Analyse conjointe du signal sonore et de sa transcription pour l'identification nommée de locuteurs

Vincent Jousse[⋆ ⋆⋆] — **Sylvain Meignier**[⋆] — **Christine Jacquin**[⋆⋆] — **Simon Petitrenaud**[⋆] — **Yannick Estève**[⋆] — **Béatrice Daille**[⋆⋆]

[⋆] *LIUM - Université du Maine, Avenue Laënnec - 72085 Le Mans Cedex*

[⋆⋆] *LINA - 2 rue de la Houssinière - BP 92208 - 44322 Nantes Cedex 03*

prénom.nom@univ-lemans.fr ou prénom.nom@univ-nantes.fr

RÉSUMÉ. *Depuis quelques années, le traitement de très grandes collections de documents multimédias devient crucial pour certaines applications comme les systèmes d'indexation ou de recherche documentaire. Mais ces collections ne peuvent être traitées manuellement avec un coût raisonnable : seuls les systèmes automatiques apportent une solution viable. Dans ce document, nous traiterons de l'extraction automatique de l'identité du locuteur (prénom et patronyme) présente dans les enregistrements sonores. À partir des résultats d'un système de transcription enrichie, nous présentons une méthode qui vise à extraire l'identité des locuteurs de la transcription et à l'assigner aux différents tours de parole. Le système a été évalué sur des enregistrements radiophoniques provenant de la campagne d'évaluation ESTER 1 phase II.*

ABSTRACT. *For some years, processing mass of multimedia documents has become a very crucial issue for applications like indexation or information retrieval. Among the focused information, speaker identity can be very useful for such applications. A huge collection of documents cannot be manually processed with a reasonable cost: only automatic systems are a relevant solution.In this paper, we consider the extraction of speaker identity (firstname and lastname) from audio records of broadcast news. Using a rich transcription system, we present a method which allows to extract speaker identities from automatic transcripts and to assign them to speaker turns. Experiments are carried out on French broadcast news records from the ESTER 1 phase II evaluation campaign.*

MOTS-CLÉS : *identification nommée du locuteur, reconnaissance du locuteur, transcription enrichie.*

KEYWORDS: *speaker named identification, speaker recognition, rich transcription.*

TAL. Volume 50 – n° 1/2009, pages 201 à 225

1. Introduction

Pour faciliter la recherche et l'accès à l'information, les très grandes collections de données audio ont besoin d'être indexées. Les annotations manuelles sont coûteuses, particulièrement pour la transcription des paroles prononcées, les thèmes des documents ou les noms des locuteurs. Il paraît donc nécessaire de s'intéresser à la réalisation automatique de ces annotations. En effet, si le taux d'erreur de ces systèmes est suffisamment faible, ces derniers permettent un gain de temps non négligeable. Le système présenté dans ces travaux s'intéresse au cas de l'annotation des documents avec l'identité des locuteurs. Cette identité est composée du prénom et du patronyme du locuteur, ce couple étant appelé dans la suite du document « nom complet ».

Dans le cadre de l'annotation automatique de documents sonores (appelée aussi transcription enrichie), la première étape consiste, à partir du signal acoustique, à segmenter le signal sonore en tours de parole. Ces derniers débutent lorsqu'un locuteur commence à parler et finissent lorsqu'un autre locuteur prend la parole ou qu'un intermède débute (jingle, chanson, publicité...).

Les différents tours de parole sont ensuite regroupés en classes contenant les segments produits par un même locuteur. Elles sont identifiées par des labels anonymes (locuteur 1, locuteur 2,...). À ce stade, aucune connaissance *a priori* sur les locuteurs n'est utilisée. Toutefois, une détection du genre (homme ou femme) de chaque classe est réalisée. L'étape suivante consiste à transcrire automatiquement les tours de parole en mots. La transcription du document peut être, en plus, complétée par une annotation en entités nommées. Les entités nommées de type « personne » constituent une source d'information sur les locuteurs du document, mais elles ne permettent pas d'identifier directement qui parle et quand.

Pour résoudre ce problème, il existe actuellement deux approches principales qui permettent d'attribuer un nom complet à un locuteur. La première se fonde sur l'analyse exclusive de l'acoustique à partir de méthodes issues de la reconnaissance du locuteur. Par exemple, les systèmes proposés pour la tâche de suivi du locuteur de la campagne d'évaluation ESTER 1 phase II (Galliano *et al.*, 2005) permettent, avec quelques modifications mineures du système de décision, d'identifier les locuteurs d'un document. Ces méthodes reposent sur des modèles de locuteurs appris à partir d'exemples de voix de chaque locuteur cible à identifier. Une des difficultés est de collecter ces échantillons de voix pour que les systèmes de reconnaissance du locuteur aient de bonnes performances. Ils doivent être représentatifs des différentes conditions acoustiques rencontrées dans la collection de documents à annoter. Ils doivent idéalement avoir été enregistrés à la même période que les documents de la collection et en quantité suffisante (plusieurs minutes). De plus, ces systèmes sont amenés à traiter des collections susceptibles d'évoluer quotidiennement ; se pose alors le problème de l'ajout des nouveaux locuteurs.

La seconde approche (Canseco-Rodriguez *et al.*, 2005 ; Tranter, 2006 ; Mauclair *et al.*, 2006 ; Chengyuan *et al.*, 2007) propose d'extraire les noms complets des locuteurs présents dans la transcription enrichie. Le principe général consiste à déterminer

si une entité nommée de type « personne » se rapporte à un locuteur du document, plus exactement à une classe contenant les segments d'un locuteur du document, ou bien à une personne qui ne parle pas dans le document. Seules les entités nommées de type « personne » constituées d'un prénom et d'un patronyme sont retenues. L'approche est fondée sur un système en deux étapes. Une première étape affecte les noms complets aux tours de parole proches. Puis, dans une seconde étape, ces informations sont propagées au niveau des classes. Les documents traités dans ce type d'approche sont des enregistrements de journaux radiophoniques car, généralement, les locuteurs s'annoncent ou sont annoncés tout au long de l'enregistrement.

