus mis en place par Tranter avant de décrire plus en détail les ajouts réalisés par Chengyuan.

Combinaison des règles linguistiques

Toutes les règles précédemment apprises sont utilisées sur le corpus de test. Lorsqu'une règle avec une probabilité p_1 est déclenchée et suggère un nom complet e_1 pour un locuteur C_a, le score pour l'hypothèse $C_a = e_1$ est incrémenté. Lorsque deux règles sont déclenchées pour le même nom de locuteur, leurs probabilités sont combinées en utilisant la formule :

$$p_{1+2} = 1 - (1 - p_1)(1 - p_2) \tag{3.2}$$

Pour déterminer la probabilité d'attribuer un nom complet e à un locuteur C, un ensemble de règles (N-grammes) applicables $\mathcal{R}(e)$ est utilisé. L'équation 3.2 peut être généralisée de cette manière :

$$P(e|C) = 1 - \prod_{r \in \mathcal{R}(e)} (1 - p_r) \tag{3.3}$$

Modèle d'entropie maximum

Les travaux de (Chengyuan et al., 2007) proposent de remplacer la combinaison présentée dans l'équation 3.2 par un modèle d'entropie maximum (Berger et al., 1996). Cette approche permet de représenter les différentes règles au sein d'un modèle statistique et d'y ajouter des connaissances supplémentaires.

Le modèle d'entropie maximum proposé s'exprime de la manière suivante :

$$P(e|C) = \frac{1}{Z_\lambda(C)} \exp\left(\sum_{k=1}^{F} \lambda_k f_k(e, C)\right) \tag{3.4}$$

Les N-grammes sont utilisés comme attributs $f_k(e, C)$ du modèle d'entropie maximum. λ_k représente les paramètres du modèle (multiplicateur de Lagrange) et $Z_\lambda(C)$ est une constante de normalisation qui permet de s'assurer que la somme des probabilités est égale à un. Les paramètres du modèle sont

appris en utilisant la méthode dite de *Generalized Iterative Scaling* afin d'optimiser la probabilité *a posteriori* des paramètres λ étant donné les données d'apprentissage (Chen et Rosenfeld, 2000). Conjointement aux N-grammes, d'autres attributs sont introduits dans le modèle d'entropie.

Une des premières connaissances ajoutée est le genre des locuteurs. Un attribut modélisant la correspondance du genre du prénom du nom complet avec le genre du tour de parole est ajouté au modèle. Des modèles acoustiques des voix des locuteurs cibles (lorsqu'ils sont disponibles) sont eux ajoutés comme attributs. Ces modèles sont appris sur les échantillons de voix des locuteurs cibles via des mélanges de Gaussiennes à 192 composantes. De plus, la position du nom complet au sein de l'enregistrement est pris en compte (avant le premier tour de parole du locuteur désigné, après le dernier tour de parole du locuteur, etc.).

Décision

Pour attribuer un nom complet à un locuteur, il faut pouvoir choisir parmi les différents candidats possibles. Ce choix est effectué en prenant le nom complet candidat e^* qui a la probabilité maximale pour le locuteur C donné. Pour ce faire, la formule 3.5 est utilisée :

$$e^* = \arg\max_e P(e|C) \text{ si } P(e^*|C) > \alpha \tag{3.5}$$

Un seuil α est utilisé pour filtrer les décisions. Lorsque la probabilité est en dessous de la valeur fixée pour α, aucune décision n'est prise pour C : le locuteur est considéré comme inconnu.

3.6.3 Corpus

Le corpus d'apprentissage utilisé par Tranter contient environ 70 heures d'enregistrements de journaux d'informations en langue anglaise provenant du corpus Hub-4 de 1996/1997. Ce corpus a été transcrit manuellement et les locuteurs identifiés (quand cela était possible). De plus, les erreurs de transcription des noms complets (mauvaise orthographe du prénom ou du nom) de locuteurs intervenant dans document ont été corrigées. Le corpus de développement correspond à celui utilisé pour la campagne RT-04 (NIST, 2004) et se compose de 12 documents enregistrés en février 2001 et de novembre à décembre 2003. Le corpus d'évaluation correspond aussi à celui de la campagne RT-04, soit 12 émissions enregistrées en décembre 2003.

À noter que les corpus utilisés par Chengyuan pour évaluer le système utilisant le modèle d'entropie maximum correspondent à ceux de Tranter mais avec un découpage légèrement différent.

3.6.4 Analyse des données

Afin de tester le système mis en place, les locuteurs de la référence sont comparés aux locuteurs de l'hypothèse générée. Trois cas différents pour ces locuteurs sont possibles :

- Disponible : le nom complet du locuteur de la référence est connu et est présent à l'identique dans la transcription de l'émission ;
- Indisponible : le nom complet du locuteur de la référence est connu mais n'est pas présent à l'identique dans la transcription de l'émission. Plusieurs cas de figure sont possibles : des personnes dont les noms sont cités dans d'autres émissions, l'utilisation de synonymes demandant une connaissance du contexte (le président = Bill Clinton) ou encore l'utilisation de tournures demandant un traitement supplémentaire (John Smith et sa femme Judy = Judy Smith) ;
- Non-nommé : le locuteur n'a pu être nommé dans la référence.

À la différence des travaux de Canseco-Rodiguez, ces travaux ne prennent pas en compte les noms partiels, c'est à dire les identifications à l'aide d'un seul nom ou prénom. Le nom complet associé par le système d'INL à un locuteur doit correspondre exactement au nom renseigné dans la transcription. Le tableau 3.5 présente les résultats pour les différents corpus utilisés par Tranter.

Données	D	I	N	Nombre d'occurrences
Apprentissage	79,3%	4,0%	16,8%	6912
Développement	78,2%	8,8%	13,0%	195
Évaluation	76,8%	12,9%	10,3%	206
Évaluation (srap)	47,3%	42,5%	10,3%	138

TABLE 3.5 – *% de données Disponibles (D), Indisponibles (I) et Non-nommées (N) dans les corpus de Tranter. Cela fixe la limite haute des performances atteignables puisque les noms complets I ne peuvent être récupérés à partir de la transcription. Les résultats en utilisant les transcriptions automatiques (srap) sont aussi donnés.*

La quantité de noms complets disponibles à partir de la transcription reste aux alentours de 78 %. Dans les travaux de Microsoft, ce chiffre monte à 83 % sur le corpus de test. L'utilisation de transcriptions automatiques multiplie par trois les noms complets non disponibles dans la transcription. Cela est du aux nombreuses erreurs de transcription des noms complets.

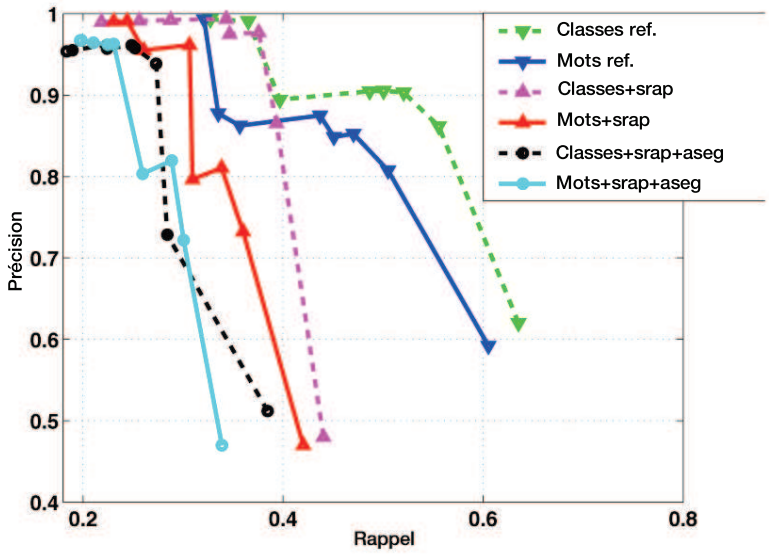

FIGURE 3.6 – *Résultats sur le corpus d'évaluation avec ou sans (mots) l'utilisation de classes sémantiques. Les résultats sont donnés pour les transcriptions de référence (ref), les transcriptions utilisant un SRAP (srap) pour obtenir les mots et les transcriptions utilisant un système de segmentation automatique pour les tours de parole et les locuteurs (aseg)*

3.6.5 Résultats

La figure 3.6 donne les différents résultats en fonction du type de transcription utilisée (*srap* : transcription automatique, *aseg* : segmentation et classification automatiques) et de l'usage des classes sémantiques ou non. Ces résultats montrent que l'utilisation de classes sémantiques pour généraliser les mots présents dans les règles apporte un gain significatif, que ce soit sur des transcriptions manuelles ou des transcriptions automatiques. Sur des transcriptions réalisées manuellement et sur le corpus de développement, à 95 % de précision le rappel est de 60 %. Sur les données d'évaluation, le rappel chute à 38 % pour la même précision. Le tableau 3.6 donne le rappel maximum obtenu sur chaque type de transcription (automatique ou manuelle), ainsi que le rappel à 95% de précision.

Transcription	Segmentation	Rappel maximum	Rappel à 95 % de Précision
ref	ref	64 %	38 %
auto	ref	44 %	38 %
auto	auto	38 %	26 %

TABLE 3.6 – *Résumé des résultats sur le corpus de test. Sont présentés le meilleur rappel obtenu et le rappel à 95 % de précision*

Le taux d'erreur, DER : Diarization Error Rate (Tranter et Reynolds, 2006),

du système de segmentation/classification utilisé est de 6,9 %. Le taux d'erreur du système de transcription, WER : Word Error Rate (McCowan et al., 2004), est quant à lui de 12,6 %. L'utilisation de transcriptions automatiques n'affecte pas le rappel à 95 % de précision mais le rappel maximum possible (en faisant baisser la précision) passe de 64 % à 44 %. L'utilisation de transcriptions et de segmentations automatiques fait chuter le rappel à 26 % (pour 95 % de précision) et fait aussi chuter le rappel maximum possible à 34 %.

Dans (Chengyuan et al., 2007) Microsoft a comparé le système de Tranter avec le système utilisant un modèle d'entropie maximum sur des transcriptions manuelles. Les résultats présentés dans la figure 3.7 et le tableau 3.7 montrent qu'en utilisant les mêmes informations, le modèle d'entropie maximum est plus performant que la simple combinaison utilisée par Tranter. En revanche, l'ajout des modèles de locuteur n'apporte aucun gain significatif. L'utilisation d'informations sur les genres et la position des noms complets donnent les meilleurs résultats.

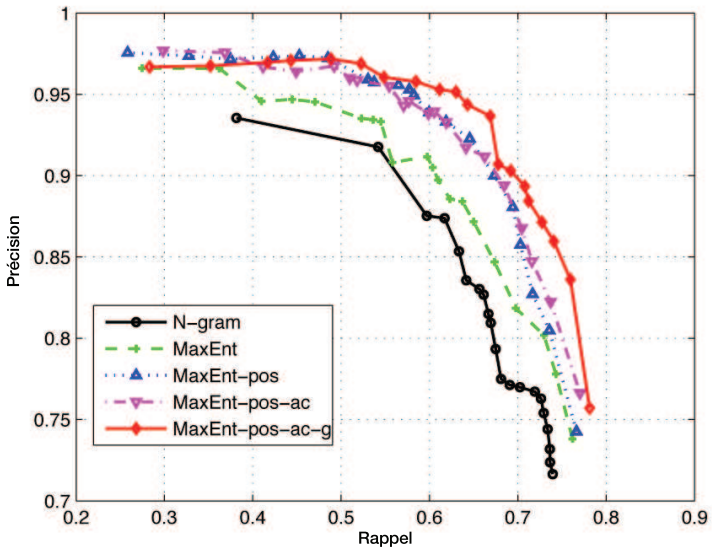

FIGURE 3.7 – *Résultats sur le corpus de test (transcription et segmentation / classification manuelles). Le modèle d'entropie maximum (maxent) et le modèle N-gramme utilisent les mêmes règles lexicales. Les résultats avec l'ajout de la position (pos), des données acoustiques (ac) et l'utilisation du genre (g) sont aussi donnés pour le modèle d'entropie maximum.*

La méthode à base de N-grammes sur fenêtre glissante permet d'automatiser l'apprentissage des règles que Canseco-Rodriguez avait défini à la main. Une propagation relativement simple des scores est effectuée au sein des tours de paroles des locuteurs dans les travaux de Tranter : les différents scores obtenus pour un nom complet n_1 et un locuteur s_a sont cumulés. Les tests réalisés

Système	Rappel
Tranter combinaison	38 %
Entropie	47 %
Entropie + pos.	58 %
Entropie + pos. + ac.	58 %
Entropie + pos. + ac. + genre	67 %

TABLE 3.7 – *Résultat des différents systèmes sur le corpus de test (transcription et segmentation/classification manuelles) à 95 % de précision.* pos. *: position du nom complet dans l'enregistrement,* ac. *: informations acoustiques sur les locuteurs*

montrent de bonnes performances sur des données manuelles, mais aussi leurs limites sur les données automatiques, alors que les systèmes automatiques de transcription et de segmentation sont très performants. Dans (Chengyuan et al., 2007), Microsoft a proposé une amélioration du processus d'attribution en utilisant un modèle d'entropie maximum et des connaissances supplémentaires comme le genre des locuteurs ou la position des noms complets dans l'enregistrement. Ces différents ajouts permettent d'améliorer le rappel de 38 % à 67 % à 95 % de précision.

Dans ces travaux, la détection des noms complets dans la transcription est réalisée grâce à une liste de locuteurs et la détection des autres classes sémantiques est réalisée à partir des données de l'apprentissage étiquetées manuellement.

3.7 Approche statistique : arbre de classification sémantique

Le LIUM (Mauclair et al., 2006; Estève et al., 2007) a proposé une alternative au modèle N-gramme en mettant en place un système d'INL basé sur un arbre de classification sémantique. Cet arbre est utilisé pour les attributions locales (Kuhn et De Mori, 1995) ainsi qu'un détecteur automatique d'entités nommées pour identifier les noms complets dans la transcription. Ces travaux sont les premiers à utiliser un détecteur d'entités nommées pour cette tâche. L'utilisation de ces deux systèmes automatiques et d'un processus de décision permet au système du LIUM de fonctionner de manière entièrement automatique pour identifier les locuteurs d'un document.

3.7.1 Détection des entités nommées

Le système de détection des entités nommées (EN) a été mis en place lors de la campagne d'évaluation ESTER I (Galliano et al., 2005) sur la détection des EN. C'est un système à base de règles. Quelques règles ont été inférées à partir du corpus d'apprentissage d'ESTER I et d'autres ont été développées manuellement (pour, par exemple, réaliser une grammaire pour détecter les dates). De plus, des listes de prénoms, de patronymes, de villes, de pays, etc. ont été ajoutées à la base de connaissances du système.

La liste des EN choisie pour la campagne ESTER I se compose de 8 catégories principales qui sont : les personnes, les lieux, les organisations, les groupes socio-politiques, les montants, les informations temporelles, les produits et les aménagements. Pour son système d'INL, le LIUM n'a retenu que 5 de ces catégories : les personnes, les lieux, les organisations les groupes socio-politiques et les informations temporelles (Estève et al., 2007). Les résultats sur le corpus de développement de la campagne ESTER I sont présentés dans le tableau 3.8 en terme de précision et de rappel.

Type d'EN	Rappel (%)	Précision (%)
personne	98,7 %	92,6 %
lieu	83,9 %	87,9 %
organisation	86,1 %	84,5 %
groupe socio-politique	84,0 %	91,8 %
information temporelle	96,7 %	87,7 %

TABLE 3.8 – *Performances du système de détection des EN du LIUM sur le corpus de développement ESTER I*

Comme pour les autres systèmes d'INL, lorsqu'une EN est détectée, la suite de mots correspondant est remplacée par le type de l'EN dans la transcription.

3.7.2 Attributions

Les arbres de classification sémantiques (SCT - Semantic Classification Tree) sont des arbres de décision binaire s'appuyant sur des expressions régulières comme paramètres de décision pour classer les échantillons **??**. Dans le cadre de l'identification nommée, ils sont utilisés pour réaliser l'attribution locale des étiquettes « précédent », « courant » et « suivant ». Ces arbres sont activés sur le contexte lexical gauche et droit de chaque nom complet détecté par le système d'EN. Au maximum cinq mots à gauche et cinq mots à droite sont gardés pour l'apprentissage et l'utilisation de l'arbre. Le SCT attribue une probabilité

à chaque étiquette possible pour un nom complet détecté dans la transcription. La figure 3.8 donne un exemple d'arbre de classification. Les détails sur l'apprentissage de l'arbre sont décrits dans (Mauclair et al., 2006).

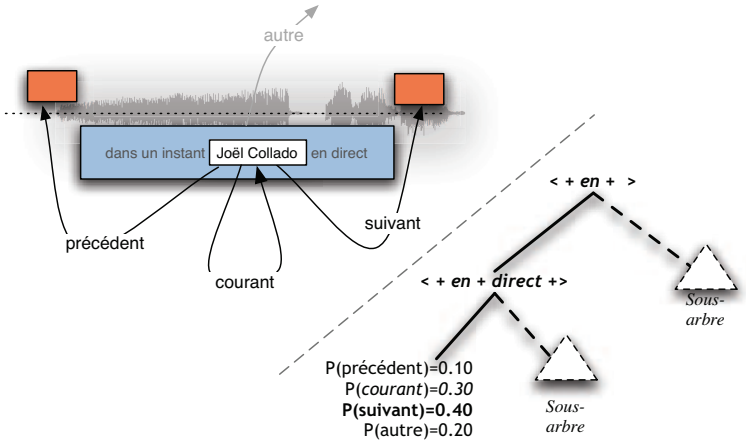

FIGURE 3.8 – *Exemple d'arbre de classification sémantique*

Seule l'étiquette avec la plus grande probabilité est prise en compte, les autres probabilités sont ignorées. Lorsque deux décisions du SCT attribuent le même nom complet à un locuteur, les probabilités sont additionnées. On obtient alors, pour chaque nom complet candidat pour nommer un locuteur, un score qui peut être utilisé pour prendre une décision.

Dans l'exemple de la figure 3.9, le tour de parole correspondant au locuteur numéro 2 (abbrégé *LOCU2* dans la figure) se voit affecter deux fois le nom complet *Maude Bayeu*. Dans ce cas, le nom complet *Maude Bayeu* aura pour score 1,22 (0.72 + 0.5).