Dans cet article, la méthode proposée s'apparente à la seconde approche. Le système utilise un arbre de classification sémantique pour déterminer à quel tour de parole se rapporte un nom complet. Nous proposons une méthode originale pour affecter ces noms complets aux classes anonymes à partir des tours de parole identifiés. Le système a été appris et évalué sur des enregistrements radiophoniques en français de la campagne d'évaluation ESTER 1 phase II (Galliano *et al.*, 2006). Les résultats sont donnés sur le corpus d'évaluation composé de dix-huit enregistrements provenant de cinq radios françaises et de Radio Télévision Marocaine.

Dans un premier temps, nous décrivons les contextes et les limites de l'identification nommée avant de présenter l'état de l'art relatif aux différents travaux réalisés sur l'identification nommée du locuteur utilisant des transcriptions enrichies. Puis nous introduisons brièvement le système de transcription enrichie utilisé avant de décrire le système d'identification nommée lui-même. Enfin, nous proposons et discutons des métriques pour évaluer de tels systèmes, puis commentons les résultats obtenus.

2. Contextes et limites

Les méthodes proposées dans (Canseco-Rodriguez *et al.*, 2005 ; Tranter, 2006 ; Mauclair *et al.*, 2006 ; Chengyuan *et al.*, 2007) s'appuient toutes sur une transcription enrichie. Comme le décrit la figure 1, il est supposé que :

– le document est découpé en tours de parole ;

– les tours de parole d'un même locuteur sont regroupés en classes identifiées par des labels anonymes (par exemple *locuteur1*, *locuteur2*, etc.) ;

– le genre (homme ou femme) est renseigné pour chaque classe ;

– la transcription en mots est disponible pour chaque tour de parole ;

– les entités nommées de type « personne » composées d'un prénom et d'un patronyme (nom complet) sont détectées dans la transcription.

L'hypothèse de travail majeure proposée initialement dans (Canseco-Rodriguez *et al.*, 2005) suppose qu'un nom complet détecté dans un tour de parole permet d'identifier le tour courant ou un des tours de parole directement contigus (tour de parole suivant ou précédent). Cependant, certains noms complets identifient des tours de parole plus éloignés ou des personnes n'intervenant pas dans le document. Les personnes

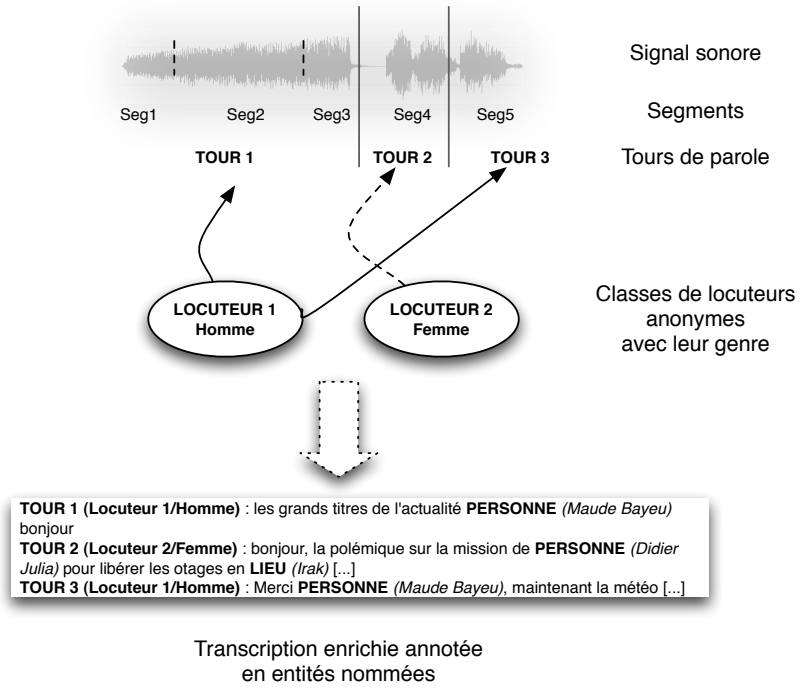

Figure 1. *Informations disponibles dans une transcription enrichie*

n'intervenant pas dans le document ne présentent pas d'intérêt dans le cadre de l'identification nommée car seuls les locuteurs du document sont recherchés. Ces personnes seront donc considérées comme « autre ». Les locuteurs intervenant dans le document, mais dans des tours de parole non contigus pourraient être utilisés pour l'identification. En revanche, le modèle retenu et présenté ci-après se focalise sur les tours de parole contigus et non sur ceux plus éloignés. Les noms de locuteurs faisant référence à des locuteurs de tours de parole plus éloignés sont donc aussi traités comme le cas « autre ».

La figure 2 illustre les quatre types d'affectations possibles pour un nom complet détecté dans un tour de parole.

Les méthodes d'identification nommée à partir d'une transcription enrichie ne sont pertinentes que si les locuteurs s'annoncent ou sont présentés. Elles sont bien adaptées aux enregistrements radiophoniques (journaux d'information par exemple) où le passage de parole est généralement indiqué en nommant le locuteur, mais le sont moins pour les enregistrements de réunion, par exemple. En effet, dans les journaux d'information les locuteurs se présentent ou annoncent le locuteur suivant, ils félicitent le

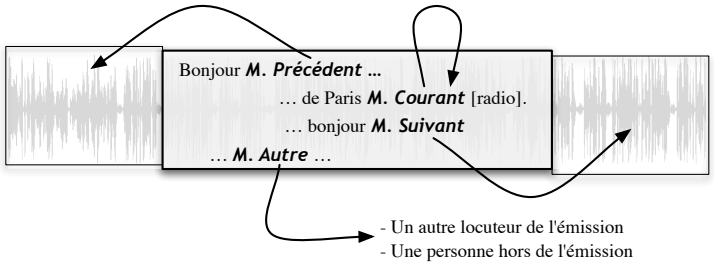

Figure 2. *Principe de base des systèmes d'identification nommée fondés sur une analyse conjointe*

précédent ou le suivant, concluent le reportage par leur nom,... Il est donc possible, en prenant en compte des marqueurs spécifiques, de déterminer à quel tour de parole se rapporte un nom complet détecté. Une étude sur le corpus a montré que les noms complets des locuteurs d'émissions radiophoniques étaient, à de rares exceptions près, systématiquement présents dans la transcription (Jousse *et al.*, 2008).

Ces approches s'appuient sur la transcription des documents. Si cette dernière est réalisée par un système automatique, la détection et l'affectation des noms complets sont tributaires de la qualité des transcriptions. En particulier, il faut que les prénoms et les patronymes des locuteurs fassent partie du vocabulaire du système de transcription. Toutefois, il est plus facile d'adapter un système de transcription pour qu'il tienne compte des nouveaux prénoms et patronymes que de fournir des échantillons de voix à un système de reconnaissance du locuteur.

Ces systèmes peuvent être utilisés dans le domaine de la recherche d'information multimédia pour rechercher, par exemple, des interventions de personnes précises dans des collections de documents. Deux cas d'utilisation peuvent être distingués : soit le système connaît les identités des locuteurs cibles et utilise au mieux cette information, soit il n'a aucune connaissance *a priori* sur ces identités.

3. Méthodes d'identification nommée à partir de transcription enrichie

Les premiers travaux ont été conduits historiquement dans (Canseco-Rodriguez *et al.*, 2005) sur des journaux d'information en langue anglaise. Les auteurs ont été les premiers à montrer que le prénom et le patronyme d'un locuteur apparaissant dans un contexte lexical donné permettaient d'identifier de manière précise l'identité des locuteurs s'exprimant dans les tours de parole proches. Leur méthode repose sur l'utilisation de règles affectant les étiquettes *« tour courant »*, *« tour précédent »*, *« tour suivant »* aux noms complets détectés. Ces étiquettes ont été reprises dans l'ensemble des systèmes décrits ci-dessous. Les règles utilisées ont été définies manuellement après analyse d'un corpus. douze règles sont utilisées pour désigner le locuteur cou-

rant, trentre-quatre pour le suivant et six pour le précédent. Cette méthode nécessite un traitement manuel du corpus : les règles sont décrites par un humain. Le temps de mise en place de telles règles peut être très long suivant la quantité de corpus à analyser. Dans les travaux décrits dans (Canseco-Rodriguez *et al.*, 2005 ; Canseco-Rodriguez, 2006), 150 heures ont été étudiées. Ces règles sont peu transposables d'un corpus à l'autre : il faut réécrire le jeu de règles pour l'utiliser sur des documents d'une autre langue, par exemple. On notera aussi que les entités nommées sont étiquetées manuellement et que le système proposé n'est pas complet. Ces études se sont axées sur la première phase, où les noms complets sont attribués aux tours de parole voisins. La seconde phase, où les locuteurs de l'enregistrement sont nommés grâce à la propagation des entités nommées, n'est pas définie.

La méthode décrite dans (Tranter, 2006) propose un système d'apprentissage automatique à base de n-grammes pour attribuer une étiquette à un nom complet. Un modèle 3-grammes est appris sur une fenêtre glissante de cinq mots autour du nom complet. Ce modèle permet ensuite d'attribuer les étiquettes *« tour courant »*, *« tour précédent »*, *« tour suivant »* ou *« autre »* aux noms complets détectés dans le document traité. Les entités nommées ont été remplacées par les catégories auxquelles elles appartiennent afin de généraliser la transcription. À la différence de celle de (Canseco-Rodriguez *et al.*, 2005), cette méthode a l'avantage d'utiliser un système d'apprentissage automatique pour attribuer les étiquettes relatives aux tours de parole. En revanche, elle utilise une détection manuelle des entités nommées.

Une autre méthode s'appuyant sur un système d'apprentissage automatique est présentée dans (Mauclair *et al.*, 2006). Elle repose sur l'utilisation d'un arbre de classification sémantique (SCT : *Semantic Classification Tree* (Kuhn et De Mori, 1995)) pour attribuer les étiquettes *« tour courant »*, *« tour précédent »*, *« tour suivant »* ou *« autre »* aux noms complets détectés dans la transcription. D'une part, l'arbre de classification sémantique utilise des expressions régulières pour l'attribution des étiquettes. Ces expressions modélisent le contexte lexical des noms complets détectés. D'autre part, l'arbre de classification sémantique utilise des questions globales conjointement aux expressions régulières portant sur la phrase. Notamment, la place du nom complet dans le tour de parole (début, fin, milieu) est un critère améliorant la qualité de la décision. La méthode fondée sur les SCT a été comparée à la méthode s'appuyant sur des n-grammes dans (Estève *et al.*, 2007). Les résultats sont similaires pour les deux méthodes sur des transcriptions manuelles, en revanche le système fondé sur les SCT s'est avéré meilleur sur des transcriptions automatiques. Il est donc plus robuste aux erreurs de transcription.