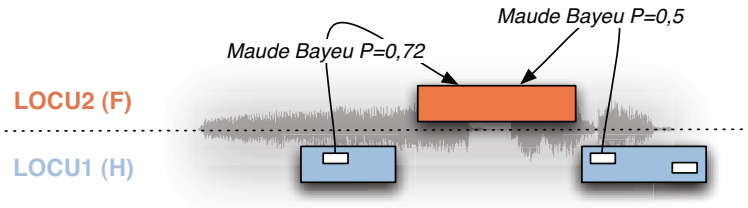

FIGURE 3.9 – *Exemple d'attributions locales du système du LIUM*

Lors du processus d'attribution globale, le nom complet avec le plus grand score pour un locuteur donné est attribué à ce locuteur. Chaque locuteur est traité indépendamment de l'autre, c'est à dire que rien n'empêche le système d'attribuer le même nom complet à deux locuteurs différents. Si la segmentation

et la classification sont considérées comme correctes, ce cas de figure devrait être impossible.

3.7.3 Expériences et résultats

Le système a été évalué sur des données provenant de la campagne d'évaluation des systèmes de transcriptions en français ESTER I (Galliano et al., 2005). Ces données regroupent six radios différentes et correspondent à 81 heures d'enregistrements pour le corpus d'apprentissage (150 émissions), 12,5 heures pour le corpus de développement (26 émissions) et 10 heures pour le corpus de test (18 émissions). Toutes ces données ont été étiquetées en entités nommées grâce au système décrit en section 3.7.1.

Ce système a été comparé à celui de Tranter dans (Estève et al., 2007). Il a été évalué sur des données segmentées et transcrites manuellement, et sur des données segmentées manuellement mais transcrites automatiquement. Ces transcriptions automatiques sont celles fournies par le LIMSI lors de la campagne ESTER I et présentent un taux d'erreur mots de 11,9 %. Les résultats sont exprimés en terme de précision et de rappel dans les figures 3.10 et 3.11. Le système de Tranter à base de N-grammes sert de comparaison.

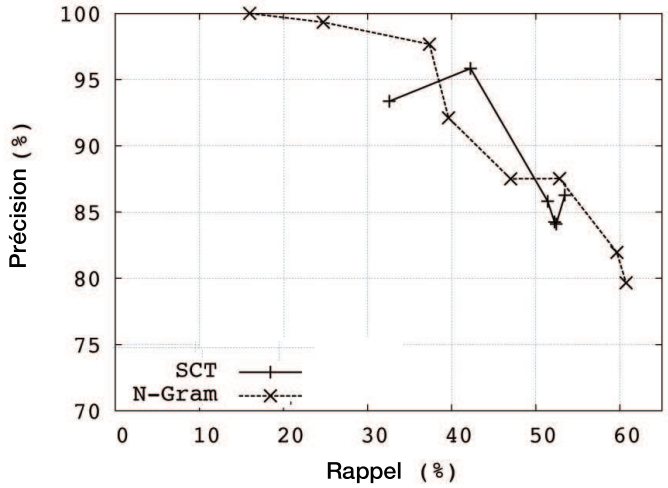

FIGURE 3.10 – *Résultats du système du LIUM et du système à base de N-grammes sur des transcriptions entièrement manuelles (segmentation / classification / transcription)*

Sur des transcriptions manuelles les deux systèmes font presque jeu égal : à 95 % de précision, le rappel est légèrement inférieur à 40 %. En revanche, sur des transcriptions automatiques, le système à base de SCT est plus performant : les N-grammes atteignent une précision maximum de 91 % pour un rappel de

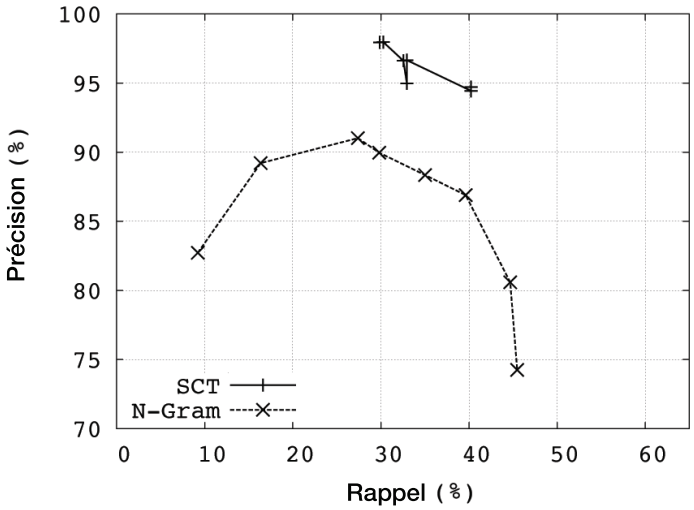

FIGURE *3.11* – *Résultats du système du LIUM et du système à base de N-grammes sur des transcriptions automatiques et une segmentation/classification manuelle*

28 % quand le système à base de SCT obtient un rappel de presque 40 % à 95 % de précision.

3.8 Bilan

Les travaux sur l'identification nommée du locuteur utilisant la transcription enrichie du signal sont relativement récents puisque les premiers travaux menés par Canseco-Rodriguez datent de 2004. Ces travaux reposent tous sur les mêmes hypothèses de base, à savoir que les locuteurs s'annoncent dans le document lui même, que leurs noms ont bien été transcrits et que le contexte lexical permet de déterminer s'il se rapporte au locuteur qui parle actuellement, à celui qui va parler juste après, ou à celui qui a parlé juste avant.

Canseco-Rodriguez a montré que lorsqu'un locuteur était annoncé dans le document, il était possible, via son contexte lexical, de décider s'il désignait le locuteur précédent, courant ou suivant. Les travaux de Tranter et Microsoft ont montré que la majorité des locuteurs s'annoncent et que le rappel maximum qu'il est possible d'atteindre se situe entre 78 % et 83 %, en fonction des corpus. Deux des principales hypothèses données en section 3.4.1 ont donc été vérifiées.

Dans le but d'automatiser au maximum les processus d'INL, plusieurs problèmes restent à résoudre. Tout d'abord, il faut disposer d'un détecteur d'entités nommées qui va permettre de s'affranchir des dictionnaires de classes sémantiques et des listes de locuteurs pré-établies utilisées par Canseco-Rodriguez ou

encore Tranter. Les travaux du LIUM ont effectué un premier pas en ce sens, mais le détecteur d'entités nommées utilisé est assez sommaire.

Le processus d'attribution globale et la gestion des différents conflits qu'il engendre nécessiteraient d'être approfondis. Seuls les travaux de Microsoft ont proposé un modèle d'entropie maximum allant dans ce sens puisqu'il permet de combiner plusieurs informations (le genre, des modèles acoustiques, la position du locuteur dans l'enregistrement) pour renforcer les décisions du modèle N-gramme. Mais les conflits (plusieurs candidats possibles pour un même locuteur, avec éventuellement beaucoup d'incertitude) apparaissant pour nommer un même locuteur n'ont pas été étudiés. À chaque fois, le locuteur ayant la plus haute probabilité (ou le plus haut score pour les travaux du LIUM) est choisi, sans prendre en compte l'incertitude et les conflits présents.

Pour finir l'hypothèse concernant la présence des noms complets (correctement transcrits) reste encore à étudier en profondeur, surtout sur des transcriptions réalisées par des SRAP. Tranter et le LIUM ont réalisé des expériences sur des transcriptions automatiques en constatant que les résultats étaient nettement inférieurs à ceux obtenus sur des transcriptions manuelles. Ces résultats mériteraient d'être approfondis afin de valider cette dernière hypothèse.

Chapitre 4

Milesin : Un système d'INL par analyse conjointe du signal et de sa transcription

Sommaire

Les travaux déjà effectués sur l'INL ont permis de prouver l'efficacité de l'approche utilisant la transcription du signal audio. Comme décrit dans la section 3.8, des progrès restent à réaliser pour automatiser le processus, notamment dans la détection des noms complets et des classes sémantiques. De plus, les travaux ont majoritairement étudié les techniques d'attribution locale (règles linguistiques, N-grammes, arbre de classification), mais peu de travaux sur le processus d'attribution globale ont été menés.

Dans ce chapitre, nous commençons par décrire le système de détection des entités nommées utilisé, puis nous décrirons ensuite l'arbre de classification sémantique mis en œuvre pour réaliser les attributions locales. Nous finirons par expliquer le processus d'attribution globale reposant sur la théorie des fonctions de croyance avant de présenter les résultats du système. Ces travaux font suite à ceux déjà menés au LIUM (cf. section 3.7).

4.1 Détection des entités nommées

Une des étapes fondamentales de tout système d'INL reposant sur la transcription du signal audio est la détection des noms complets et des différentes classes sémantiques. Dans le but d'automatiser au maximum les systèmes d'INL, cette phase peut être réalisée à l'aide de systèmes de détection des entités nommées. Les entités nommées de type *personne* vont pouvoir être exploitées pour extraire les noms complets de la transcription, les autres entités nommées vont servir de classes sémantiques pour généraliser la transcription.

4.1.1 La campagne d'évaluation ESTER 2

En 2008, lors de la campagne d'évaluation ESTER 2 (Galliano et al., 2009), une tâche d'évaluation des systèmes de détection des entités nommées a été proposée. Cette tâche a permis d'évaluer les différents systèmes de détection des entités nommées sur des transcriptions manuelles mais aussi sur des transcriptions automatiques provenant de SRAP. Plusieurs laboratoires universitaires français ont participé à cette tâche comme le LINA, le LIMSI, le LIA, le LI (Tours) ou encore le LSIS. Deux entreprises privées ont aussi participé à l'évaluation : Synapse et Xerox.

La mesure utilisée pour l'évaluation est le Slot Error Rate (SER) décrit dans (Makhoul et al., 1999). La liste des entités nommées à étiqueter pour la campagne ESTER 2 est la suivante : les personnes, les lieux, les organisations, les productions humaines (moyens de transports, récompenses, œuvres artistiques, ...), les montants, le temps et les fonctions (politiques, administratives, ...).

Sur des transcriptions réalisées manuellement, Synapse et Xerox ont (de loin) les systèmes le plus performants avec un SER inférieur à 10 %. Le troisième système est celui du LIA (LIA_NE) qui obtient un SER de 23,9 %. En revanche, sur des transcriptions réalisées automatiquement, c'est LIA_NE qui arrive en première position avec un taux d'erreur de 43,4 % sur les transcriptions du LIMSI qui affichent un taux d'erreur mot (WER) de 12,1 % (soit le meilleur système de la campagne d'évaluation ESTER 2). D'autres tests ont été réalisés avec des transcriptions contenant plus d'erreurs (17,83 % et 26,09% de WER) : LIA_NE obtient systématiquement les meilleurs performances. Les résultats sont présentés dans le tableau 4.1.

%WER	Transcription manuelle de référence			Transcription LIMSI 12.11			Transcription LIUM 17.83			Transcription IRISA 26.09		
Participants	%S	%P	%R	%S	%P	%R	%S	%P	%R	%S	%P	%R
LIA	23.9	86.46	71.85	**43.4**	79.52	59.45	**51.6**	76.51	55.02	**56.8**	72.26	49.02
LIMSI	30.9	81.15	70.94	45.3	75.13	62.33	55.5	70.50	57.52	61.2	66.13	50.67
LINA	37.1	80.75	55.48	54.0	71.98	44.01	60.4	68.76	40.84	65.2	63.66	35.66
LI Tours	33.7	79.39	65.82	50.7	71.36	54.16	80.8	56.59	46.46	82.9	51.28	42.38
LSIS	35.0	82.65	73.07	55.3	70.23	58.39	86.5	70.36	28.66	88.6	67.03	25.22
Synapse	9.9	93.02	89.37	44.9	76.39	67.16	60.7	70.26	59.21	66.2	65.95	52.71
Xerox	**9.8**	93.61	91.50	44.6	58.91	70.06	—	—	—	—	—	—

TABLE *4.1* – *Performances de chaque participant à la tâche de détection des entités nommées de la campagne ESTER 2 (Galliano et al., 2009). Les résultats sont donnés en terme de Slot Error Rate (%S), Rappel (%R) et Précision (%P)*

4.1.2 Le système LIA_NE

Lors de la campagne ESTER 2, le système le plus performant étant LIA_NE (Bechet et Charton, 2010), il a été choisi comme détecteur d'entités nommées pour les travaux de cette thèse. LIA_NE a la particularité d'avoir été conçu spécifiquement pour être utilisé sur des sorties de transcriptions automatiques. Ces transcriptions ont des caractéristiques qui peuvent mettre en défaut les systèmes de détection d'EN prévus pour être utilisés sur des textes écrits tels que les livres ou les journaux.

Les transcriptions automatiques proviennent de documents audio qui contiennent de la parole plus ou moins spontanée et qui sont soumis à des phénomènes propres à la parole orale comme les disfluences ou les faux départs. Ces transcriptions ne sont pas parfaites et contiennent des erreurs commises par le système de reconnaissance automatique de la parole. La ponctuation et les majuscules sont généralement absents des transcriptions. Même si des modules de post-traitement existent pour les remettre, ils sont eux aussi sujet à erreurs. De plus, les transcriptions automatiques ne contiennent que des mots qui appar-

tiennent au dictionnaire du SRAP alors que les EN sont essentiellement composées de noms propres pour lesquels il est très difficile d'obtenir une liste exhaustive.

LIA_NE utilise tout d'abord un processus génératif grâce à un étiqueteur basé sur des HMM (*Hidden Markov Model*) qui permet d'annoter le document avec les parties du discours et des classes sémantiques (personne, organisation, lieu, produit). Une fois le document annoté, un processus discriminatif basé sur les CRF (McCallum et Li, 2003) (*Conditional Random Field*) est utilisé. Ce processus va permettre de détecter les EN à partir des annotations réalisées par l'étiqueteur.

Afin d'apprendre les caractéristiques de ce modèle CRF, un maximum de connaissances sur les entrées du dictionnaire du SRAP est collecté (de manière non supervisée). Pour réaliser ces collectes, deux méthodes sont mises en œuvre :

- Un processus d'extraction automatique des EN sur de larges corpus de textes.
  Ce processus utilise un étiqueteur en parties du discours associé à des données étiquetées manuellement en entités nommées. Ils vont permettre d'apprendre un modèle CRF pour ensuite étiqueter automatiquement un corpus de 1,3 giga mots. Ce corpus étiqueté automatiquement sera utilisé pour l'apprentissage du modèle CRF.
- L'extraction automatique de données provenant de l'encyclopédie en ligne Wikipedia.
  Pour chaque entrée de Wikipedia, un graphe est généré représentant toutes les formes du mot présentes dans Wikipedia. Par exemple, pour *Paris*, 39 formes ont été extraites comme : *Ville Lumière, Ville de Paris, Paname, Capitale de la France, Département de Paris,* etc. Chaque mot et ses différentes formes sont ensuite associés à une des catégories d'EN de la campagne ESTER (*personne, lieu, organisation, produit*). *Paris* et ses différentes formes seront par exemple associés à la catégorie *lieu*. Cette association est réalisée grâce aux données d'apprentissage étiquetées manuellement et à des lexiques préalablement constitués au LIA.

Les 7 catégories d'entités nommées utilisées pour ESTER 2 ont été gardées comme classes sémantiques.

4.2 Attributions locales : arbre de classification sémantique

Les noms complets étiquetés via le système de détection d'EN sont utilisés comme point de départ pour la phase d'attribution locale. Cette phase est réa-

lisée grâce à un arbre de classification sémantique (SCT - *Semantic Classification Tree*).

Les arbres de classification sémantiques sont des types particuliers d'arbre de décision binaire. Ils sont beaucoup utilisés dans le traitement du langage naturel, comme pour des systèmes de compréhension (Kuhn et De Mori, 1995) ou des détecteurs de noms propres inconnus (Béchet et al., 2000). Dans le cadre de l'identification nommée, ils sont utilisés pour réaliser l'attribution locale des étiquettes « précédent », « courant » et « suivant » décrites dans la section 3.4.2. Le fonctionnement et l'apprentissage des SCT sont décrits de manière exhaustive dans (Kuhn, 1993).

4.2.1 Arbre de classification sémantique

Un SCT est un type particulier d'arbre de décision (Breiman et al., 1984; Cornuéjols et Miclet, 2002). Un arbre de décision permet de classer un objet à l'aide de questions : chaque nœud de l'arbre représente une question, chaque lien est une réponse à la question, et chaque feuille est une classe. La figure 4.1 donne un exemple d'arbre de décision permettant de décider si une expérience peut avoir lieu.

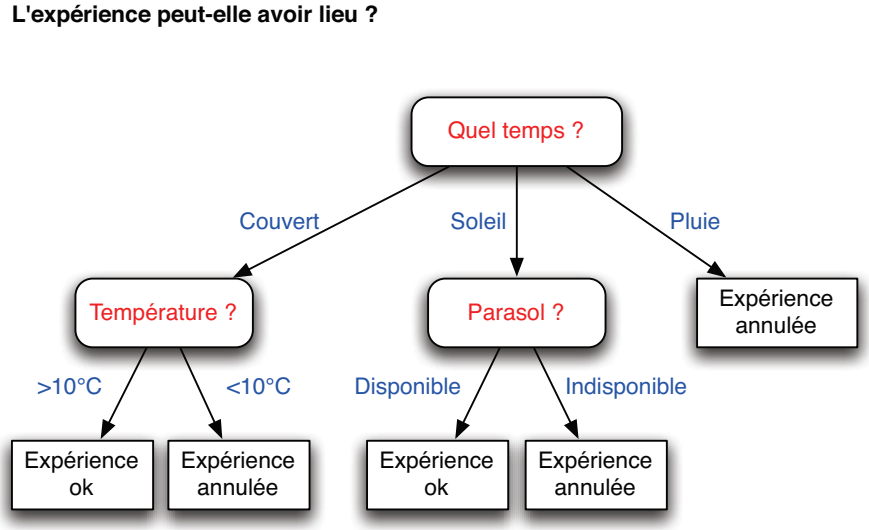

FIGURE 4.1 – *Exemple d'arbre de décision : les questions sont en rouge, les réponses en bleu.*

Les SCT sont des arbres de classification dit binaires. Les questions d'un arbre de décision binaire sont particulières car il n'y a que deux types de réponses possibles : *oui* ou *non*. Les questions des SCT sont des motifs lexicaux

auquel l'échantillon testé doit correspondre. Chaque feuille d'un arbre de classification est associée à une distribution de probabilité sur toutes les classes possibles. Dans le cas de l'identification nommée, les quatre classes possibles sont : *précédent*, *courant*, *suivant* et *autre*. Un exemple de SCT est donné dans la figure 4.2.