Récemment, un modèle d'entropie maximale a été proposé dans (Chengyuan *et al.*, 2007). L'attribution des étiquettes utilise des n-grammes identiques à ceux proposés dans (Tranter, 2006). En revanche la prise des décisions est effectuée grâce à ce modèle d'entropie maximale. La position dans le tour de parole du nom complet, comme proposé dans (Mauclair *et al.*, 2006), est aussi intégrée au processus. La décision vérifie en plus la cohérence du genre du prénom du locuteur par rapport au genre du locuteur s'exprimant dans le tour de parole comme l'avait proposé (Canseco-

Rodriguez, 2006). L'ajout des scores d'un système de reconnaissance automatique du locuteur au système d'entropie maximale a aussi été étudié. Le système dispose de 150 modèles de locuteurs communs aux corpus d'évaluation et d'apprentissage représentant 10 % des locuteurs du corpus d'évaluation. L'ajout d'une telle connaissance n'a pas montré de réelle amélioration. Bien que le système d'entropie maximale apporte des gains par rapport au modèle n-grammes, cette méthode ne semble pas meilleure que notre système de référence (*cf.* 6.4.1). Ce système est proche du système utilisant des SCT décrit dans (Estève *et al.*, 2007). Toutefois, il faut nuancer ce propos : en effet les expériences ont été conduites sur des corpus de langues différentes (français et anglais). Contrairement aux systèmes (Canseco-Rodriguez *et al.*, 2005 ; Tranter, 2006 ; Mauclair *et al.*, 2006), on notera que ce système a uniquement été évalué sur des transcriptions enrichies manuellement.

En conclusion, l'attribution des noms complets aux classes dans les systèmes d'identification nommée a été abordée dans la littérature avec diverses méthodes : système à base de règles, modèle n-gramme, arbre de classification. Ces systèmes partagent tous la même architecture en trois étapes :

– une première étape d'attribution des étiquettes *« tour courant »*, *« tour précédent »* ou *« tour suivant »* aux noms complets ;

– une deuxième étape propage ces noms complets dans les tours de parole correspondants et aux classes associées ;

– une troisième étape sélectionne au niveau des classes le nom complet candidat.

D'une part pour la première étape, les travaux de (Estève *et al.*, 2007) ont montré la supériorité de la méthode utilisant un arbre de classification sémantique par rapport aux modèles n-grammes proposés dans (Tranter, 2006). D'autre part pour la troisième étape, le système d'entropie maximale (Chengyuan *et al.*, 2007) s'appuyant aussi sur les scores des modèles n-grammes n'a pas montré des gains probants par rapport au système (Estève *et al.*, 2007).

Au vu de la littérature, l'approche par arbre de classification sémantique est la plus performante pour résoudre la première étape. En revanche, les étapes suivantes, propageant les noms complets et les attribuant aux locuteurs du document, ont été moins étudiées.

4. Système de transcription enrichie utilisé

La méthode d'identification nommée proposée s'appuie sur des documents préalablement transcrits et enrichis. Cette transcription nécessite de découper le document en segments qui seront ensuite classifiés en locuteurs. Ces segments, groupés en tours de parole, sont transcrits et les entités nommées sont annotées. La figure 3 illustre ces trois étapes.

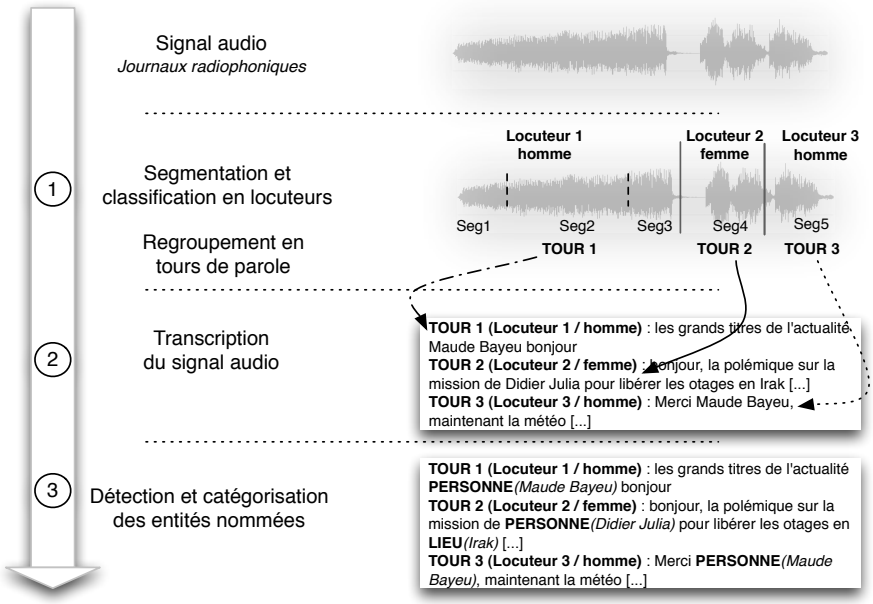

Figure 3. *Description du système de transcription enrichie*

4.1. *Segmentation et classification en locuteur*

L'étape de segmentation et de classification peut être réalisée de manière automatique ou manuelle.

Lorsque cette étape est réalisée par un humain, les frontières des segments correspondent généralement aux inspirations du locuteur ou à des silences. Le découpage en segments ne correspond pas forcément à des phrases. Les étiquettes des classes correspondent ici aux noms des locuteurs, quand il est possible de les déterminer à partir des paroles prononcées ou d'informations annexes comme les grilles de programmes. De la même manière, les segments sont regroupés en tours de parole. Cette étape sera appelée par la suite étape de segmentation et de classification manuelle.

Lorsque le traitement est réalisé de manière automatique, la segmentation consiste à découper le signal audio en segments cohérents contenant les paroles d'un même locuteur prononcées dans des conditions acoustiques similaires (même canal de transcription, même bruit sonore). Ces segments sont relativement courts, généralement de quelques secondes. Puis, ils sont regroupés en classes de locuteurs désignées par des étiquettes anonymes du type *locuteur1*, *locuteur2*, etc. Les segments sont ensuite regroupés en tours de parole. Cette étape sera appelée étape de segmentation et de classification automatique par la suite.

Le système de segmentation et de classification en locuteur du LIUM utilisé pour la tâche a été évalué lors de la campagne d'évaluation ESTER 1 phase II en 2005. Il a été classé deuxième avec un taux d'erreur de segmentation et de classification en locuteur des 16,9 % (DER, Diarization Error Rate[1]). Depuis 2005, le système a évolué vers le système décrit dans (El Khoury *et al.*, 2008) ; les changements majeurs portent sur l'ajout d'une seconde passe de classification utilisant des modèles plus complexes, comme il est fait dans (Barras *et al.*, 2006). Ce système permet d'obtenir 11,5 % de DER sur le même corpus d'évaluation.

4.2. *Transcription*

Les segments sont transcrits soit par un annotateur humain, soit par un système de reconnaissance automatique de la parole. De la même manière que précédemment, nous parlerons respectivement de transcription manuelle et de transcription automatique.

Le système de transcription utilisé est celui du LIUM décrit dans (Deléglise *et al.*, 2005). Ce système repose sur deux décodages successifs avec un modèle de langage 3-grammes et sur une réévaluation d'un graphe généré lors du dernier décodage avec un modèle de langage 4-grammes. Le vocabulaire contient environ 64 000 mots. Lors de la campagne d'évaluation ESTER 1 phase II en 2005, pour la tâche de transcription, le système a été classé deuxième avec un taux d'erreur sur les mots de 23,6 % (WER, Word Error Rate[2]). Depuis 2005, le système a évolué ; les changements majeurs portent sur les paramètres acoustiques et l'apprentissage des modèles acoustiques. Ce système permet d'obtenir 20,5 % de WER sur ce même corpus d'évaluation.

4.3. *Détection d'entités nommées : Nemesis*

Dans l'optique d'un travail sur des transcriptions automatiques, nous avons choisi d'utiliser un étiquetage automatique en entités nommées au lieu de l'étiquetage manuel. L'outil utilisé est Nemesis (Fourour, 2004), développé par l'équipe TALN (traitement automatique des langues naturelles) du LINA (laboratoire d'informatique de Nantes Atlantique).

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 Ne-

1. *« Diarization Error Rate »*, métrique utilisée dans les campagnes d'évaluation du NIST et d'ESTER (NIST, 2003).

2. *« Word Error Rate »*, métrique utilisée dans les campagnes d'évaluation du NIST et d'ESTER.

mesis se compose principalement de quatre 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.

Il est évident que les entités nommées de type « personne » sont primordiales pour le fonctionnement du système d'identification nommée proposé. Toutefois, d'autres entités nommées sont conservées, à savoir les lieux, les radios, et les organisations. Ces dernières permettent en effet de généraliser les informations utilisées par l'arbre de classification en remplaçant des informations spécifiques (les mots) par leurs catégories plus génériques. À noter que la liste des entités nommées retenues que nous venons d'énumérer est très proche de celle proposée dans le système de (Tranter, 2006).

5. Méthode d'identification proposée

La méthode d'identification proposée ici commence par affecter les étiquettes (*« tour courant »*, *« tour précédent »*, *« tour suivant »*) aux noms complets détectés dans la transcription *via* un arbre de classification sémantique. Les noms complets sont ainsi affectés aux tours de parole correspondants avec leurs scores. Ces informations sont agrégées dans les classes fournies par l'étape de segmentation et de classification en locuteurs. Enfin, pour chaque classe, un et un seul nom complet est sélectionné à partir des scores.

5.1. *Arbre de classification sémantique*

La méthode utilise un arbre de décision binaire reposant sur le principe des arbres de classification sémantique (SCT – (Kuhn et De Mori, 1995)), qui apprend automatiquement des règles lexicales à partir des noms complets détectés dans le corpus d'apprentissage. L'arbre permet, lors de la phase d'évaluation, d'attribuer l'étiquette *« tour courant »*, *« tour précédent »*, *« tour suivant »* ou *« autre »* à chaque nom complet détecté.

Lors de la phase d'apprentissage et de test, les cinq mots à gauche et les cinq mots à droite d'un nom complet constituent le contexte lexical analysé par l'arbre de classification sémantique. Cette taille de contexte lexical a été optimisée expérimentalement à partir du corpus de développement. Les articles considérés comme non porteurs de sens pour la tâche d'identification nommée (*l', le, la, les, un, une, uns*) ont été éliminés du contexte analysé. En revanche, les prépositions comme *de, dans, à* ont été gardées puisqu'elles peuvent potentiellement concerner une information importante comme un lieu, une radio ou une émission. Les entités nommées de type « personne »,

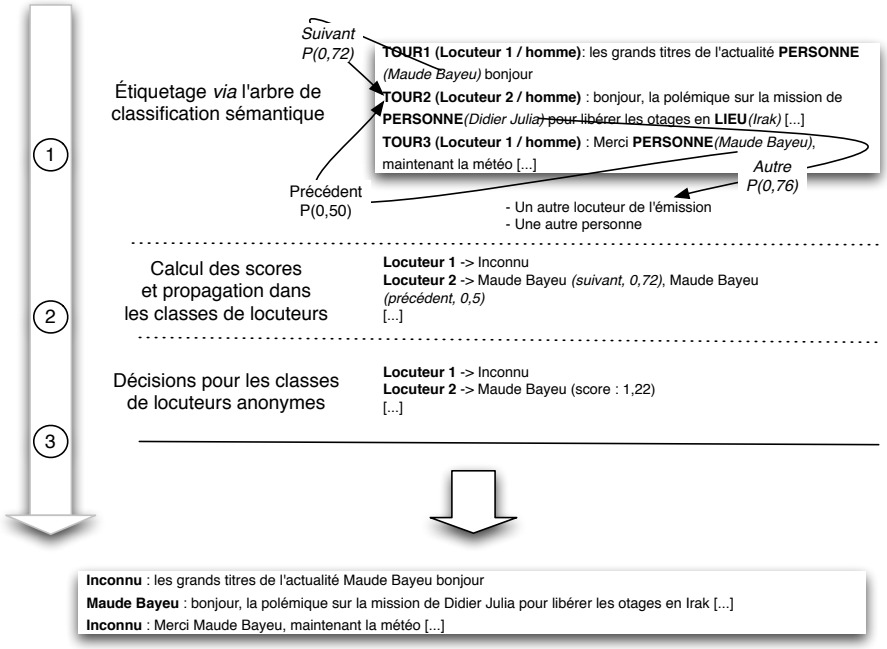

Figure 4. *Description du système d'identification nommée*

« lieu », « radio » et « organisation » ont été remplacées par les catégories correspondantes. De plus, lors de la phase d'apprentissage, un poids beaucoup plus faible est attribué aux échantillons correspondant à l'étiquette *« autre »* ; en effet, ce cas ne nous intéresse pas directement et sert uniquement à rejeter des noms complets. Il a été fixé expérimentalement à partir du corpus de développement.