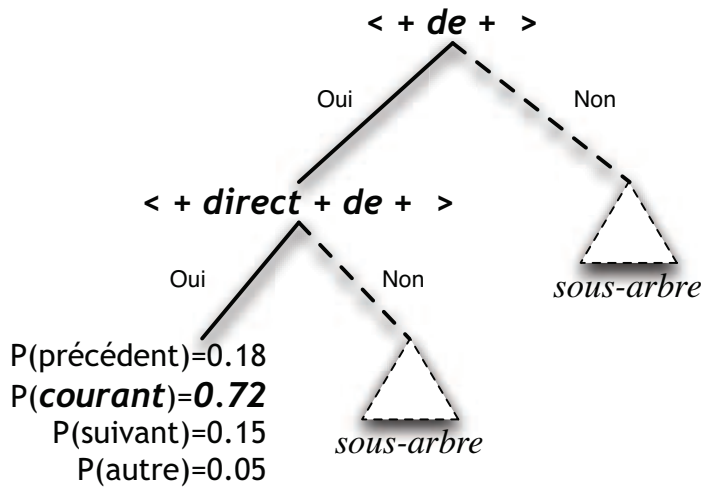

FIGURE 4.2 – *Exemple d'arbre de classification sémantique*

Les SCT apprennent les motifs lexicaux à partir de corpus annotés avec les classes correspondantes. L'apprentissage et la sélection de ces questions est réalisée de manière entièrement automatique : n'importe quel mot du corpus peut apparaître dans une question.

4.2.2 Apprentissage

Pour construire un arbre de classification sémantique, plusieurs éléments sont nécessaires (Breiman et al., 1984) : un corpus d'apprentissage annoté, un ensemble de questions (aussi appelées test), un critère de séparation et une condition d'arrêt. Nous décrivons ces différents éléments ci-dessous.

Corpus d'apprentissage

Le corpus d'apprentissage a été constitué à partir de transcriptions de journaux radiophoniques réalisées manuellement. Ces transcriptions ont été enrichies avec les tours de paroles et le nom des locuteurs lorsque ceux-ci étaient

disponibles. Des traitements préliminaires sont tout d'abord effectués sur ces transcriptions :

- Les transcriptions sont traitées via le détecteur d'entités nommées LIA_NE. Les mots composants les entités nommées retenues sont remplacés par la classe sémantique correspondante dans la transcription.
- Les transcriptions manuelles utilisées pour réaliser le corpus d'apprentissage contiennent plus d'informations que celles produites par un système de transcription automatique. Les transcriptions sont donc normalisées afin de se rapprocher au maximum de celles fournies par un système de transcription automatique. Par exemple, la ponctuation et les majuscules ont été retirées.
- Les articles considérés comme non porteurs de sens (*l', le, la, les, un, une, uns*) ont été éliminés du contexte analysé. En revanche, les prépositions locatives comme *de, dans, à* ont été gardés puisqu'ils peuvent potentiellement concerner une information importante comme un lieu, une radio ou une émission.

Une fois ces pré-traitements effectués, le corpus d'apprentissage est constitué en conservant 5 mots (maximum) à gauche et 5 mots (maximum) à droite de chaque entité nommée de type personne détectée. L'entité nommée détectée et son contexte lexical constituent un exemple d'apprentissage. Le nombre de mots à conserver pour chaque exemple a été déterminé grâce au corpus de développement.

Ensuite, l'arbre a besoin de savoir si une entité nommée de type personne détectée correspond au locuteur suivant, précédent ou courant. Cette phase d'étiquetage de chaque entité nommée est réalisée automatiquement en comparant l'entité nommée détectée aux noms complets des locuteurs suivant, précédent et courant. Cette phase d'étiquetage n'est pas vérifiée manuellement, il est supposé que les noms des locuteurs indiqués dans les transcription manuelles sont corrects.

Pour finir, les questions de l'arbre peuvent porter sur des informations plus globales. Comme proposé dans (Estève et al., 2007), la position du nom complet détecté au sein du tour de parole est ajoutée à l'exemple d'apprentissage. La position correspond aux situations où le nom complet apparaît dans les dix premiers mots (*START_TURN*), au milieu (*MID_TURN*) ou dans les dix derniers mots (*END_TURN*) d'un tour de parole. Une autre question globale permet de savoir si le tour de parole est considéré comme très court (*SHORT_TURN*), à savoir plus petit que le contexte lexical utilisé par l'arbre (cinq mots à gauche du nom complet, cinq mots à droite).

Des exemples d'apprentissage réalisés à partir des transcriptions manuelles sont donnés dans la figure 4.3.

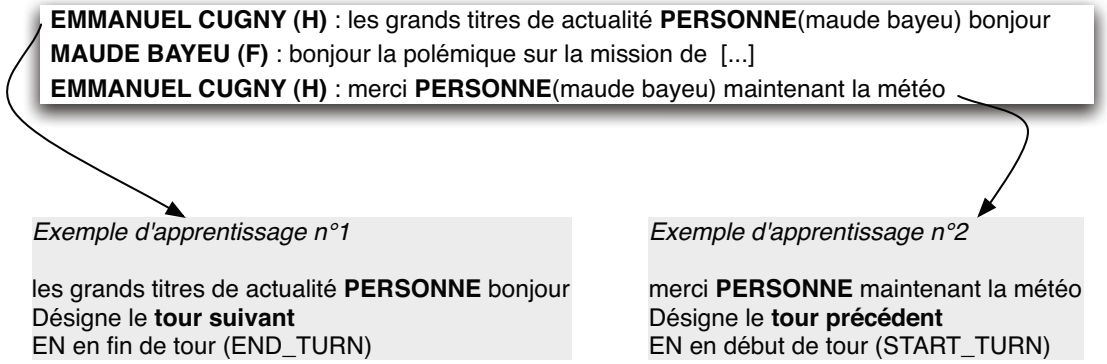

FIGURE 4.3 – *Exemples d'apprentissage du SCT constitués à partir de transcriptions manuelles*

Ensemble de questions

À partir du corpus d'apprentissage, le processus de génération de l'arbre doit construire un ensemble de questions permettant de classer les différents exemples d'apprentissage.

Chaque question correspond à un motif lexical (une sorte d'expression régulière simple) auquel l'exemple est soumis. Si l'exemple correspond au motif lexical, la réponse à la question est *oui*, sinon la réponse est *non*. Ces motifs lexicaux peuvent contenir des mots et des caractères spéciaux (<, > et +). Le < (respectivement >) est une ancre qui désigne le début (respectivement la fin) d'un exemple tandis que + se rapporte à n'importe quelle suite de mots. Ainsi, le motif lexical < + *de* + > correspond à chaque exemple contenant le mot *de*, alors que < + *en direct* + *de* + > correspond à chaque exemple contenant les mots *en direct* et *de* apparaissant dans cet ordre.

La figure 4.4 montre une partie d'un arbre appris automatiquement pour l'INL. Les questions globales et les motifs lexicaux utilisés sont indiqués.

Critère de séparation

Le processus de construction du SCT doit choisir pour chaque question de l'arbre un motif lexical. Pour choisir le meilleur motif, le critère d'*impureté* de Gini $i(T)$ d'un nœud T (Breiman et al., 1984) est utilisé. La meilleure question (ici la meilleure expression régulière) est celle qui présente la plus grande diminution d'*impureté* ΔI entre le nœud T et ses enfants.

Si T désigne un nœud quelconque, et $f(j|T)$ le pourcentage d'exemples de ce nœud qui appartiennent à la classe j (dans notre cas, les quatre classes possibles

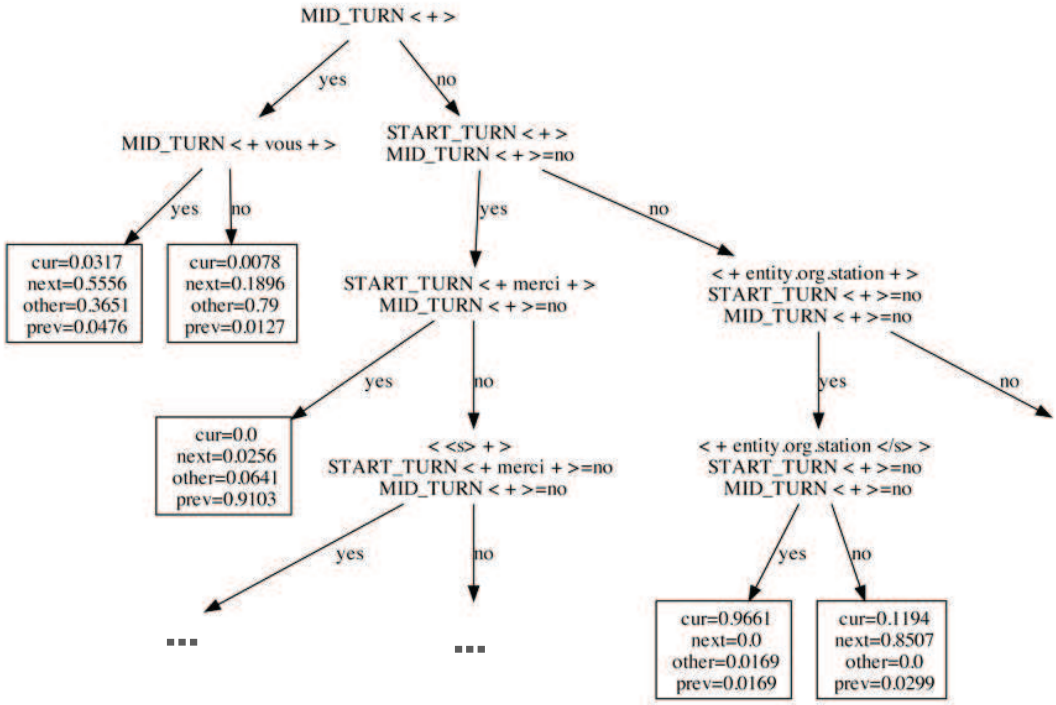

FIGURE 4.4 – *Extrait d'un arbre de classification sémantique appris automatiquement pour l'INL*

sont *précédent*, *courant*, *suivant* et *autre*), alors le critère d'impureté de Gini $i(T)$ est défini par :

$$i(T) = 1 - \sum_j f^2(j|T) \tag{4.1}$$

Par exemple, prenons les 4 classes *précédent*, *courant*, *suivant* et *autre*. 10 exemples du corpus d'apprentissage arrivent dans un nœud T. Sur ces 10 exemples, 4 appartiennent à la classe *précédent*, 3 à la classe *courant*, 2 à la classe *suivant* et 0 à la classe *autre*. Alors $f(précédent|T) = 0,5$, $f(courant|T) = 0,3$, $f(suivant|T) = 0,2$ et $f(autre|T) = 0$. À partir de ces valeurs et de l'équation 4.1, nous déduisons que le critère d'impureté de Gini $i(T)$ est égal à $0,62 = 1 - (0,5^2 + 0,3^2 + 0,2^2)$.

Si les deux nœuds fils d'un nœud T sont appelés T_{oui} et T_{non}, et le pourcentage d'exemples de T qui seront attribués aux fils T_{oui} et T_{non}, sont notés p_{oui} et p_{non} respectivement, la changement d'impureté ΔI est défini par la formule suivante :

$$\Delta I = i(T) - p_{oui} \times i(T_{oui}) - p_{non} \times i(T_{non}) \tag{4.2}$$

La question associée à un nœud T sera la question qui maximise ΔI.

Ce processus de choix des questions est répété pour chaque nouveau nœud fils d'un nœud T donné. La construction de l'arbre s'arrête lorsqu'une des conditions d'arrêt décrites ci-dessous est satisfaite.

Conditions d'arrêt

Le processus de construction de l'arbre s'arrête lorsque l'impureté d'un nœud est sous un certain seuil ou lorsque le nombre d'exemples restant dans un nœud est égal à une limite fixée au préalable.

Dans notre cas, les expériences menées sur le corpus de développement nous ont permis de fixer le seuil minimum du critère de Gini à 0,01 et le nombre minimum d'exemples à 50.

4.2.3 Étiquetage et attributions locales

Une fois l'arbre appris, il est utilisé pour étiqueter chaque nom complet détecté dans la transcription avec les quatre étiquettes possibles : *précédent*, *courant*, *suivant* ou *autre*. Le SCT ne choisit pas une étiquette parmi les quatre possibles, mais attribue une probabilité à chacune des étiquettes, comme le montre la figure 4.4. Une des possibilités retenue dans (Mauclair et al., 2006; Estève et al., 2007) est de ne retenir que l'étiquette qui a la plus forte probabilité. Le problème de cette approche est qu'elle occulte complètement les informations fournies par les probabilités des autres étiquettes.

Dans les travaux présentés dans cette thèse, les probabilités affectées à toutes les étiquettes sont prises en compte. En effet, nous estimons que les quatre probabilités attribuées aux étiquettes sont porteuses d'informations et doivent donc être prises en compte.

Imaginons la répartition des masses de probabilités suivante pour un nom complet : courant $0,31$, précédent $0,28$, suivant $0,30$, autre $0,11$. Décider que le nom complet désigne le locuteur courant car c'est l'étiquette qui a la plus forte probabilité ($p(courant) = 0,31$) occulte complètement le fait que l'arbre est très incertain sur ces étiquettes. En effet, le conflit entre les différentes étiquettes est très élevé et l'arbre affecte des probabilités proches aux étiquettes courant, précédent et suivant.

Chacune de ces probabilités, ainsi que le nom complet concerné, est affectée au locuteur à nommer correspondant (le suivant, le courant ou le précédent). Chaque locuteur se retrouve alors avec un ensemble de noms complets (couplés avec les probabilités issues de l'arbre) qui représente les candidats potentiels. La figure 4.5 donne un exemple d'attribution.

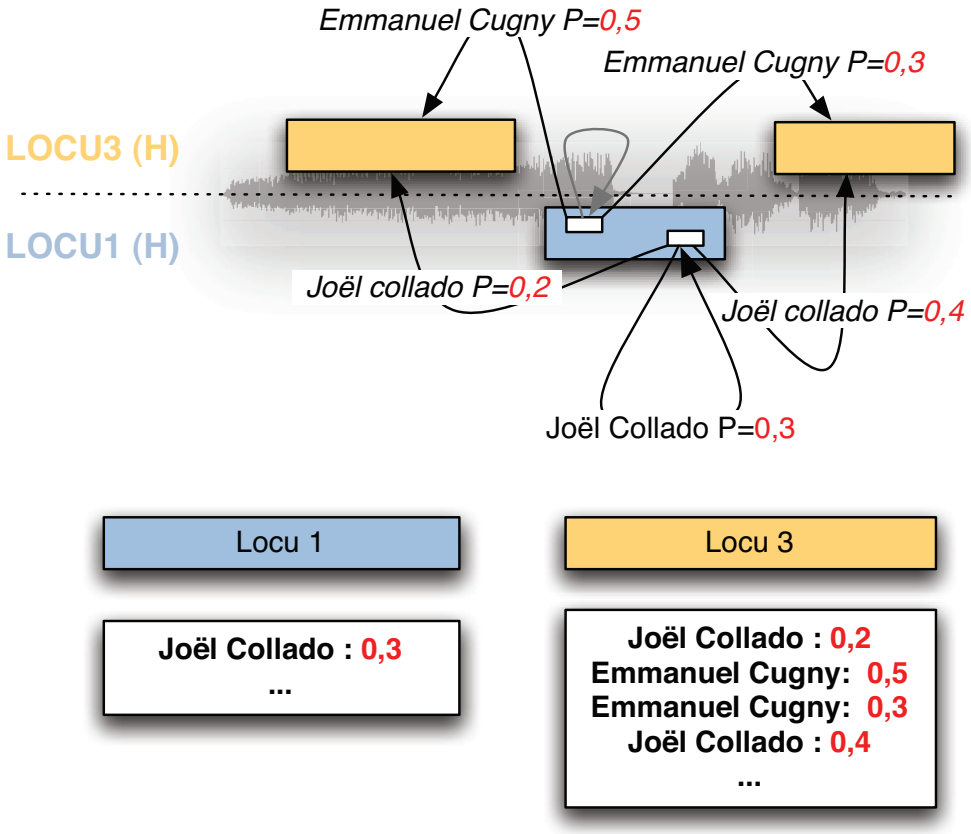

FIGURE 4.5 – *Exemple de propagation des attributions locales aux locuteurs à nommer LOCU1 et LOCU3.*

Les noms complets qui étaient étiquetés *autre* ne sont pas pris en compte, puisqu'il n'est pas possible de les attribuer à un tour de parole du document.

L'enjeu d'un processus d'attribution globale est de décider quel nom complet candidat doit être attribué à un locuteur. Lors de ce processus, différentes incertitudes apparaissent au sein des locuteurs à nommer, puisque plusieurs noms complets peuvent être candidats. L'attribution globale doit prendre en compte toutes ces informations de manière à réaliser l'attribution finale d'un nom complet à un locuteur. Dans cette thèse, la théorie des fonctions de croyances est utilisée pour prendre en compte ces informations.

4.3 Attribution globale : processus de décision et fonctions de croyance pour l'INL

Une des principales difficultés de la phase d'attribution globale (cf. 3.4.3) est de résoudre les conflits qui apparaissent à la suite de la phase d'attribution locale : lorsqu'il y a plusieurs candidats possibles pour un même locuteur. Les règles apprises par l'arbre de classification peuvent ne pas être assez discrimantes (manque d'exemples d'apprentissage, contexte lexical d'un nom complet imprécis, etc.) ce qui l'amènera à commettre des erreurs. Ces erreurs sont à l'origine des conflits qui apparaissent au sein d'un locuteur lorsque plusieurs noms complets sont candidats.

Il arrive que l'arbre de décision se trompe et désigne un locuteur comme étant, par exemple, le locuteur suivant (en lui affectant une forte probabilité), alors qu'il ne l'est pas. La théorie des fonctions de croyance va permettre, dans un même formalisme, de modéliser ces connaissances et ces incertitudes de manière à les utiliser lors de la phase d'attribution globale.

4.3.1 Formalisme et notations

Tout d'abord, il convient d'expliciter les différentes notations qui seront utilisées par la suite.

Soit $\mathcal{E} = \{e_1, \ldots, e_I\}$ correspondant à l'ensemble des noms complets candidats pour nommer un locuteur. L'ensemble $\mathcal{O} = \{o_1, \ldots, o_J\}$ correspond aux occurrences successives des noms complets détectés dans les transcriptions, $\mathcal{T} = \{t_1, \ldots, t_K\}$ désigne l'ensemble des tours de parole dans l'ordre chronologique, et $\mathcal{C} = \{c_1, \ldots, c_L\}$ l'ensemble des locuteurs à nommer. Le but est donc d'attribuer un nom complet issu de $\mathcal{E}$ aux locuteurs de $\mathcal{C}$. Chaque locuteur c_l peut intervenir une ou plusieurs fois dans un enregistrement, ce qui correspond donc à un ou plusieurs tours de parole : $c_l = \{t \in \mathcal{T} \mid t \text{ est un tour de parole de } c_l\}$. Chaque tour de parole est prononcé par un et un seul locuteur, c'est-à-dire que les tours de parole sont successifs et ne peuvent être simultanés. Pour chaque occurrence d'un nom complet o_j (pour $j = 1, \ldots, J$) détecté dans un tour de parole t_k, on désigne par $P(o_j, t_k)$ la probabilité (obtenue grâce à l'arbre de classification sémantique) que o_j soit le locuteur du tour de parole t_k. Ainsi, $P(o_j, t_{k-1})$ et $P(o_j, t_{k+1})$ représentent la probabilité que o_j soit le locuteur du tour respectivement précédent et suivant celui où il a été détecté. Par hypothèse, la probabilité que o_j soit un autre locuteur est donc :

$$1 - \sum_{r \in \{-1,0,1\}} P(o_j, t_{k+r})$$

4.3.2 Fonctions de croyance

Dans cette thèse, le modèle des croyances transférables proposé par (Smets et Kennes, 1994) est adopté. Le but de ce modèle est de déterminer la croyance concernant différentes propositions, à partir d'un ensemble d'informations disponibles. Soit $\mathcal{E}$ un ensemble fini, appelé cadre de discernement de l'expérience. La représentation de l'incertitude se fait grâce à une fonction de croyance, définie comme une fonction m de $2^{\mathcal{E}}$ dans $[0,1]$ telle que :

$$\sum_{A \subseteq \mathcal{E}} m(A) = 1. \tag{4.3}$$

La quantité $m(A)$ représente la part de croyance allouée exactement à la proposition A. Les sous-ensembles A de $\mathcal{E}$ tels que $m(A) > 0$ sont les *éléments focaux* de m. Une structure de croyance est à support simple si elle est focalisée au maximum sur $\mathcal{E}$ et un seul sous-ensemble A. Elle est alors définie par :

$$\begin{cases} m(A) = w, \quad w \in [0,1] \\ m(\mathcal{E}) = 1 - w \end{cases} \tag{4.4}$$

L'ignorance totale est représentée par la fonction de croyance vide telle que $m(\Omega) = 1$. Si tous les éléments focaux de m sont des singletons, m devient alors une mesure de probabilité (dite masse bayésienne).

La théorie des croyances possède plusieurs outils d'agrégation permettant de combiner des fonctions de croyance définies sur un même cadre de discernement (Smets et Kennes, 1994). En particulier, la combinaison de deux structures m_1 et m_2, définies de façon "indépendante" sur $\mathcal{E}$ utilisant l'opérateur binaire conjonctif non normalisé $\cap$, définit une fonction résultante m' par la formule suivante (Smets et Kennes, 1994) :

$$\forall A \subseteq \mathcal{E},\ m'(A) = \sum_{B \cap C = A} m_1(B) m_2(C). \tag{4.5}$$

On peut alors définir la combinaison de n structures $m_1, \ldots, m_n$ sur $\mathcal{E}$ par : $m = m_1 \cap \ldots \cap m_n$, telle que

$$m(A) = \sum_{A_1 \cap \ldots \cap A_n = A} \prod_{i=1}^{n} m_i(A_i) \quad \forall A \subseteq \Omega. \tag{4.6}$$

Si on obtient une fonction de croyance m non normalisée (i. e. telle que $m(\emptyset) \neq 0$), on peut la convertir par la procédure de normalisation de Dempster

(Shafer, 1976) en une structure normalisée m^* en répartissant proportionnellement la masse de l'ensemble vide sur les autres éléments focaux :

$$m^*(A) = \begin{cases} \dfrac{m(A)}{1 - m(\emptyset)} & \text{si } A \neq \emptyset \\ 0 & \text{si } A = \emptyset. \end{cases} \tag{4.7}$$

Une fonction de croyance m décrit l'état d'une croyance sur un phénomène. Une fois qu'une structure m est définie, il est possible de la transformer en distribution de probabilité, en particulier pour des aspects décisionnels. Une de ces distributions, appelée probabilité *pignistique*, consiste à répartir équitablement la masse d'un sous-ensemble de Ω entre ses éléments. Cette distribution, notée P_m, est définie pour tout $\omega \in \Omega$ par (Smets et Kennes, 1994) :

$$P_m(\{\omega\}) = \sum_{A \subset \Omega} \frac{m^*(A)}{|A|} \delta_A(\omega), \tag{4.8}$$

où $|A|$ est le cardinal de A, $\delta_A(\omega) = 1$ si $\omega \in A$ et $\delta_A(\omega) = 0$ si $\omega \notin A$.

4.3.3 Définition des masses de croyance

On s'intéresse à un tour t_k prononcé par le locuteur c_l. n_k occurrences de noms complets ont été attribuées à ce tour de parole par l'arbre de classification. Soient n_{k+r}, le nombre d'occurrences provenant d'une attribution de type *locuteur précédent* ($r = -1$) et *locuteur suivant* ($r = 1$). Soient $\{o^k_{j,r}\}$, avec $r = -1, 0, 1$ et $j = 1, \ldots, n_{k+r}$, les occurrences de noms complets détectés dans ces trois tours. Chaque occurrence $o^k_{j,r}$, correspondant à une étiquette e_i, représente une information particulière sur le locuteur potentiel du tour t_k qui peut être décrite par une masse de croyance $m^{jr}_{t_k}$ sur $\mathcal{E}$ à support simple, focalisée sur e_i et $\mathcal{E}$:

$$\begin{cases} m^{jr}_{t_k}(\{e_i\}) = P(o^k_{j,r}, t_{k-r}) \text{ si } o^k_{j,r} = e_i \\ m^{jr}_{t_k}(\mathcal{E}) = 1 - P(o^k_{j,r}, t_{k-r}), \end{cases} \tag{4.9}$$

4.3.4 Combinaison par tour de parole et par locuteur

La première étape de la combinaison consiste à agréger les informations au sein d'un tour de parole donné. La combinaison des $n_{k-1} + n_k + n_{k+1}$ masses ciblées sur le tour t_k et obtenues par l'équation 4.9 peut se faire par la règle conjonctive de Dempster *non normalisée* (cf. équations 4.5- 4.6), afin d'assurer

l'associativité et la commutativité de la combinaison : on obtient une masse de croyance m_{t_k} sur l'identité du locuteur du tour t_k, définie par

$$m_{t_k} = \bigcap_{r=-1}^{1} \bigcap_{j=1}^{n_{k+r}} m_{t_k}^{jr} \tag{4.10}$$

et la masse de croyance normalisée $m_{t_k}^*$ correspondante (cf. équation 4.7).

La deuxième étape de la combinaison consiste à agréger les résultats obtenus par tour de parole pour l'ensemble de l'enregistrement. Plusieurs possibilités sont envisageables *a priori* pour combiner ces informations. Il est possible de procéder de façon analogue à celle de l'équation 5.1. Pour l'assignation d'un nom complet e_i à un locuteur c_l donné, nous pourrions calculer un "score" pour chaque e_i, dénoté $sm_l(e_i)$, comme la somme des masses de croyance concernant les tours de parole du locuteur c_l :

$$sm_l(e_i) = \sum_{t_k \in c_l} m_{t_k}^*(\{e_i\}). \tag{4.11}$$

Les scores obtenus ne sont donc pas nécessairement normalisés (i.e. ≤ 1), ce qui rendra là encore difficile l'interprétation des résultats lors du processus de décision. Nous préférons rester dans le cadre de la théorie des croyances : une manière assez naturelle consiste en effet à continuer de combiner toutes les masses de croyance relatives au même locuteur c_l par la même règle conjonctive de Dempster et donc à fusionner toutes les masses de croyance relatives aux différents tours de parole t_k de ce locuteur. On obtient alors une masse de croyance globale M_l sur l'identité du locuteur c_l de l'enregistrement, définie par

$$M_l = \bigcap_{t_k \in c_l} m_{t_k} \tag{4.12}$$

et la masse de croyance normalisée M_l^* correspondante.

4.3.5 Processus de décision

Grâce au processus décrit en section 4.3.4, chaque locuteur C_l dispose d'une liste de noms complets candidats e_i associés à leurs probabilités pignistiques $P_{M_l}(e_i)$. Le processus de décision doit permettre d'assigner les noms complets aux locuteurs de la manière la plus optimale possible.

Soit $f : \mathcal{C} \rightarrow \mathcal{E}$ la fonction d'assignation des noms complets aux locuteurs. Les locuteurs étant censés être tous différents, une caractéristique importante que l'on voudrait imposer pour f est l'injectivité : $f(c_l) = f(c'_l) \Rightarrow c_l = c'_l$. Chaque nom complet ne peut être attribué à plus d'un locuteur. Bien sûr, certains noms complets peuvent rester non attribués et inversement, certains locuteurs peuvent rester sans étiquette.

Ce type d'assignation est un problème classique en recherche opérationnelle (Burkard et al., 2009). Ce problème est souvent présenté de la manière suivante : une entreprise dispose de n employés pour réaliser n tâches. Pour chaque tâche, sa réalisation par un employé donné représente un certain coût, le but de l'entreprise étant de minimiser le coût de réalisation des différentes tâches. Il lui faut donc choisir les couples employé/tâche qui minimise le coût global de réalisation.

Dans notre cas, nous disposons de n locuteurs à nommer, et de m noms complets candidats potentiels. À chaque couple locuteur/nom complet correspond une probabilité pignistique, qui représente le bénéfice d'attribuer ce nom complet à ce locuteur. La probabilité des noms complets n'étant pas candidats pour un locuteur donné est fixée à 0. Le tableau 4.2 donne un exemple de matrice obtenue pour un enregistrement avec 3 locuteurs et 3 noms complets possibles.

TABLE 4.2 – *Exemple de matrice des coûts utilisée pour le processus de décision (algorithme de Kuhn-Munkres)*

Locuteur C_l	Probabilités $P_{M_l}(e_i)$		
	Abdelamjid Kadouri	*Mohammed Bellaouchi*	*Aziz Al-Bouazaoui*
c_1	**0,06**	0,34	0,00
c_2	0,01	0,17	**0,05**
c_3	0.00	**0,49**	0,02

Un des algorithmes parmi les plus utilisés pour résoudre ce problème est « l'algorithme de Kuhn-Munkres (Munkres, 1957) », aussi appelé « algorithme Hongrois ». Il consiste à trouver les associations optimales qui minimisent un certain coût. La fonction de coût utilisée est la suivante :

$$C(e_1, ..., e_I) = \sum_{l=1}^{L} P_{M_l}(e_i) \tag{4.13}$$

Nous utilisons ici une variante de l'algorithme original qui permet d'utiliser des matrices rectangulaires et de maximiser la fonction de coût (Nedas, 2005). Puisque les affectations locuteur/nom complet sont représentées par des probabilités, il est plus intéressant de maximiser le coût de la fonction que de la minimiser.

La matrice des coûts présentée dans le tableau 4.2 donne un exemple de problème d'attribution. Dans cet exemple, le document comporte 3 locuteurs c_1, c_2, c_3 et 3 noms complets sont candidats. Dans ce cas, le nom complet *Abdelamjid Kadouri* sera attribué à c_1, *Aziz Al-Bouazaoui* à c_2 et *Mohammed Bellaouchi* à c_3. Ce sont en effet les associations qui maximisent la fonction de coût : $0,06 + 0,05 + 0,49 = 0,61$.

4.3.6 Prise en compte du genre

Les travaux présentés dans cette thèse prennent en compte le genre des locuteurs pour filtrer les probabilités $P(o_j, t_k)$ données par l'arbre de classification sémantique. Pour chaque locuteur de l'enregistrement, cette caractéristique est disponible puisqu'elle est déterminée de manière automatique lors des phases de segmentation et de classification (avec un taux d'erreur inférieur à 5 % sur des données ESTER 1 phase II). De plus, le genre des noms complets extraits de la transcription peuvent être déterminés grâce au prénom, lorsque celui-ci le permet. La comparaison de ces deux informations, obtenues de deux manières différentes, nous permet d'affiner le processus de décision. En cas d'incohérence, le couple prénom et patronyme n'est pas retenu et il est supprimé de la liste des candidats potentiels. Cependant, pour le cas des prénoms ambigus comme « Dominique », l'entité nommée sera conservée.

Une base de données (extraite du Web) composée d'environ 20 000 prénoms est utilisée pour obtenir le genre de chaque prénom. À chaque prénom est associé le nombre de fois où il a été attribué au genre féminin et au genre masculin depuis 1900 en France. Cette base de données ne semble pas exempte d'erreur : par exemple le prénom « Vincent » apparaît 227180 fois comme prénom masculin et 373 fois comme prénom féminin. Le genre retenu correspondra au genre majoritaire. Si ce dernier a une fréquence inférieure à 75 %, alors le genre est considéré comme indéterminé.

Au niveau de la définition des masses de croyance présentée en section 4.3.3, le filtre suivant est appliqué : si le genre de l'occurrence o_j (et donc du nom complet e_i correspondant) et celui du tour de parole (et donc de la classe c_l à laquelle il appartient) sont différents, les scores $P(o_j, t_r)$ sont ignorés.

4.4 Évaluation du système proposé

4.4.1 Description des corpus

L'évaluation du système proposé est réalisée à partir d'émissions radiophoniques en français de la campagne ESTER 1 phase II (Galliano et al., 2006; Gravier et al., 2004) et du projet ANR EPAC (Estève et al., 2010a). Les émissions d'ESTER 1 contiennent essentiellement de la parole lue ou préparée, et peu de parole spontanée : 15 % du corpus correspond à des interventions de personnes parlant au téléphone. En revanche les émissions du corpus EPAC contiennent plus de parole spontanée : le projet EPAC se focalisait sur le traitement de la parole conversationnelle.

Les émissions proviennent de cinq radios françaises et de Radio Télévision Marocaine et durent de 10 à 60 min. Elles sont réparties en trois corpus utilisés pour l'apprentissage de l'arbre de classification, le développement et l'évaluation du système. Le corpus d'apprentissage est celui de la campagne ESTER 1, et le corpus de développement utilisé correspond au corpus test de la campagne ESTER 1. Le corpus de test est constitué de données issues d'EPAC.

Le corpus d'apprentissage contient 76 heures de données (7 420 tours de parole) dans lesquels 12 457 noms complets sont détectés. 755 locuteurs différents interviennent dans ce corpus et 40 n'ont pu être nommés. Le corpus de développement contient 10 heures (1 082 tours de parole) dans lesquels 1 660 noms complets ont été détectés. 211 locuteurs différents interviennent dans ce corpus et 24 n'ont pu être nommés. Le corpus de test contient 10 heures de données (1 194 tours de parole) dans lesquels 1102 noms complets sont détectés. 118 locuteurs différents interviennent dans ce corpus et 2 n'ont pu être nommé.

La majorité locuteurs du corpus de test ont la particularité d'être nommés. Lorsque seul le prénom ou le nom étaient disponibles ils ont été utilisés pour nommer le locuteur. De plus, des connaissances supplémentaires comme la grille des programmes ont été utilisées lors de la réalisation de la transcription manuelle. Le tableau 4.3 résume la composition de ces différents corpus.

Corpus	D	I	N	Nombre d'occurrences
Apprentissage	70,68 %	10,20 %	19,12 %	2357
Développement	86,77 %	5,89 %	7,34 %	321
Évaluation	89,23 %	10,58 %	0,19 %	141

TABLE 4.3 – *% de données Disponibles (D), Indisponibles (I) et Non-nommées (N) dans les différents corpus. Cela fixe la limite haute des performances atteignables, puisque les noms complets I ne peuvent être récupérés à partir de la transcription.*

	apprentissage	***développement***	***test***
Tour précédent	6,9 % (250)	8,5 % (53)	11,5 % (46)
Tour courant	5,6 % (206)	2,0 % (12)	1,5 % (6)
Tour suivant	34,9 % (1 275)	32,5 % (200)	32,9 % (131)
Autre	52,6 % (1 922)	57,0 % (351)	54,1 % (216)
Total	100 % (3 653)	100 % (616)	100 % (399)

TABLE 4.4 – *Répartition des étiquettes sur les différents corpus, statistiques sur les noms complets (fréquence et effectif).*

Les transcriptions fournies avec les corpus ont été créées pour l'évaluation des tâches de segmentation et de classification en locuteurs, ainsi que pour la tâche de transcription.

Le tableau 4.4 montre la répartition *a priori* des quatre étiquettes calculées pour les différents corpus de développement. L'étiquette « *autre* » est la plus fréquente et représente plus de 50 % des cas ; vient ensuite l'étiquette « *tour suivant* » avec plus de 30 % des, tandis que les deux dernières étiquettes « *tour précédent* » et « *tour courant* » sont les moins fréquentes : de 2 % à 10 % en fonction du corpus.