Les arbres de classification sémantique intègrent dans chaque nœud une expression régulière. La suite d'expressions régulières activées depuis la racine jusqu'à une feuille de l'arbre permet de classifier les contextes lexicaux suivant les quatre étiquettes décrites précédemment. En complément des expressions régulières, l'arbre peut intégrer des questions globales. Le système proposé utilise la position du nom complet dans le tour de parole comme question globale. La position correspond aux situations où le nom complet apparaît au début (dix premiers mots, soit la taille du contexte lexical), au milieu ou à la fin (dix derniers mots) d'un tour de parole. Une autre question globale permet de savoir si le tour de parole est considéré comme très court, à savoir plus petit que le contexte lexical utilisé par l'arbre (cinq mots à gauche du nom complet, cinq mots à droite).

La figure 5 représente un arbre de classification avec deux exemples d'expressions régulières ainsi que les scores associés aux étiquettes dans une des feuilles. Lorsqu'un

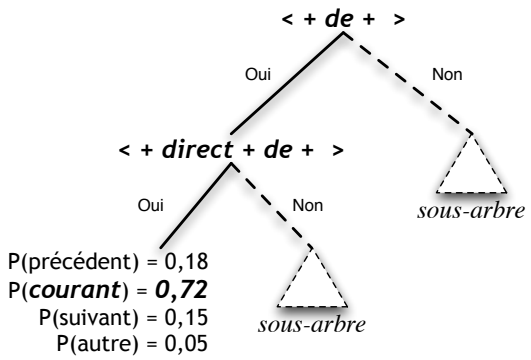

Figure 5. *Arbre de classification sémantique*

exemple atteint une feuille de l'arbre, les scores correspondent aux probabilités de chacune de ces étiquettes. Ces scores sont établis lors de l'apprentissage de l'arbre et reflètent les cas observés dans le corpus d'apprentissage.

Le système proposé utilise l'outil LIA_SCT développé par Frédéric Béchet au laboratoire d'informatique d'Avignon (Bechet *et al.*, 2000). Pour apprendre l'arbre de classification, le corpus d'apprentissage est étiqueté en entités nommées de manière automatique grâce à Nemesis. L'étiquetage en entités nommées donné avec les transcriptions de référence n'est pas utilisé. Des expériences préliminaires ont montré que nous obtenions de meilleures performances ainsi. De plus, les types d'entités nommées disponibles dans le corpus (étiqueté manuellement) et ceux utilisés par notre système sont difficilement compatibles. Pour cette phase d'apprentissage, la segmentation et la classification utilisées sont toujours celles fournies avec le corpus : elles ont été réalisées manuellement. Cela permet de s'affranchir des problèmes induits par une segmentation automatique qui, en cas d'erreur, perturbe les tours de parole : ils n'appartiennent alors plus à un et un seul locuteur. Seul l'étiquetage en entités nommées est donc réalisé automatiquement pour l'apprentissage de l'arbre, le reste de la transcription est réalisé manuellement, elle ne contient donc pas d'erreur.

Pour éviter les problèmes de surapprentissage, une expression régulière est retenue par l'arbre de classification sémantique lorsqu'elle a été rencontrée au moins cinquante fois dans le corpus d'apprentissage. Le critère d'arrêt optimisé lors de l'apprentissage est le critère de Gini (Kuhn et De Mori, 1995).

5.2. *Système de décision*

Dans l'étape précédente, à chaque nom complet détecté dans la transcription est associée une liste d'étiquettes (tour courant,...) valuée par un score. Mais le but final du système est d'affecter à chaque classe un nom complet détecté. Dans ce but, la première étape consiste à associer à chaque tour de parole, la liste des noms complets

qui ont une probabilité non nulle d'en être les locuteurs. Ensuite, un processus de décision détermine le nom complet attribué à chaque classe tout en prenant en compte des contraintes liées d'une part, au regroupement des tours de parole en classes et d'autre part au genre des locuteurs.

Nous prenons comme hypothèse que chaque classe fournie par le système de segmentation et de classification en locuteurs contient des segments mono-locuteurs générés par un même locuteur. On suppose que les segments et les classes sont sans erreur. Cette hypothèse sera discutée dans la section 6.

5.2.1. *Notations*

$\mathcal{E} = \{e_1, \ldots, e_I\}$ correspond à l'ensemble des noms complets candidats pour nommer une classe. Ces candidats sont issus d'une liste des locuteurs possibles connue du système.

Soit $\mathcal{O} = \{o_1, \ldots, o_J\}$ les occurrences des noms complets détectés par Nemesis dans les transcriptions. $\mathcal{T} = \{t_1, \ldots, t_K\}$ désigne l'ensemble des tours de parole, et $\mathcal{C} = \{c_1, \ldots, c_L\}$ l'ensemble des classes à nommer. Les traitements vont permettre d'attribuer un nom complet issu de $\mathcal{E}$ aux classes de $\mathcal{C}$. Une classe est définie par $c_l = \{t_k \in T/\ c_l\ est\ le\ locuteur\ de\ t_k\}$, chaque classe c_l regroupant un ou plusieurs tours de parole t_k. Chaque tour de parole appartient à une et une seule classe.