4.4.2 Métriques utilisées

Les métriques utilisées dans les expériences (Rappel et Précision) sont celles définies dans la section 3.2. Comme il a été proposé dans (Tranter, 2006), ces valeurs peuvent être complétées par un taux d'erreur *Err* global également calculé à partir de ces 5 erreurs. Ce taux s'inspire du calcul du *WER* utilisé pour l'évaluation de la transcription. Il a l'avantage de mesurer la qualité des résultats du système d'identification nommée en une seule valeur, facilitant les comparaisons entre les systèmes par rapport aux mesures de précision et de rappel.

$$Err = \frac{S + I + D}{S + I + D + C_2 + C_1} \ ; \tag{4.14}$$

Les erreurs peuvent être calculées en terme de durée ou en terme de nombre de locuteurs. Pour une évaluation en durée, dans le cas où un locuteur parlant 90% du temps est correctement nommé et que les six autres locuteurs parlant seulement 10% du temps ne le sont pas, le système présentera un taux d'erreur de 10%.

Pour une évaluation en terme de nombre de locuteurs, dans le même cas de figure, le système aura un taux d'erreur de 87,5%.

D'un point de vue applicatif, la métrique exprimée en durée est préférable si les locuteurs considérés comme importants correspondent aux locuteurs s'exprimant beaucoup. En revanche, si l'application cherche à nommer le plus possible de locuteurs, il est plus intéressant d'évaluer les performances en terme de nombre de locuteurs.

4.4.3 Système de transcription automatique du LIUM

Les expériences réalisées sur des transcriptions automatiques ont été effectuées grâce au système de transcription automatique de la parole du LIUM. Ce système est constitué de deux modules : un module de segmentation et de classification en locuteur et un module de transcription.

Tout d'abord, la tâche de segmentation et de classification en locuteur consiste à annoter des régions du signal audio (cf. section 2.3.6). Ces annotations peuvent être de plusieurs types : des étiquettes représentant les différents locuteurs, le genre des locuteurs, le type de canal (radio, studio) ou encore le type d'environnement sonore comme le bruit, la musique, ...Dans la tâche de segmentation, les étiquettes représentant les locuteurs sont anonymes. Le système utilisé au LIUM (LIUM_SpkDiarization) utilise une segmentation basée sur GLR (Generalized Likelihood Ratio) et BIC (Siegler et al., 1997) (Bayesian Information Criterion). Il est entièrement décrit dans (Meignier et Merlin, 2010). Ce système a été classé premier lors de la campagne d'évaluation ESTER 2 avec un DER (*Diarization Error Rate*) de 10 %.

Le module de transcription a pour but de transcrire le signal audio en mots. Celui du LIUM (Estève et al., 2010b) est basé sur le décodeur CMU Sphinx 3.7. Il utilise un vocabulaire de plus de 65 000 mots, ainsi qu'un modèle de langage 3-gramme. Des ajouts au décodeur de base ont permis d'utiliser SAT (Speaker Adaptative Training) et des modèles 4-gramme pour réévaluer les graphes de mots. Le système est décrit de manière exhaustive dans (Deléglise et al., 2005) et (Estève et al., 2010b). Ce système s'est classé 2 ème lors de la campagne d'évaluation ESTER 2 avec un taux d'erreur mot de 17,83 %.

4.4.4 Résultats

Comme dans les systèmes présentés dans le chapitre 3, une liste des locuteurs potentiels (ainsi que leur genre) est utilisée. Cela permet notamment de comparer le système à celui du LIUM présenté dans la section 3.7. Les résultats sans cette liste de locuteur seront donnés et discutés dans la section 5.3.

Transcriptions manuelles

Le tableau 4.5 présente les résultats du système sur les corpus de développement et de test, en utilisant des transcriptions manuelles. La liste de locuteurs utilisée pour le corpus de développement correspond aux locuteurs du corpus d'apprentissage, de développement et de test de la campagne ESTER 1 (soit 1016 locuteurs). La liste de locuteurs utilisée pour le test est celle du corpus de test EPAC (soit 101 locuteurs). Les résultats en durée et en nombre de locuteurs sont donnés.

	En durée			En nb de Locuteur
	R	**P**	**ErrDur**	**ErrLoc**
Corpus de Développement	87,85%	95,24%	11,44%	12,43 %
Corpus de Test	77,10%	92,97%	22,85%	28,23%

TABLE 4.5 – *Résultats de* Milesin *sur des transcriptions entièrement manuelles en utilisant une liste des locuteurs du corpus*

Ces résultats sont comparables à ceux donnés pour le système du LIUM dans la figure 3.10 : à 95 % de précision, le meilleur système obtenait un rappel d'environ 40 %. Le corpus utilisé dans 3.10 correspond au corpus de développement du système présenté. Pour la même précision, le rappel a plus que doublé pour atteindre 86,00 %. Sur le corpus de test, à 93 % de précision, le rappel est d'environ 77 %.

Les résultats sur le corpus de test s'expliquent par la composition du corpus décrit dans le tableau 4.3. En effet, le taux de locuteurs Indisponibles dans la transcription est deux fois supérieur à celui du corpus de développement.

Transcriptions automatiques : résultats préliminaires

L'utilisation de systèmes d'INL prend tout son sens lorsqu'elle peut être réalisée sur des transcriptions produites par des systèmes automatiques. Même si quelques expériences préliminaires ont été réalisées par Tranter et le LIUM, l'utilisation de systèmes d'INL sur des transcriptions reste problématique.

Le tableau 4.6 présente les résultats de *Milesin* sur des transcriptions entièrement automatiques en utilisant une liste des locuteurs du corpus. Seuls les résultats en durée sont présentés, puisque la segmentation et la classification ne sont pas identiques entre la référence et l'hypothèse proposée. Les systèmes de segmentation / classification et de transcription automatiques utilisés sont ceux présentés dans la section 4.4.3. Sur le corpus de développement, le DER est de

11,5 % et le WER de 20,5 %. Sur le corpus de test, le DER est de 5,71 % et le WER de 18,8 %.

	En durée		
	R	**P**	**ErrDur**
Corpus de Développement	25,15 %	55,65 %	69,43 %
Corpus de Test	38,77 %	70,87 %	61,23 %

TABLE *4.6* – *Résultats de* Milesin *sur des transcriptions entièrement automatiques en utilisant une liste des locuteurs du corpus*

Le taux d'erreur (en durée) obtenu varie de 69,43 % à 61,23 % en fonction du corpus utilisé. Lorsque l'on utilise *Milesin* avec des transcriptions entièrement automatiques, les erreurs des différents systèmes se cumulent. Aucune adaptation de *Milesin* pour des transcriptions automatiques n'a été effectuée dans ces expériences.

L'influence des différentes composantes d'une transcription automatique sur les résultats des systèmes d'INL est discutée plus en détails dans la section 5.4.2.

4.5 Bilan

Milesin est un système d'INL utilisant un détecteur d'entités nommées pour étiqueter les noms complets et les classes sémantiques, un arbre de classification sémantique pour réaliser les attributions locales et la théorie des fonctions de croyances couplée à l'algorithme de Kuhn-Munkres pour le processus d'attribution globale.

L'utilisation des fonctions de croyance pour combiner les croyances et incertitudes de la phase d'attribution locale permet de disposer d'un solide modèle mathématique pour calculer les probabilités d'affectation d'un nom complet à un locuteur. Grâce à l'utilisation de l'algorithme de Kuhn-Munkres, l'assignation des noms complets aux locuteurs est optimale. Ces différentes améliorations donnent de bons résultats, puisque sur le même corpus transcrit manuellement, *Milesin* obtient un rappel de 86,22 % à 95 % de précision, là où le système antérieur à cette thèse obtenait un rappel d'environ 40 %. En revanche, les performances sur les transcriptions automatiques sont encore mauvaises.

Afin de se comparer aux travaux précédemment effectués, une liste de locuteurs potentiels a été utilisée. Cette liste constitue une connaissance a priori qui peut se justifier si les intervenants des enregistrements sont déjà connus et qu'il s'agit uniquement de les nommer dans les fichiers. En revanche, si l'on veut

disposer d'un système sans connaissances a priori sur les documents à traiter, il faut pouvoir se passer de cette liste de locuteurs. De plus, il serait intéressant de tester *Milesin* sur des transcriptions automatiques afin d'analyser son comportement et d'apporter de nouvelles pistes de travail sur ce type de transcriptions.

Ces différentes pistes sont décrites en détail dans le chapitre suivant, ainsi que les avancées et limites du travail effectué dans cette thèse.

Chapitre 5

Milesin : avancées et limites

Sommaire

Un nombre significatif d'avancées ont été réalisées depuis le système de base du LIUM (cf. section 3.7), comme l'utilisation de LIA_NE comme détecteur d'entités nommées ou la mise en place des fonctions de croyances pour combiner les informations issues de l'arbre de classification sémantique. En revanche, l'utilisation de transcriptions automatiques reste encore problématique.

Les principales étapes et réflexions qui ont conduit au développement de *Milesin*, et notamment une analyse des erreurs, sont tout d'abord présentées.

5.1 Analyse préliminaire

Les travaux de cette thèse ont été réalisés à partir d'un système déjà existant. Ce système disposait d'un détecteur d'entités nommées sommaire, il a donc tout d'abord été décidé de le remplacer par Nemesis, le système de détection d'entités nommées du LINA. Une analyse des erreurs a ensuite été conduite de manière à comprendre les erreurs réalisées par le système.

5.1.1 Nemesis, un outil prévu pour le TAL

Nemesis (Fourour, 2004) est un système d'identification et de catégorisation d'entités nommées pour le français. Ses spécifications ont été élaborées à la suite d'une étude en corpus et s'appuient sur des critères graphiques et référentiels. Ces derniers ont permis de construire une typologie des entités la plus fine et la plus exhaustive possible, fondée sur celle de Grass (Grass, 2000). L'architecture logicielle de Nemesis se compose principalement de 4 modules (prétraitement lexical, première reconnaissance, apprentissage, seconde reconnaissance) qui effectuent un traitement immédiat des données à partir de textes bruts. L'identification des entités nommées est réalisée en analysant leur structure interne et leurs contextes gauche et droit immédiats à l'aide de lexiques de mots déclencheurs, ainsi que de règles de réécriture. Leur catégorisation s'appuie quant à elle sur la typologie construite précédemment. L'outil atteint environ 90% de précision et 80% de rappel sur des textes écrits en langage naturel.

Nemesis a été développé par l'équipe TALN (Traitement Automatique des Langues Naturelles) du LINA (Laboratoire d'Informatique de Nantes Atlantique). À noter qu'aucune modification ou adaptation n'a été réalisée pour utiliser Nemesis sur des transcriptions de journaux radiophoniques.

5.1.2 Analyse des erreurs

Le système du LIUM, présenté dans la section 3.7, s'appuie sur un arbre de classification sémantique pour réaliser les attributions locales. Le nom complet dont l'étiquette (« précédent », « courant », « suivant ») a la probabilité maximale est propagé au tour de parole puis au locuteur correspondant. Les probabilités sont ainsi additionnées au sein du locuteur, et le nom complet ayant le score maximum pour un locuteur lui est affecté.

Ce système a été utilisé comme base de ces travaux, seul le détecteur d'entités nommées a été remplacé par Nemesis. Afin de comprendre le comportement du système, nous avons choisi de l'étudier avec des transcriptions et des segmentations / classifications manuelles. Le corpus utilisé correspond au corpus de développement de *Milesin*. Sur ce corpus, il obtient un rappel de 65 % à 91 % de précision. Nous chercherons donc à comprendre pourquoi ce système a un rappel faible (65 %) bien que les conditions soient optimales (transcriptions et segmentation / classification manuelles).

Méthode

Chaque entité nommée de type personne détectée a été vérifiée, et les scores générés par l'arbre pour ces entités nommées ont été étudiés. Cette étude a permis de mettre en évidence deux grandes catégories d'erreurs : les erreurs de détection d'entités nommées et les erreurs de décision de l'arbre. D'autres types d'erreurs ont aussi été constatés mais elles sont moins significatives.

Erreurs de détection d'entités nommées

La première catégorie d'erreur est due à des problèmes rencontrés lors de la détection d'entités nommées. En effet, les entités nommées de type personne sont primordiales pour un système d'INL. Ces entités doivent contenir le prénom et la patronyme de la personne concernée, et uniquement ceux-ci. Un mauvais étiquetage de ces entités nommées a une influence directe sur les performances du système en terme de rappel :

(N1) Un locuteur n'est pas correctement étiqueté pour être utilisé par le système d'identification nommée quand : il n'est pas détecté, seul son prénom est étiqueté ou quand ses nom et prénom sont étiquetés avec un ou plusieurs autres mots en plus (comme la fonction de la personne). Lorsque le nom du locuteur détecté ne correspond pas exactement au nom du locuteur de la référence, il est alors comptabilisé comme une erreur de suppression et par conséquent le rappel chute. Par exemple, l'entité nommée *Joël Collado de Météo France* est étiquetée

comme une personne et ne correspond pas exactement au couple prénom / nom attendu lors de l'identification.

(N2) Un ensemble de mots ne contenant pas de nom et prénom est détecté comme une personne. Cette détection fait chuter le rappel en attribuant cette *fausse identité* à un locuteur. Par exemple, Nemesis étiquette l'entité nommée *Président du Fetia Api* comme personne qui est ensuite attribuée à un locuteur. *Président du Fetia Api* est bien une personne, mais cette entité nommée n'est pas un locuteur au sens de l'identification. Il est identifié par sa fonction au lieu de son prénom et nom.

Erreurs de décision de l'arbre

La seconde catégorie d'erreurs provient des erreurs d'étiquetage commises par l'arbre. Elles ont une incidence à la fois sur le rappel et sur la précision du système :

(A1) Les erreurs affectant le rappel proviennent majoritairement des noms complets dont l'étiquette avec la probabilité maximale était l'étiquette *autre*, alors qu'ils correspondent à une des 3 autres étiquettes.
"*Valérie Crova* dit : (...) qui a toujours entretenu selon **Jean Christophe Bouisson** des rapports tendus avec le gouvernement. *Jean Christophe Bouisson* dit : On ne peut pas penser (...)"
Dans cet exemple *Jean Christophe Bouisson* est étiqueté comme *autre* alors qu'il est le *suivant*.

(A2) Les erreurs affectant principalement la précision sont dues à des locuteurs qui sont mal étiquetés avec les étiquettes *suivant* et *précédent*. Dans ce cas, l'étiquette avec la probabilité maximale ne désigne pas le bon tour de parole. Généralement, c'est l'étiquette *suivant* qui a été attribuée au lieu de l'étiquette *autre*. Par exemple, l'annonce d'interview correspond à une erreur typique :
"**Jean Michel Hibon** au micro de **Jérôme Susini.**"
On s'attend à ce que *Jean Michel Hibon* parle ensuite (ce qui est le cas), mais le système détecte *Jérôme Susini* comme le locuteur du prochain tour de parole car son score est plus élevé.

Erreurs	Fréquence	Pourcentage Total
N1	12	14%
N2	4	4,6%
A1	32	37,2%
A2	30	34,9%
E	8	9,3%
Total	86	100%

FIGURE 5.1 – *Répartition des erreurs sur le corpus de test.*
N1, N2 : erreurs issues de la détection d'entités nommées.
A1, A2 : erreurs issues de l'arbre de classification.
E : autres erreurs.

Autres erreurs

D'autres erreurs plus indépendantes du système ont été relevées :
(E) Certaines personnes ne sont citées que partiellement (seul leur prénom est cité dans la transcription), d'autres ont été mal orthographiées par le transcripteur ou encore certaines ne sont pas du tout citées ou annoncées dans la transcription. Sur les 11 heures du corpus, ce dernier cas ne concerne que 3 personnes. Cette constatation permet de valider l'hypothèse de départ : les noms des locuteurs sont présents dans les transcriptions d'émissions radiophoniques.

Répartition des erreurs

Le tableau 5.1 montre la répartition des différents types d'erreurs sur le corpus de test. Les erreurs provenant de l'arbre de classification (A1 et A2) sont clairement dominantes avec plus de 72% des erreurs totales alors que les erreurs dues à Nemesis (N1 et N2) représentent un peu moins de 19% des erreurs. Pour l'arbre de classification, les erreurs A1 et A2 ont le même ordre de grandeur (environ 36% en moyenne). En revanche, en ce qui concerne Nemesis les erreurs de type N1 sont beaucoup plus fréquentes que celles de type N2 (3 fois plus).

Cas de l'anaphore pronominale

Avant cette étude, nous pensions que la résolution d'anaphore pronominale à l'intérieur d'un tour de parole pouvait conduire à améliorer les prises de décisions. Lors de l'analyse des erreurs, nous n'avons relevé que deux cas où cette prise en compte aurait été bénéfique. L'un des deux cas correspond à la situation suivante :
" *Fabrice Drouelle* : écoutez ce qu'en pense [Jérôme Savary] (...) **il** le dit à [Christine Siméone]."
Comme le pronom personnel *il* est en rapport avec Jérôme Savary, l'arbre aurait pu l'étiqueter comme *suivant*.

Bilan

L'analyse des erreurs conduite ici a permis de mettre en évidence les faiblesses de l'arbre de classification sémantique et du processus de décision associé : lorsque l'on choisit de prendre une décision en se basant sur l'étiquette qui obtient la probabilité maximale, l'arbre de décision est responsable de plus de 70 % des erreurs du système.

Nemesis quant à lui, représente environ 20 % des erreurs commises par le système. Ces erreurs mettent en évidence l'importance de ne détecter que le couple prénom / nom pour les entités de type personne. Pour les systèmes d'INL, dès qu'un autre mot vient s'ajouter au couple prénom / nom (comme la fonction de la personne par exemple), l'identification sera considérée comme erronée.

En complément, nous donnons une comparaison de LIA_NE et de Nemesis dans le tableau 5.1. *Milesin* est utilisé pour ces expériences sur le corpus de développement. Les meilleurs résultats concernant l'utilisation de LIA_NE pour l'INL s'expliquent par une concision plus importante de ses entités nommées ainsi que par les bonnes performances obtenues lors de la campagne d'évaluation ESTER 2. Nemesis a par la suite été remplacé par LIA_NE dans *Milesin*.