Pour chaque occurrence d'un nom complet o_j (pour $j = 1, \ldots, J$) détecté dans un tour de parole t_k, on notera t_{k-1} (respectivement t_{k+1}) le tour de parole précédant (respectivement suivant) celui où il a été détecté. De la même manière, le score $P(o_j, t_r)$ désigne la probabilité que o_j soit le locuteur du tour de parole t_r avec $r \in \{k-1, k, k+1\}$. Ces scores sont fournis par l'arbre de classification, ils correspondent respectivement à l'étiquette *« tour précédent »*, *« tour courant »*, *« tour suivant »* et *« autre »*.

5.2.2. *Calcul des scores*

Afin d'éviter l'utilisation d'informations inutiles ou bruitées, les scores $P(o_j, t_r)$ pour $r = k-1, k, k+1$ sont filtrés. Les trois seuils α_r , à partir desquels les scores $P(., t_r)$ sont pris en compte, sont appris à partir du corpus de développement. Leurs valeurs sont fixées à $0,09$ pour $r = k-1$, $0,2$ pour $r = k$ et $0,2$ pour $r = k+1$. Si $P(o_j, t_r)$ est en dessous du seuil α_r, alors $P(o_j, t_r)$ est mis à 0. Les expériences sur le corpus de développement ont montré que cette précaution évitait l'accumulation de petites erreurs de l'arbre de classification (scores très faibles). De plus, les scores de l'étiquette *« autre »* ne permettant pas d'attribuer un nom complet à une classe, ils sont aussi mis à 0.

Pour l'assignation d'un nom complet e_i à une classe c_l donnée, nous calculons un score pour chaque nom complet e_i, dénoté $s_l(e_i)$, qui n'est autre que la somme des scores concernant les tours de parole de la classe c_l :

$$s_l(e_i) = \sum_{\{(o_j,t_r)|o_j=e_i,\, t_r\in c_l\}} P(o_j, t_r) \quad [1]$$

5.2.3. *Processus de décision*

Comme il a déjà été dit, nous faisons l'hypothèse que la segmentation et le regroupement en classes sont corrects. Le but est maintenant d'attribuer à chaque classe c_l un nom complet e_i.

Notre solution propose de réorganiser le partage des noms complets entre les classes anonymes. Elle s'effectue en deux étapes qui sont répétées jusqu'à ce que toutes les classes aient été nommées, ou jusqu'à ce qu'il n'y ait plus de candidat :

– tri des classes candidates pour un nom complet en fonction des scores ;

– assignation du nom complet à la classe la plus probable.

Plusieurs stratégies peuvent être utilisées pour trier les classes c_l en concurrence pour un nom complet e_i donné. Prendre le score maximal $s_l(e_i)$ semble être la solution la plus naturelle. Mais si l'on veut pouvoir comparer ces scores, il faut les normaliser au préalable. Cela peut poser des problèmes, notamment dans le cas fréquent où un mauvais nom complet avec un score faible mais sans concurrent pourrait être affecté à une classe.

Afin d'éviter ce type d'erreur, nous proposons d'utiliser un compromis, à savoir le produit des scores normalisés et non normalisés. Soit $\mathcal{D} = \{c_l \in \mathcal{C} |\ \forall e_i \in \mathcal{E}, s_l(e_i) = 0\}$ l'ensemble des classes non nommées, alors :

$$\text{SC}_l(e_i) = \frac{s_l^2(e_i)}{\sum_{q=1}^{I} s_l(e_q)} \quad \text{si } c_l \notin \mathcal{D} \quad [2]$$

et

$$\text{SC}_l(e_i) = 0 \text{ si } c_l \in \mathcal{D}. \quad [3]$$

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

Tous les noms complets possibles sont pris en compte *a priori* et triés en fonction de leur score $\text{SC}_l(e_i)$. Premièrement, le nom complet avec le score maximal (noté e_i^*)

est choisi, et si plusieurs classes sont associées au même e_i^*, alors ce nom complet sera assigné à la classe dont le score $SC_l(e_i^*)$ est maximal. Ensuite, tous les noms complets choisis sont supprimés des classes qui n'ont pas encore été nommées.

Un exemple concret est donné dans le tableau 1. Le nom complet « Jacques Derrida » a été assigné à trois classes différentes. Dans cet exemple, c_{13} a le meilleur score et « Jacques Derrida » devrait donc être affecté à c_{13} ; mais le score ne représente que 39 % des scores totaux parmi tous les candidats possibles pour c_{13}, alors que le score pour c_{15} représente 79 %. Finalement « Jacques Derrida » est assigné à c_{15} et d'autres noms complets seront attribués aux classes c_{13} et c_{14}.

Classe	Nom complet e_i^*	$s_l(e_i^*)$	$\mathrm{SC}_l(e_i^*)$
c_{13}	Jacques Derrida	$\mathbf{8,58}$	3,36
c_{14}	Jacques Derrida	1,67	1,09
c_{15}	Jacques Derrida	4,94	$\mathbf{3,88}$

Tableau 1. *Exemple d'une assignation initiale multiple*

Lors de l'itération suivante, les noms complets restants sont examinés de la même manière pour les classes restantes et ainsi de suite, jusqu'à ce que toutes les classes soient nommées ou que la liste des noms complets à attribuer soit vide. Le tableau 2 montre les résultats obtenus pour l'exemple précédent.

Classe	Nom complet e_i^* (1^{re} itération)	2^{e} itération
c_{13}	Jacques Derrida (3,36)	**Nicolas Demorand** (0,99)
c_{14}	Jacques Derrida (1,09)	**Alexandre Adler** (0,30)
c_{15}	**Jacques Derrida** (3,88)	-
c_{16}	**Olivier Duhamel** (0,93)	-

Tableau 2. *Exemple du processus de décision avec deux itérations (décision en gras, scores entre parenthèses).*

5.2.4. *Prise en compte du genre*

Le processus de décision précédent ne prend pas en compte une information disponible et qui peut être déterminante pour la validation du locuteur associé à une classe : le genre des locuteurs. En effet, pour chaque classe, cette caractéristique est disponible car, d'une part, 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), d'autre part, les genres des noms complets extraits de la transcription peuvent être déterminés à travers celui de leur prénom associé. 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.