	En durée			En nb de Locuteur
	R	**P**	**ErrDur**	**ErrLoc**
LIA_NE	86,00%	94,07%	13,17%	13,69%
Nemesis	80,05%	94,68%	18,69%	19,11%

TABLE 5.1 – *Résultats de* Milesin *sur des transcriptions entièrement manuelles en utilisant Nemesis ou LIA_NE comme détecteur d'entités nommées*

5.2 Processus de décision : variantes

Lors des travaux de cette thèse, le processus de décision a beaucoup évolué. Nous commençons par rappeler et formaliser le processus de décision utilisé dans les précédents travaux du LIUM, pour ensuite présenter la solution intermédiaire développée pendant cette thèse. Nous finissons par exposer le cheminement qui nous a conduit à utiliser la théorie des fonctions de croyance.

5.2.1 Utilisation d'un maximum

Lors de l'utilisation de l'arbre de classification sémantique sur les entités nommées de type personne, une distribution de probabilités sur les 4 classes possibles (*précédent*, *courant*, *suivant* et *autre*) est attribuée à l'entité nommée concernée. Dans les précédents travaux du LIUM, seule le nom complet avec l'étiquette ayant la probabilité maximale est propagé. Les autres probabilités ne sont pas prises en comptes.

Lors de la propagation des attributions locales de l'arbre aux tours de paroles puis aux locuteurs, une somme des probabilités $P(o_j, t_r)$ (cf. 4.3.1 pour la signification des notations) est effectuée. Ce score concernant un locuteur e_i pour un locuteur donné c_l s'exprime :

$$s_l(e_i) = \sum_{\left\{(o_j,t_r)|o_j=e_i,\, t_r\in c_l\right\}} P(o_j, t_r) \tag{5.1}$$

Ensuite, le nom complet e_i ayant le score maximum pour un locuteur donné c_l lui est attribué. Dans ces travaux, la fonction d'assignation des noms complets aux locuteurs $f(c_l) = f(c'_l) \Rightarrow c_l = c'_l$ n'est pas injective : il est donc possible d'affecter deux noms complets identiques à des locuteurs différents.

Soit $\mathcal{D} = \{c_l \in \mathcal{C} |\ \forall e_i \in \mathcal{E}, s_l(e_i) = 0\}$, l'ensemble des locuteurs n'ayant pas de candidats potentiels. La règle $\mathbf{R_1}$ permettant d'affecter le nom complet e_i ayant le score maximal $s_l(e_i)$ est définie par :

$$\begin{gathered} \forall c_l \in \mathcal{C} \setminus \mathcal{D}, e_i^* = \arg\max_{e_i \in \mathcal{E}} \mathrm{s}_l(e_i) \Rightarrow f(c_l) = e_i^* \\ \forall c_l \in \mathcal{D}, f(c_l) = \text{Anonyme} \end{gathered} \tag{5.2}$$

Cette méthode de combinaison a plusieurs inconvénients. Tout d'abord, le fait de réaliser la somme des probabilités issues de l'arbre ne permet plus de travailler dans l'espace probabiliste. En effet, une fois les sommes effectuées, seuls des scores difficilement comparables entre eux sont disponibles.

De plus, la règle $\mathbf{R_1}$ ne prend pas en compte la contribution du score $s_l(e_i)$ pour un locuteur c_l. Prenons l'exemple décrit dans le tableau 5.2. En utilisant la règle $\mathbf{R_1}$, *Jacques Derrida* est attribué au locuteur c_{13} car ce locuteur obtient le score maximum. β_{il} désigne le pourcentage que représente $s_l(e_i^*)$ par rapport à la somme de tous les scores s_l pour le locuteur c_l. Même si *Jacques Derrida* a le score le plus élevé pour c_{13}, il ne représente que 35% des scores totaux. En revanche, au sein du locuteur c_{15}, *Jacques Derrida* représente 80% des scores. Avant d'utiliser les fonctions de croyance, nous avons tout d'abord proposé une méthode permettant de normaliser les scores pour prendre en compte ces différentes aspects.

TABLE 5.2 – *Exemple d'une assignation initiale multiple*

Locuteur	nom complet e_i^*	$s_l(e_i^*)$	β_{il}
c_{13}	Jacques Derrida	**8, 58**	35%
c_{14}	Jacques Derrida	1, 67	56%
c_{15}	Jacques Derrida	4, 94	**80**%

5.2.2 Normalisation des scores

La contribution du score de chaque nom complet pour un locuteur (modélisé par β_{il}) peut être une information intéressante à prendre en compte lors du processus de décision. Mais l'utilisation de β_{il} comme unique critère de décision ne peut être une option. En effet, dans le cas où un mauvais nom complet serait le seul candidat pour une classe donnée, β_{il} vaudrait 100 %. Ce nom complet serait alors systématiquement choisi pour le locuteur concerné, même si $s_l(e_i)$ est très faible pour ce locuteur. Il faut donc prendre en compte conjointement $s_l(e_i)$ et β_{il}. Nous proposons l'utilisation d'une nouvelle règle $\mathbf{R_2}$:

$$\mathrm{SC}_l(e_i) = s_l(e_i)\beta_{il} \tag{5.3}$$

Cette règle permet de normaliser le score de chaque nom complet e_i en fonction du score global du locuteur à nommer. Lorsqu'un nom complet est le seul candidat pour un locuteur, son score reste inchangé. Lorsqu'il y a plusieurs candidats pour un même locuteur, la normalisation permet de prendre en compte la contribution du score par rapport à l'ensemble des scores du locuteur. Les scores en forte concurrence sont alors pénalisés. Cette normalisation est particulièrement efficace lorsque plusieurs noms complets potentiels ont des scores élevés.

Reprenons l'exemple décrit dans le tableau 5.3 en utilisant cette fois ci $SC_l(e_i)$ pour prendre la décision. Le tableau 5.3 décrit ce cas de figure.

TABLE 5.3 – *Exemple d'une assignation initiale multiple prenant en compte le conflit*

Locuteur	nom complet e_i^*	$s_l(e_i^*)$	β_{il}	$SC_l(e_i^*)$
c_{13}	Jacques Derrida	**8,58**	35%	3,00
c_{14}	Jacques Derrida	1,67	56%	0,94
c_{15}	Jacques Derrida	4,94	**80%**	**3,95**

Dans le précédent exemple, *Jacques Derrida* était affecté au locuteur c_{13}. Avec l'utilisation de $SC_l(e_i)$ pour prendre la décision, *Jacques Derrida* est maintenant attribué au locuteur c_{15} : celui où il obtient un score $s_l(e_i^*)$ élevé et pour lequel il y a peu de conflit (β_{il} est égal à 80%).

5.2.3 Expériences et résultats

Le tableau 5.4 donne les résultats sur les corpus de développement et de test (transcriptions enrichies manuelles) en utilisant les différents systèmes de combinaison des informations : simple maximum, normalisation des scores ou fonctions de croyances. Une liste de locuteur et de leur genre est utilisée (cf. section 4.4.4). *Milesin* a été utilisé pour ces expériences, ainsi que les différentes avancées qu'il contient : injectivité de $f(c_l) = f(c'_l) \Rightarrow c_l = c'_l$, prise en compte du genre et utilisation de l'algorithme Hongrois. Seule la manière de combiner les informations de l'arbre de classification au niveau du locuteur a été changée.

	En durée			En nombre	
	R	**P**	**ErrDur**	**ErrLoc**	**Nb. corrects**
Corpus de Développement					
Maximum	75,37 %	97,31 %	22,51 %	32,51 %	164
Normalisation	87,41 %	95,71 %	11,67 %	13,99 %	209
F. de croyance	87,85 %	95,24 %	11,44 %	12,43 %	206
Corpus de Test					
Maximum	77,83 %	100 %	21,10 %	28,30 %	76
Normalisation	81,57 %	100 %	17,54 %	21,70 %	83
F. de croyance	77,10%	92,97%	22,85%	28,23 %	81

TABLE 5.4 – *Différence entre les méthodes de combinaison globales sur des transcriptions enrichies manuelles en utilisant une liste de locuteurs*

Tout d'abord, on observe le même comportement des différentes méthodes sur les deux corpus. La méthode qui présente les meilleurs résultats est la méthode utilisant la normalisation : 17,54 % de taux d'erreur en durée sur le corpus de test (11,67 % pour le corpus de développement) pour 21,70 % de taux

d'erreur en nombre de locuteurs sur ce même corpus (13,99 % sur le corpus de développement). La méthode utilisant les fonctions de croyance est quant à elle légèrement moins performante % : 22,85 % de taux d'erreur en durée sur le corpus de test pour 28,23 % de taux d'erreur en nombre de locuteurs.

Lors de l'analyse de ces résultats, le nombre de locuteurs concernés doit être pris en compte. Par exemple, sur le corpus de développement, la différence de locuteurs correctement nommés entre la méthode utilisant la normalisation et celle utilisant les fonctions de croyance est faible : on passe de 209 locuteurs correctement nommés à 206 (contre 164 pour la méthode du maximum). De manière identique, sur le corpus de test, le nombre de locuteurs correctement nommés passe de 83 à 81 avec l'utilisation des fonctions de croyance (contre 76 pour la méthode du maximum).

Bien que ce ne soit pas la méthode qui obtienne les meilleurs résultats, ce sont les fonctions de croyance qui ont été utilisées pour *Milesin*. Ce choix est justifié dans la section suivante.

5.2.4 Critiques et théorie des fonctions de croyance

Plusieurs critiques de natures diverses peuvent être émises quant à la pertinence de la méthode de combinaison utilisant un maximum et une addition des probabilités de l'arbre de classification (exposée précédemment en section 5.2.1). Même si elle a donné de bons résultats (cf. section 5.2.3), la notion de score est difficile à interpréter. Ces quantités obtenues ne représentent pas de degré de confiance, ni de probabilité que tel nom complet corresponde à tel locuteur. Elles conduisent à un manque de lisibilité de la décision. L'équation 5.3 sur laquelle s'appuie la décision représente un compromis qu'il est difficile de justifier.

Mais la critique principale concerne la méthode même de combinaison : le conflit d'informations au sein d'un tour de parole donné n'est pas pris en compte. L'information disponible n'est pas combinée dans sa globalité de façon satisfaisante. La méthode de combinaison proposée est susceptible de propager des imperfections pour l'étape suivante de la décision. En effet, l'affectation d'un nom complet à un locuteur ne tient pas compte des liens entre les différentes informations fournies par l'arbre de classification, en particulier quand un même locuteur prononce plusieurs noms complets et peut donc aboutir à des résultats erronés. Le tableau 5.5 présente un exemple de tour de parole t_k où 8 noms complets sont détectés. Les probabilités indiquées correspondent au locuteur du tour suivant t_{k+1}, qui est de genre masculin. L'un des noms complets (féminin) est donc éliminé. Certaines occurrences sont redondantes, car elles correspondent à un même nom complet (*Jean-Claude Pajak* et *Jacques Chi-*

rac) et une seule occurrence a une probabilité élevée. Il reste donc 4 noms complets en compétition, ce qui représente une incompatibilité importante. Or, la contribution de ce tour produit des scores élevés pour *Jean-Claude Pajak* (1, 25) et *Jacques Chirac* (0, 87), scores semblables à celui qu'on obtiendrait si on disposait d'une information sans ambigüité, comme par exemple un tour ne contenant qu'une seule occurrence de probabilité élevée sans concurrence. En effet, même si *Jacques Chirac* obtient un score relativement élevé (0, 87) avec la méthode de combinaison proposée, le fait que ce score provienne de 3 scores relativement faibles de l'arbre (0, 29) est complètement occulté.

TABLE 5.5 – *Contribution du score dans un tour de parole (le genre du locuteur t_{k+1} est masculin).*

Occurrence o_j	sexe	$P(o_j, t_{k+1})$	score
Oscar Temaru	M	0, 29	0, 29
Hamid Karzaï	M	0, 29	0, 29
Jacques Chirac	M	0, 29	0, 87
Jacques Chirac	M	0, 29	
Jacques Chirac	M	0, 29	
Jean-Claude Pajak	M	0, 29	
Jean-Claude Pajak	M	**0,96**	1, 25
Véronique Rebeyrotte	*F*	*0,29*	–

Cet exemple met en évidence le fait que cette méthode ne prend pas en compte la contradiction des informations délivrées par certains tours de parole. Même si une solution intermédiaire a été présentée dans 5.2.2, il convient de réajuster le calcul des scores en tenant mieux compte de l'incertitude de ces informations. Un formalisme probabiliste basé par exemple sur des probabilités conditionnelles peut être envisagé pour ce type de situations, mais le manque d'informations *a priori* rend ce type de modélisation difficile. Bien que les sorties de l'arbre soient de nature probabiliste, la théorie des croyances nous a paru mieux adaptée et moins contraignante, grâce en particulier à la souplesse de son utilisation. C'est pourquoi c'est cette théorie qui a finalement été retenue dans les travaux présentés dans 4.3.

5.3 Liste de locuteurs et applications

Tous les travaux menés jusqu'ici utilisent une liste de noms complets représentant les locuteurs potentiels (généralement appelée « liste de locuteurs »). Elle peut être utilisée pour étiqueter les noms complets directement dans la transcription (comme pour les travaux de Canseco ou de Tranter) ou pour rejeter certaines décisions qui ne concernent pas des locuteurs cibles (comme *Mile-*

sin). L'utilisation de telles listes est difficilement justifiable si le contexte d'utilisation du système d'INL n'est pas précisé.

5.3.1 Contexte

Imaginons des enregistrements radiophoniques (ou télévisuels) où les intervenants sont connus car ils peuvent être obtenus via le programme des émissions ou via des connaissances externes. Dans ce cas un système d'INL peut être utilisé pour étiqueter automatiquement les locuteurs et les interventions de manière précise dans chaque enregistrement. Il est alors tout à fait acceptable d'utiliser une liste de locuteurs par enregistrement, puisque leur identité est déjà connue. De plus, le genre des locuteurs peut généralement être récupéré de la même manière et peut donc être utilisé comme connaissance a priori.

En revanche, lorsque le système d'INL est utilisé pour nommer des enregistrements dont les locuteurs ne sont pas connus, utiliser une liste de locuteurs revient à disposer d'informations *a priori*. Lorsque l'on cherche à nommer automatiquement de grandes collections de documents audio, il est impossible d'obtenir une liste de tous les locuteurs au préalable. Constituer cette liste est, dans ce cas, un des objectifs du processus d'INL. Puisqu'aucune information a priori sur les locuteurs ne peut être obtenue au préalable, le genre des locuteurs doit être déterminé grâce au prénom du nom complet (cf. section 4.4.4) lorsque cela est possible.

5.3.2 Expérimentations

Afin de prendre ces deux cas d'utilisation en compte, nous présentons dans le tableau 5.6 les résultats avec et sans listes de locuteurs. Les résultats avec la liste globale présentée pour *Milesin* (cf. section 4.4.4) servent de référence. Ensuite, une liste de locuteurs différente par enregistrement est utilisée. Cette liste ne contient que les locuteurs du document, ils ont été extraits (ainsi que leur genre) automatiquement des transcriptions de référence. Lorsqu'aucune liste de locuteurs n'est utilisée, le genre des locuteurs est déterminé grâce à la méthode décrite dans 4.3.6.

Lorsqu'une liste de locuteurs par enregistrement est utilisée, les résultats obtenus sont supérieurs à ceux présentés dans 4.4.4 où la liste de locuteurs contient plus de noms complets. Ce comportement s'explique aisément dans la mesure où l'utilisation d'une liste ne contenant que les locuteurs de l'enregistrement permet d'éliminer au maximum les locuteurs parasites : tous les étiquetages

	En durée			En nombre
	R	**P**	**ErrDur**	**ErrLoc**
Corpus de Développement				
Liste globale	87,85%	95,24%	11,44%	12,43 %
Liste par enregistrement	88,35 %	96,46 %	10,95 %	11,46 %
Sans liste	72,86 %	73,39 %	29,27 %	30,22 %
Corpus de Test				
Liste globale	77,10%	92,97%	22,85%	28,23%
Liste par enregistrement	78,00 %	95,68 %	21,95 %	25,81 %
Sans liste	61,77%	64,38 %	38,16 %	35,48 %

TABLE 5.6 – *Différence entre les méthodes de combinaison globales sur des transcriptions enrichies manuelles en utilisant une liste de locuteurs*

de l'arbre concernant des locuteurs ne faisant pas partie de l'enregistrement ne sont pas pris en compte.

En revanche, lorsqu'aucune liste de locuteurs n'est utilisée, les performances chutent considérablement. Même si le rappel chute de l'ordre de 15 %, c'est surtout la chute de la précision qui est problématique pour les systèmes d'INL. En effet, identifier moins de locuteurs mais de façon précise sera préféré à identifier plus de locuteurs en commettant beaucoup d'erreurs. Dans les résultats présentés dans le tableau 5.6, la précision chute de 22 % pour le corpus de développement et de 31 % pour le corpus de test.

Ces résultats s'expliquent de plusieurs manière. Tout d'abord, sans l'utilisation de liste de locuteurs, les erreurs du détecteur d'entités nommées sont propagées au sein du processus. Par exemple, au lieu d'étiqueter *Pierre Kupferman* seul comme entité nommée de type personne, c'est *Bonjour Pierre Kupferman* qui est étiqueté. De même, au lieu d'étiqueter *Hélène Cardin*, LIA_NE étiquette *Shanghai Hélène Cardin*. La délimitation des entités nommées est un des problèmes auquel les détecteurs ont à faire face, cf. section 2.5.3. Même si étiqueter un locuteur comme *Bonjour Pierre Kupferman* au lieu de *Pierre Kupferman* permettrait à un relecteur humain d'obtenir l'information qu'il cherche, ce cas de figure est compté comme une erreur dans notre évaluation.

Ensuite, beaucoup de locuteurs sont nommés avec des noms partiels : *Dominique de Villepin* nommé *Villepin*, *Mohammed Belhaj Soulami* nommé *Mohammed Kamal Belhaj Soulami* ; ou uniquement avec leur prénom : *Joël Collado* nommé *Joël*, *Christophe Godinot* nommé *Christophe*, etc. Pour finir, des locuteurs non nommés dans la référence se retrouvent nommés avec des noms complets qui présentent une probabilité faible pour ce locuteur, mais qui sont les seuls candidats potentiels.

5.3.3 Perspectives

En fonction du contexte, l'utilisation de connaissances a priori peut tout à fait être justifiable. Dans l'hypothèse où la liste des locuteurs ne peut être connue à l'avance, les expériences montrent que le système est beaucoup moins performant. Lorsque les systèmes d'INL sont utilisés sans liste de locuteurs, ils servent à collecter les informations des documents, notamment à des fin de recherche documentaire. Dans ce type de cadre applicatif, il est important de ne pas commettre d'erreurs. Même si le système ne détecte pas tous les locuteurs, il faut que ceux qu'il étiquette soient fiables.

Dans cette optique, les systèmes doivent présenter la plus grande précision possible, même si le rappel reste relativement faible. Dans Milesin, il n'est actuellement pas possible de favoriser la précision au dépend du rappel par exemple. En revanche, les derniers travaux et notamment l'utilisation de la théorie des fonctions de croyance permettent d'avoir des outils à disposition pour ce faire. Cette théorie permet de modéliser l'ignorance et le conflit au sein d'un locuteur. Dans un premier temps, un simple seuil sur ces valeurs devrait permettre de faire augmenter la précision lors de la prise de décision.

Dans un deuxième temps, il sera intéressant de regarder comment il est possible d'améliorer le détecteur d'entités nommées pour diminuer les erreurs de délimitation mentionnées dans la section 5.3.2. Un traitement des coréférences entre prénom, patronyme et nom complet devrait aussi pouvoir améliorer les résultats du système sur certains corpus.

5.4 Transcriptions automatiques

La section 4.4.4 a mis en évidence les mauvaises performances de *Milesin* lorsqu'il est utilisé sur des transcriptions entièrement automatiques. Bien que l'utilisation de systèmes d'INL automatiques comme *Milesin* sur des transcriptions automatiques n'ait pas encore été étudié en profondeur, les sections suivantes présentent une première analyse de ces résultats.

5.4.1 De la pertinence des métriques utilisées

Lorsque l'évaluation est réalisée sur des transcriptions entièrement automatiques (segmentation et classification), il est difficile d'évaluer en nombre de locuteurs comme cela a pu être fait sur des transcriptions manuelles. En effet, il est compliqué de considérer un locuteur comme correct, quand les segmentations et classifications peuvent varier entre la référence et l'hypothèse.

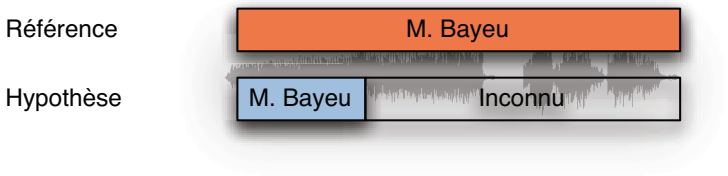

FIGURE 5.2 – *Problème d'identification nommée dû à une mauvaise segmentation*

La figure 5.2 illustre un de ces problèmes. Dans la référence, un seul tour de parole est affecté au locuteur *Maude Bayeu*. Dans l'hypothèse, le système de segmentation a commis une erreur et seule une partie du tour de parole de référence a pu être étiqueté. Même si un tel étiquetage permet de repérer dans l'enregistrement le début d'une des interventions de Maude Bayeu, il présentera un taux d'erreur très élevé.

En fonction du type d'application visée, l'évaluation doit aussi être adaptée. Si le but est d'étiqueter les interventions des locuteurs de manière précise dans les enregistrements, la mesure en durée est pertinente. En revanche, si le but est d'obtenir uniquement une liste des locuteurs de l'enregistrement, sans se soucier d'où il parle, la mesure en durée présente moins d'intérêt. Par exemple, s'il on considère qu'obtenir le début de chaque intervention de chaque locuteur est suffisant pour repérer un locuteur dans en enregistrement, une métrique évaluant cela pourrait être mise en place.

Toutes ces problématiques sont en cours d'étude dans le projet ANR REPERE 2010 [1], et feront l'objet de futur travaux au LIUM.

5.4.2 Influence de la qualité des transcriptions enrichies

Afin d'étudier l'impact des différentes composantes d'une transcription automatique (segmentation/classification et transcription), le tableau 5.7 présente les résultats de *Milesin* en ne faisant varier qu'une des deux composantes à la fois. Cette étude à été menée en utilisant une liste de locuteur par enregistrement.

Milesin se comporte de façon similaire sur le corpus de développement ou sur le corpus de test : plus le système tend vers un système automatique et plus les performances se dégradent. La qualité de la transcription a plus d'impact sur les performances que celle de la segmentation / classification. Par exemple, sur

1. `http://www.agence-nationale-recherche.fr/programmes-de-recherche/projets-selectionnes/programme-contenus-et-interactions-defi-multimedia-reconnaiss`

		En durée			En nb de Locuteur
Trans.	**Seg/Class.**	**R**	**P**	**ErrDur**	**ErrLoc**
Corpus de développement					
M	**M**	88,35 %	96,46 %	10,95 %	11,46 %
M	**A**	59,66 %	79,47 %	38,35 %	- %
A	**M**	41,74 %	85,44 %	54,79 %	61,35 %
A	**A**	25,15 %	65,99 %	69,85 %	-
Corpus de test					
M	**M**	78,00 %	95,68 %	21,95 %	25,81 %
M	**A**	34,41 %	64,62 %	65,59 %	-
A	**M**	54,66 %	89,03 %	45,34 %	49,19 %
A	**A**	42,50 %	85,48 %	57,50 %	-

TABLE 5.7 – *Résultats de* Milesin *utilisant une transcription enrichie manuelle ou automatique sur le corpus d'évaluation ESTER 1 phase II*

Trans. *: Transcription* ***M****anuelle ou* ***A****utomatique.*
Seg/Class. *: segmentation/classification manuelles ou automatiques.*
R, P, F *: rappel, précision et F-mesure calculés en durée.*
ErrDur *: Taux d'erreur en durée.*
ErrLoc *: Taux d'erreur en nombre de locuteurs.*

le corpus de développement, utiliser une transcription automatique augmente le taux d'erreur d'environ 43 points alors qu'utiliser une segmentation / classification automatique ne l'augmente que de 27 points (par rapport aux transcriptions manuelles de référence).

Concernant la classification et la segmentation automatique en locuteur, nous avons constaté que la segmentation automatique engendrait plus d'erreurs que la classification automatique. Actuellement, ce module a été développé pour minimiser le taux de DER et n'a pas subit de modifications en vue de l'adapter aux techniques d'INL.

Sur le corpus de développement transcrit automatiquement, les locuteurs dont il est impossible de récupérer le nom à cause d'erreurs de transcriptions représentent 47 % du temps de parole total contre environ 5 % lorsque les transcriptions sont réalisées manuellement. Sur le corpus de test, ils représentent 32 % contre 10 % sur les transcriptions entièrement manuelles.

Le tableau 5.8 compare les résultats d'identification nommée obtenus en utilisant les transcriptions réalisées par le système du LIUM et celles réalisées par le système du LIMSI (Gauvain et al., 2005). La transcription du LIMSI correspond à la transcription générée par leur système durant la campagne d'évaluation ESTER 1 phase II en 2005, où ce système a obtenu les meilleurs résultats avec un taux d'erreur sur les mots de 11,9 %. Les transcriptions ont été générées à partir de segmentations et de classifications automatiques. Pour supprimer les erreurs de segmentation et de classification, les mots ont été replacés dans les segments de référence avant d'appliquer le système de détection d'entités nommées.

La transcription du LIMSI permet de réduire les taux d'erreur en durée *ErrDur* de 13 points et le taux d'erreur en nombre de locuteurs *ErrLoc* d'environ dix points. Elle contient notamment plus de noms complets correctement transcrits que celle du LIUM, ce qui a un impact non négligeable sur les résultats du processus d'identification nommée. En effet, alors que 1 541 noms complets sont détectés dans les transcriptions de référence, seulement 970 sont détectés dans les transcriptions du LIUM contre 1 192 dans les transcriptions du LIMSI.

5.5 Bilan

Afin de mieux comprendre les erreurs commises par le système sur lequel s'appuyait cette thèse, une analyse sur corpus a tout d'abord été menée. Cette analyse a permis de mettre en évidence les problèmes relatifs au détecteur d'entités nommées (Nemesis) ainsi qu'au processus de décision. L'utilisation de Nemesis, un détecteur d'entités nommées issues du TAL, a été un premier pas

	En durée			En nb de locuteurs
Transcription	**Rappel**	**Précision**	**ErrDur**	**ErrLoc**
LIUM	41,74 %	85,44 %	54,79 %	61,35 %
LIMSI	56,31 %	86,17 %	41,24 %	51,79 %

TABLE 5.8 – *Comparaison des résultats avec deux systèmes de transcription différents sur le corpus de développement*

Rappel, ***précision*** *et* ***f-mesure*** *calculés en en durée.*
ErrDur *: taux d'erreur en durée.*
ErrLoc *: taux d'erreur en nombre de locuteurs.*

vers l'automatisation de la détection des entités nommées. Mais cet outil, prévu pour être utilisé sur des textes écrits (correctement ponctués, accentués, ...), s'est avéré peu adapté à des transcriptions d'enregistrements audio. C'est pourquoi LIA_NE, développé spécifiquement pour travailler sur des transcriptions, lui a été préféré.

Le système de décision a été étudié et amélioré à plusieurs reprises pendant ces travaux. L'utilisation d'un simple maximum a tout d'abord été remplacée par une méthode de normalisation des scores. Bien que donnant de bons résultats, il a été décidé de remplacer cette méthode par la théorie des fonctions de croyance. En effet, certains points de la méthode à base de normalisations sont difficilement justifiables. La théorie des fonctions de croyance a permis d'apporter un cadre mathématique au processus de décision, pour des performances légèrement inférieures à la technique utilisant la normalisation. La nouvelle méthode n'utilise pas encore toutes les possibilités de la théorie des fonctions de croyance (exploitation du conflit, de l'incertitude), de futurs travaux sur cette partie sont à prévoir.

Les différents contextes applicatifs des systèmes d'INL ont ensuite été présentés, et plus particulièrement la pertinence de l'utilisation de listes de locuteurs. En fonction de l'utilisation souhaitée d'un système d'INL, il peut tout à fait être pertinent de disposer d'une liste de locuteurs à nommer connue à l'avance.

Pour finir, des premiers résultats sur des transcriptions entièrement automatiques ont été donnés. La majeure partie du travail réalisé pour *Milesni* l'a été sur des transcriptions entièrement manuelles. De ce fait, le passage à l'automatique n'a pas encore été totalement approfondi et les performances sont moins bonnes sur ce type de transcriptions.

Chapitre 6

Conclusion et perspectives

Cette thèse s'inscrit dans le domaine de l'identificaton automatique de locuteurs. Plus particulièrement, les travaux présentés ici s'intéressent à l'identification de locuteurs par leur prénom et leur nom dans des documents audio en français. Ce processus est appelé identification nommée du locuteur (INL).

L'INL contribue à l'indexation automatique de documents audio en fournissant des informations sur les locuteurs du document. Elle permet d'identifier les différents locuteurs par leur prénom et leur nom, et par la même occasion d'obtenir la liste des interventions de ces locuteurs dans les enregistrements.

Plusieurs approchent ont été étudiées dans la littérature, et notamment une approche utilisant des techniques de suivi du locuteur. Ces techniques nécessitent de connaître les locuteurs à identifier au préalable, de manière à obtenir des échantillons de leurs voix. Ces échantillons seront ensuite utilisés pour apprendre des modèles de voix afin d'identifier les locuteurs dans les enregistrements. Ce type d'approche a un inconvénient majeur : il faut disposer de données a priori qui sont difficile à obtenir.

Afin de ne pas devoir disposer de connaissances a priori sur les documents à traiter, un autre type de système d'INL a été étudié. En s'appuyant sur des techniques de reconnaissance automatique du locuteur et sur la transcription du signal fournie par des systèmes de reconnaissance automatique de la parole, ces systèmes permettent d'identifier les locuteurs par leur prénom et leur nom sans disposer de connaissances a priori sur ces locuteurs. L'hypothèse fondatrice de ces travaux est que les locuteurs du document s'annoncent entre eux. Il est alors possible d'utiliser cette information pour nommer les locuteurs du document.

Lex travaux présentés ici ont permis plusieurs avancées. Tout d'abord *Milesin* est le premier système d'INL (basé sur la transcription) pouvant fonction-

ner de manière entièrement automatique. L'utilisation d'un détecteur d'entités nommées comme LIA_NE permet de s'affranchir des phases d'étiquetages manuelles réalisées dans les précédents travaux.

Ensuite l'utilisation d'un arbre de classification en lieu et place du modèle n-grammes utilisé dans la littérature permet de rendre le système plus robuste aux erreurs de transcriptions issues des systèmes de reconnaissance automatique de la parole.

Pour finir, la mise en place d'un formalisme mathématique comme la théorie des fonctions de croyance a permis de combiner les différentes informations issues de l'arbre de manière efficace. Ce formalisme, associé à un algorithme de décision optimal (l'algorithme Hongrois, aussi appelé algorithme de Kuhn-Munkres) a permis d'obtenir un taux d'erreur d'environ 11 % sur le corpus de développement et d'environ 22 % sur le corpus de test.

Pespectives

Milesin est le fruit de diverses améliorations apportées au système pré-existant au LIUM. Le système d'entités nommées a tout d'abord été remplacé par LIA_NE, le détecteur d'entités nommées le plus performant de la campagne d'évaluation ESTER 2. Ensuite, l'étape d'attribution locale ainsi que le processus d'attribution globale ont été améliorés à plusieurs reprises. Le choix du système à base de fonctions de croyances a été motivé par la modélisation mathématique que permet cette théorie. Grâce à la théorie des fonctions de croyance, il est maintenant possible de modéliser le conflit, l'ignorance ou la croyance au sein d'un locuteur à nommer. Bien que ces informations soient potentiellement très utiles, elles n'ont pour l'instant été que partiellement exploitées : le conflit et l'ignorance ne sont pour l'instant pas pris en compte. La prise en compte de ces informations devrait permettre d'augmenter la précision du système.

Les travaux sur l'identification nommée du locuteur doivent maintenant porter sur la prise en compte des spécificités des transcriptions automatiques. Il conviendrait par exemple de remettre en cause l'injectivité de la fonction d'assignation d'un nom complet aux locuteurs $f(c_l) = f(c'_l) \Rightarrow c_l = c'_l$. Si cette hypothèse est tout à fait pertinente pour des transcriptions manuelles, elle l'est beaucoup moins pour des transcriptions automatiques, à cause des erreurs de classification et de segmentation. Ensuite, les systèmes de reconnaissance automatique de la parole commettent beaucoup d'erreur lors de la transcription de noms propres. De récents travaux sur la phonétisation automatique de noms propres (Laurent et al., 2010) ce sont intéressés à ce problème. Il conviendrait de les étudier en détail afin de les exploiter pour les systèmes d'INL.

À court terme, il serait intéressat d'adapter les outils de transcription et de détection des entités nommées aux spécificités de la tâche (importance de bien

transcrire les noms complets, de détecter uniquement le prénom et le nom des locuteurs, etc.). Ensuite, il serait intéressant de s'intéresser aux journaux télévisés et par la même occasion de pouvoir exploiter les informations contenues dans la vidéo (incrustation des noms des locuteurs dans les sous-titres) pour améliorer le processus.

À plus long terme, un travail sur de grandes collections de données serait intéressant. Étudier dans quelle mesure les informations extraites d'un document peuvent être utilisées pour améliorer le processus d'INL au sein de la collection entière. Ces travaux pourraient aussi être étendus à la détection du rôle des locuteurs dans un enregistrement ou à la détection des différentes interactions entre ces locuteurs.

Liste des illustrations

Liste des tableaux

Bibliographie

(Allauzen, 2003) A. Allauzen, 2003. *Modélisation linguistique pour l'indexation automatique de documents audiovisuels*. Thèse de Doctorat, Paris XI/LIMSI CNRS.

(Alto et al., 1987) P. Alto, M. Brandetti, M. Ferretti, G. Maltese, et S. S., 1987. A speech recognition system for the italian language. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, USA, 941–943.

(Alvin et Martin, 2004) M. P. Alvin et A. Martin, 2004. Nist speaker recognition evaluation chronicles. Dans les actes de *Odyssey 2004, The Speaker and Language Recognition Workshop*, 12–22.

(Atal, 1976) B. S. Atal, 1976. Automatic recognition of speakers from their voices. Dans les actes de *IEEE*, Volume 64(4), 460–475.

(Averbuch et al, 1987) A. Averbuch et al, 1987. Experiments with the TANGORA 20,000 word speech recognizer. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Dallas, Texas, USA, 701–704.

(Bahl et al., 1995) L. Bahl, S. Balakrishnan-Aiyer, J. Bellgarda, M. Franz, P. Gopalakrishnan, D. Nahamoo, M. Novak, M. Padmanabhan, M. Picheny, et R. S., 1995. Performance of the ibm large vocabulary continuous speech recognition system on the arpa wall street journal task. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Volume 1, Detroit, USA, 41–44.

(Bazillon et al., 2008) T. Bazillon, Y. Estève, et D. Luzzati, 2008. Manual vs assisted transcription of prepared and spontaneous speech. Dans les actes de *LREC, Language Evaluation and Resources Conference*, Marrakech, Maroc.

(Bechet et Charton, 2010) F. Bechet et E. Charton, 2010. Unsupervised knowledge acquisition for extracting named entities from speech. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Dallas, USA.

(Béchet et al., 2000) F. Béchet, A. Nasr, et F. Genet, 2000. Tagging unknown proper names using decision trees. Dans les actes de *38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong, Chine, 77–84.

(Beeferman et al., 1998) D. Beeferman, A. Berger, et J. Lafferty, 1998. Cyberpunc : A lightweight punctuation annotation system for speech. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, 689–692.

(Berger et al., 1996) A. L. Berger, V. J. D. Pietra, et S. A. D. Pietra, 1996. A maximum entropy approach to natural language processing. *Computational Linguistics 22*(1), 39–71.

(Bimbot et al., 2004) F. Bimbot, J.-F. Bonastre, C. Fredouille, G. Gravier, I. Magrin-Chagnolleau, S. Meignier, T. Merlin, J. Ortega-Garcia, D. Petrovska, et D. A. Reynolds, 2004. A tutorial on text-independent speaker verification. *EURASIP Journal on Applied Signal Processing, Special issue on biometric signal processing 4*, 430–451.

(Bonastre et al., 2003) J.-F. Bonastre, F. Bimbot, L.-J. Boe, J.-P. Campbell, D.-A. Reynolds, et I. Magrin-Chagnolleau, 2003. Person authentication by voice : A need for caution. Dans les actes de *European Conference on Speech Communication and Technology (Eurospeech)*, Genève, Suisse.

(Bonastre et al., 2000) J.-F. Bonastre, P. Delacourt, C. Fredouille, S. Meignier, T. Merlin, et C. J. Wellekens, 2000. Différentes stratégies pour le suivi de locuteur. Dans les actes de *RFIA'2000, Reconnaissance des Formes et Intelligence Artificelle*, Paris, France.

(Boë et al., 1999) L.-J. Boë, F. Bimbot, J.-F. Bonastre, et P. Dupont, 1999. De l'évaluation des systèmes de vérification du locuteur à la mise en cause des expertises vocales en identification juridique. *Langues 2(4)*, 270–288.

(Breiman et al., 1984) L. Breiman, R. Friedman, R. Olshen, et C. Stone, 1984. *Classification and Regression Trees*. Wadsworth.

(Bridle et Brown, 1974) J. S. Bridle et M. D. Brown, 1974. An experimental automatic word-recognition system. Dans les actes de *Joint Speech Research Unit Report No. 1003*, Ruislip, England.

(Brun et Ehrmann, 2010) C. Brun et M. Ehrmann, 2010. Un système de detection d'entités nommées adapté pour la campagne d'évaluation ESTER 2. Dans les actes de *TALN (Traitement Automatique des Langues Naturelles)*, Montréal (Canada).

(Burkard et al., 2009) R. Burkard, M. Dell'Amico, et S. Martello, 2009. Society for Industrial and Applied Mathematics.

(Canseco-Rodriguez, 2006) L. Canseco-Rodriguez, 2006. *Speaker Diarization in Broadcast News*. Thèse de Doctorat, École doctorale sciences et sechnologies de l'information des télécommunications et des systèmes, Université Paris XI.

(Canseco-Rodriguez et al., 2005) L. Canseco-Rodriguez, L. Lamel, et J. Gauvain, 2005. A comparative study using manual and automatic transcriptions for diarization. Dans les actes de *ASRU, Automatic Speech Recognition and Understanding*, San Juan, Porto Rico, USA.

(Cerf-Danon et al., 1990) H. Cerf-Danon, P. De La Noue, L. Diringer, M. El-Bèze, et J. Marcadet, 1990. A 20,000 words, automatic speech recognizer. Adaptation to French of the US TANGORA system. Dans les actes de *Nato*.

(Cerf-Danon et al., 1991) H. Cerf-Danon, M. El-Bèze, et B. Merialdo, 1991. Reconnaissance automatique de la parole. *Informatique et Santé : Nouvelles Technologies et Traitement de l'Information en médecine 4*, 84–100.

(Chen et Rosenfeld, 2000) S. F. Chen et R. Rosenfeld, 2000. A survey of smoothing techniques for me models. *IEEE Transactions on Speech and Audio Processing 8*(2), 37–50.

(Chengyuan et al., 2007) M. Chengyuan, P. Nguyen, et M. Mahajan, 2007. Finding speaker identities with a conditional maximum entropy model. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA.

(Coates-Stephens, 1992) S. Coates-Stephens, 1992. The analysis and acquisition of proper names for the understanding of free text. *Computers and the Humanities 26(5)*, 441–456.

(Cornuéjols et Miclet, 2002) A. Cornuéjols et L. Miclet, 2002. *Apprentissage artificiel : concepts et algorithmes*. Eyrolles.

(Daille et Morin, 2000) B. Daille et E. Morin, 2000. Reconnaissance automatique des noms propres de la langue écrite : Les récentes réalisations. Dans les actes de *Traitement automatique des langues*, Volume 41(3), 601–621.

(Deléglise et al., 2005) P. Deléglise, Y. Estève, S. Meignier, et T. Merlin, 2005. The LIUM speech transcription system : a CMU Sphinx III-based system for french broadcast news. Dans les actes de *Interspeech'05*, Lisbon.

(Dempster et al., 1977) A. P. Dempster, N. M. Laird, et D. B. Rubin, 1977. Maximum likelihood from incomplete data via the em algorithm. *Journal Of The Royal Statistical Society, Series B 39*(1), 1–38.

(Doddington, 1985) G.-R. Doddington, 1985. Speaker recognition : identifying people by their voices. *Proceedings of the IEEE 73(11)*, 1651–1664.

(Estève et al., 2007) Y. Estève, S. Meignier, P. Deléglise, et J. Mauclair, 2007. Extracting true speaker identities from transcriptions. Dans les actes de *Interspeech, European Conference on Speech Communication and Technology*, Antwerp, Belgium.

(Estève et al., 2010a) Y. Estève, T. Bazillon, J.-Y. Antoine, F. Béchet, et J. Farinas, 2010a. The EPAC corpus : manual and automatic annotations of conversational speech in French broadcast news. Dans les actes de *LREC 2010*, Malta.

(Estève et al., 2010b) Y. Estève, P. Deléglise, S. Meignier, S. Petitrenaud, H. Schwenk, L. Barrault, F. Bougares, R. Dufour, V. Jousse, A. Laurent, et A. Rousseau, 2010b. Some recent research work at lium based on the use of cmu sphinx. Dans les actes de *CMU SPUD Workshop*, Dallas (Texas).

(Favre et al., 2010) B. Favre, B. Bohnet, et D. Hakkani-Tür, 2010. Evaluation of Semantic Role Labeling and Dependency Parsing of Automatic Speech Recognition Output. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Dallas (USA).

(Fourour, 2004) N. Fourour, 2004. *Identification et catégorisation automatiques des entités nommées dans les textes français*. Thèse de Doctorat, Thèse en informatique de l'université de Nantes.

(Gales, 1998) M.-J.-F. Gales, 1998. Maximum likelihood linear transformations for hmm-based speech recognition. *Computer Speech and Language 12*, 75–98.

(Galliano et al., 2006) S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, et K. Choukri, 2006. Corpus description of the ESTER evaluation campaign for the rich transcription of french broadcast news. Dans les actes de *LREC, Language Evaluation and Resources Conference*, Genoa, Italy.

(Galliano et al., 2005) S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J. F. Bonastre, et G. Gravier, 2005. The ESTER phase II evaluation campaign for the rich transcription of french broadcast news. Dans les actes de *Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal.

(Galliano et al., 2009) S. Galliano, G. Gravier, et L. Chaubard, 2009. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. Dans les actes de *Interspeech*, Brighton, UK, 2583–2586.

(Gauvain et al., 2005) J.-L. Gauvain, G. Adda, M. Adda-Decker, A. Allauzen, V. Gendner, L. Lamel, et H. Schwenk, 2005. Where are we in transcribing french broadcast news ? Dans les actes de *Interspeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal.

(Gauvain et al., 1994) J. L. Gauvain, L. F. Lamel, G. Adda, et M. Adda-Decker, 1994. Speaker-independent continuous speech dictation. *Speech Communication 15*, 21–37.

(Graff, 1996) D. Graff, 1996. The 1996 broadcast news speech and language-model corpus. Dans les actes de *1997 DARPA Speech Recognition Workshop*, 11–14.

(Grass, 2000) T. Grass, 2000. Typologie et traductibilité des noms propres de l'allemand vers le français. Dans les actes de *Traitement automatique des langues*, Volume 41(3), 643–670.

(Gravier et al., 2004) G. Gravier, J.-F. Bonastre, S. Galliano, et E. Geoffrois, 2004. The ESTER evaluation campaign of rich transcription of french broadcast news. Dans les actes de *LREC, Language Evaluation and Resources Conference*, Lisbon, Portugal.

(Grishman et Sundheim, 1996) R. Grishman et B. Sundheim, 1996. Message understanding conference-6 : a brief history. Dans les actes de *Proceedings of the 16th conference on Computational linguistics*, Morristown, NJ, USA, 466–471. Association for Computational Linguistics.

(Hermansky et Cox, 1991) H. Hermansky et L. A. Cox, 1991. Perceptual linear predictive (plp) analysis-resynthesis technique. Dans les actes de *IEEE ASSP Workshop*, 0_37–0_38.

(Hollien, 1990) H. Hollien, 1990. *The Acoustics of Crime, The New Science of Forensic Phonetics*. Numéro ISBN 0-306-43467-9. Plenum Publishing Corporation.

(Huang et Zweig, 2002) J. Huang et G. Zweig, 2002. Maximum entropy model for punctuation annotation from speech. Dans les actes de *ICSLP, International Conference on Spoken Language Processing*, Denver, USA, 917–920.

(Jean-luc Gauvain et Lee, 1994) J.-L. Jean-luc Gauvain et C.-H. Lee, 1994. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing 2*, 291–298.

(Jelinek, 1976) F. Jelinek, 1976. Continuous speech recognition by statistical methods. *Proceedings of IEEE*.

(Joly, 2004) A. Joly, 2004. *Recherche par Similarité Statistique dans une grande base de signatures locales pour l'identification rapide d'extraits vidéo*. Thèse de Doctorat, L3I/Université de La Rochelle.

(Klatt, 1977) D. H. Klatt, 1977. Review of the arpa speech understanding project. *The Journal of the Acoustical Society of America 62*(6), 1345–1366.

(Kuhn et De Mori, 1995) R. Kuhn et R. De Mori, 1995. The application of semantic classification trees to natural language understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence 17*(5), 449–460.

(Kuhn et Demori, 1998) R. Kuhn et R. Demori, 1998. *Sentence interpretation*. Academic Press.

(Kuhn, 1993) R. E. W. Kuhn, 1993. *Keyword classification trees for speech understanding systems*. Thèse de Doctorat, Montreal, Que., Canada, Canada.

(Lamel et al., 2002) L. Lamel, J.-L. Gauvain, et G. Adda, 2002. Lightly supervised and unsupervised acoustic model training. *Computer Speech and Language 16*, 115–129.

(Lamel et al., 1991) L. Lamel, G. J.L., et E. M., 1991. Bref, a large vocabulary spoken corpus for french. Dans les actes de *Eurospeech*, Genova, Italy, 505–508.

(Larcher et al., 2010) A. Larcher, C. Lévy, D. Matrouf, et J.-F. Bonastre, 2010. Reconnaissance automatique du locuteur embarquée dans un téléphone portable. Dans les actes de *Journées d'Etudes sur la Parole (JEP)*, Mons, Belgique.

(Laurent et al., 2010) A. Laurent, S. Meignier, T. Merlin, et P. Deléglise, 2010. Acoustics-based phonetic transcription method for proper nouns. Dans les actes de *International Conference on Spoken Language Processing (Interspeech 2010)*, Japon (Makuhari).

(Lehuen et Lemeunier, 2010) J. Lehuen et T. Lemeunier, 2010. A robust semantic parser designed for spoken dialog sytems. Dans les actes de *IEEE International Conference on Semantic Computing (ICSC 2010)*, Pittsburgh, PA (USA).

(Makhoul et al., 1999) J. Makhoul, F. Kubala, R. Schwartz, et R. Weischedel, 1999. Performance measures for information extraction. Dans les actes de *DARPA Broadcast News Workshop*, 249–252.

(Mauclair et al., 2006) J. Mauclair, S. Meignier, et Y. Estève, 2006. Speaker diarization : about whom the speaker is talking ? Dans les actes de *IEEE Odyssey 2006*, San Juan, Puerto Rico, USA.

(McCallum et Li, 2003) A. McCallum et W. Li, 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. Dans les actes de *7^{th} conference on Natural language learning at HLT-NAACL 2003*, Morristown, NJ, USA, 188–191. Association for Computational Linguistics.

(McCowan et al., 2004) I. A. McCowan, D. Moore, J. Dines, D. Gatica-Perez, M. Flynn, P. Wellner, et H. Bourlard, 2004. On the use of information retrieval measures for speech recognition evaluation. *Idiap-RR Idiap-RR-73-2004*.

(Meignier, 2002) S. Meignier, 2002. *Indexation en locuteurs de documents sonores : Segmentation d'un document et Appariement d'une collection*. Thèse de Doctorat, Université d'Avignon et des Pays de Vaucluse.

(Meignier et Merlin, 2010) S. Meignier et T. Merlin, 2010. LIUM_SpkDiarization : An open source toolkit for diarization. Dans les actes de *CMU SPUD Workshop*, Dallas, USA.

(Meignier et al., 2008) S. Meignier, T. Merlin, C. Lévy, A. Larcher, E. Charton, J.-F. Bonastre, L. Besacier, J. Farinas, et B. Ravera, 2008. Mistral : plateforme open source d'authentification biométrique. Dans les actes de *Journées d'Etudes sur la Parole (JEP)*, Avignon, France, 81–84. Association Francophone de la Communication Parlée (AFCP).

(Mermelstein, 1976) P. Mermelstein, 1976. Distance measures for speech recognition, psychological and instrumental. *Pattern Recognition and Artificial Intelligence (C. H. Chen, Ed.)*.

(Munkres, 1957) J. Munkres, 1957. Algorithms for the Assignment and Transportation Problems. *Journal of the Society for Industrial and Applied Mathematics 5*(1), 32–38.

(Nedas, 2005) K. A. Nedas, 2005. Munkres' (Hungarian) Algorithm. `http://konstantinosnedas.com/dev/soft/munkres.htm`.

(NIST, 2004) NIST, 2004. Fall 2004 rich transcription (RT-04F) evaluation plan, (version 4, updated 02/25/2003). `http://www.nist.gov/speech/tests/rt/rt2004/fall/docs/rt04f-eval-plan-v14.pdf`.

(O'Shaughnessy, 1986) D. O'Shaughnessy, 1986. Speaker recognition. *IEEE ASSP Magazine*.

(Paik et al., 1996) W. Paik, E. D. Liddy, E. Yu, et M. McKenna, 1996. Categorizing and standardizing proper nouns for efficient information retrieval. *Corpus processing for lexical acquisition*, 61–73.

(Pallett, 1996) D. Pallett, 1996. 1995 benchmark tests for the ARPA spoken language program. Dans les actes de *SLST Workshop*, Harriman NY, USA.

(Pallett, 2003) D. Pallett, 2003. A look at nist's benchmark asr tests : past, present and future. Dans les actes de *ASRU, Automatic Speech Recognition and Understanding*, St. Thomas, U.S. Virgin Islands.

(Poibeau, 2002) T. Poibeau, 2002. *Extraction d'informations à base de connaissances hybrides*. Thèse de Doctorat, Université Paris-Nord.

(Poli, 2007) J.-P. Poli, 2007. *Structuration automatique de flux télévisuels*. Thèse de Doctorat, LSIS, Université Aix-Marseille III.

(Rabiner, 1989) L. R. Rabiner, 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Volume 77, 257–286.

(Rosenberg et Soong, 1992) A.-E. Rosenberg et F.-K. Soong, 1992. Recent research in automatic speaker recognition. *Advances in Speech Signal Processing*, 701–737.

(Sambur, 1975) M. Sambur, 1975. Selection of acoustical features for speaker identification. *IEEE Transactions on Acoustics, Speech, and Signal Processing 23*, 176–182.

(Sekine et Eriguchi, 2000) S. Sekine et Y. Eriguchi, 2000. Japanese named entity extraction evaluation : analysis of results. Dans les actes de *Proceedings of the 18th conference on Computational linguistics*, Morristown, NJ, USA, 1106–1110. Association for Computational Linguistics.

(Shafer, 1976) G. Shafer, 1976. *A Mathematical Theory of Evidence*. Princeton : Princeton University Press.

(Siegler et al., 1997) M. A. Siegler, U. Jain, B. Raj, et R. M. Stern, 1997. Automatic segmentation, classification and clustering of broadcast news audio. Dans les actes de *Proc. DARPA Speech Recognition Workshop*, 97–99.

(Siu et al., 1992) M.-H. Siu, R. Rohlicek, et H. Gish, 1992. An unsupersised, sequential learning algorithm for segmentation of speech waveforms with multi speakers. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, San Francisco, USA.

(Siu et al., 1991) M.-H. Siu, G. Yu, et H. Gish, 1991. Segregation of speakers for speech recognition and speaker identification. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Toronto, Canada.

(Smets et Kennes, 1994) P. Smets et R. Kennes, 1994. The transferable belief model. *Artificial Intelligence 66*(2), 191–234.

(Standford et al., 2003) V. Standford, J. Garofolo, O. Galibert, M. Michel, et L. C., 2003. The nist smart space and meeting room projects : Signal, acquisition, annotation and metrics. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Hong-Kong, China.

(Sundheim, 1996) B. M. Sundheim, 1996. Overview of results of the muc-6 evaluation. Dans les actes de *Proceedings of a workshop on held at Vienna, Virginia*, Morristown, NJ, USA, 423–442. Association for Computational Linguistics.

(Tranter et Reynolds, 2006) S. Tranter et D. Reynolds, 2006. An overview of automatic speaker diarisation systems. *IEEE Transactions on Audio, Speech and Language Processing 14(5)*, 1557–1565.

(Tranter, 2006) S. E. Tranter, 2006. Who really spoke when ? Finding speaker turns and identities in broadcast news audio. Dans les actes de *IEEE ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Volume 1, Toulouse, France, 1013–1016.

(Valtchev et al., 1997) V. Valtchev, J. J. Odell, P. C. Woodland, et S. J. Young, 1997. MMIE training of large vocabulary recognition systems. *Speech Communication 22*(4), 303–314.

(Veneau, 2001) E. Veneau, 2001. *Macro-segmentation multi-critère et classification de séquences par le contenu dynamique pour l'indexation vidéo*. Thèse de Doctorat, IRISA/INRIA Rennes.

(Wacholder et al., 1997) N. Wacholder, Y. Ravin, et M. Choi, 1997. Disambiguisation of proper names in text. Dans les actes de *5th Conference on Applied Natural Language Processing (ANLP'97)*, Washington DC, États-Unis, 202–208.

(Wothke et al., 1989) K. Wothke, U. Bandara, J. Kempf, E. Keppel, K. Mohr, et G. Walch, 1989. SPRING speech recognition system for german. Dans les actes de *Eurospeech, European Conference on Speech Communication and Technology*, Paris, France.