Pour avoir la connaissance relative aux genres des prénoms, nous utilisons une base de données extraite du Web composée d'environ 20 000 prénoms. 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 du processus de décision, la comparaison des genres est directement incluse dans le calcul des scores (5.2.2, formule [1]) se traduisant par le filtre suivant : 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 $P(o_j, t_r)$ de la formule [1] ne sont pas pris en compte lors du calcul. Soit $g(e_i)$ et $g(t_r)$ les genres (féminin, masculin ou indéterminé) d'un nom complet e_i (ou d'un tour de parole t_r), alors : $(o_j = e_i,\ t_r(o_j) \in c_l$ et $g(e_i) \neq g(c_l)) \Rightarrow P(o_j, t_r) = 0$.

6. Évaluation du système proposé

6.1. *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). La majorité de ces émissions contient essentiellement de la parole lue ou préparée, et peu de parole spontanée : 15 % du corpus correspond à des interventions de personnes parlant au téléphone.

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 de développement a été utilisé pour fixer les différents paramètres du système comme la taille du contexte lexical de l'arbre ou le poids donné aux échantillons lors de l'apprentissage (*cf.* 5.1).

Le corpus d'apprentissage contient 76 heures de données (75 095 segments et 7 416 tours de parole) dans lesquels 11 292 noms complets sont détectés. 755 locuteurs différents interviennent dans ce corpus dont 40 qui n'ont pu être nommés. Le corpus de développement contient 30 heures (27 149 segments et 2 931 tours de parole) dans lesquels 4 533 noms complets ont été détectés. 359 locuteurs différents interviennent dans ce corpus dont 38 qui n'ont pu être nommés. Le corpus d'évaluation contient 10 heures (10 335 segments et 1 082 tours de parole) dans lesquels 1 541 noms complets ont été détectés. 213 locuteurs différents interviennent dans ce corpus dont 24 qui n'ont pu être nommés. 26,5 % de ces locuteurs sont communs au corpus d'apprentissage seul et 28,4 % sont communs aux corpus d'apprentissage et de

	Évaluation
Tour précédent	2,0 % (31)
Tour courant	2,0 % (30)
Tour suivant	16,5 % (255)
Autre	79,5 % (1 225)
Total	100 % (1 541)

Tableau 3. *Répartition des étiquettes sur le corpus d'évaluation, statistiques sur les noms complets (fréquence et effectif).*

développement. Ce découpage correspond au découpage de la campagne d'évaluation officielle ESTER 1 PHASE II 2005.

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. Les références proposées sont d'une grande qualité, les annotateurs ont essayé de nommer le maximum de locuteurs par des identifiants permettant d'en déduire leurs noms complets. Ces noms complets ont été extraits automatiquement et n'ont pas fait l'objet de validation manuelle approfondie.

Le tableau 3 montre la répartition *a priori* des quatre étiquettes calculée pour le corpus de test. L'étiquette *« autre »* est la plus fréquente et représente 79,5 % des cas ; vient ensuite l'étiquette *« tour suivant »* avec 16,5 %, tandis que les deux dernières étiquettes *« tour précédent »* et *« tour courant »* sont les moins fréquentes : environ 2 % chacune.

6.2. *Métriques utilisées*

Le système d'identification nommée est évalué en comparant l'hypothèse générée par celui-ci à la référence distribuée avec le corpus. Cette comparaison met en évidence cinq cas (d'erreur ou de succès) possibles relatifs aux situations suivantes :

– l'identité proposée est correcte (C_1) : le système propose une identité correspondant à celle indiquée dans la référence ;

– erreur de substitution (S) : le système propose une identité différente de l'identité présente dans la référence ;

– erreur de suppression (D) : le système ne propose pas d'identité alors que le locuteur est identifié dans la référence ;

– erreur d'insertion (I) : le système propose une identité alors que le locuteur n'est pas identifié dans la référence ;

– il n'y a pas d'identité (C_2) : le système ne propose pas d'identité et la référence ne contient pas d'identité.

Une mesure de précision et de rappel peut être définie à partir des cinq cas d'erreur :

$$P = \frac{C_1}{C_1 + S + I} \;\; ; \;\; R = \frac{C_1}{C_1 + S + D} \qquad [4]$$

La précision et le rappel peuvent être synthétisés en calculant la F-mesure : $F = (2 \times P \times R)/(P + R)$.

Comme il a été proposé dans (Tranter, 2006), nous complétons ces valeurs 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.

$$Err = \frac{S + I + D}{S + I + D + C_2 + C_1} \;\; ; \qquad [5]$$

Les erreurs peuvent être calculées en terme de durée ou en terme de nombre de locuteurs (classes correctement nommées). 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 %. À noter que ce taux d'erreur ne peut être calculé qu'avec une segmentation et une classification manuelles (donc identique à la référence).

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.

6.3. *Protocole d'évaluation*

Le système décrit dans (Estève *et al.*, 2007), et développé par les auteurs, est utilisé comme système de référence. Ce système ne bénéficie pas du processus de décision décrit précédemment et de la prise en compte des genres. Il permet ainsi d'évaluer l'apport de ces deux modifications. Dans le système de référence, seule l'étiquette ayant la probabilité maximale est prise en compte pour chaque nom complet. Chaque nom complet détecté n'est donc propagé qu'à un et un seul tour de parole, avant d'être ensuite propagé au sein de la classe. Si plusieurs probabilités pour un même nom complet sont présentes au sein de la classe, elles sont additionnées pour donner le score du nom complet au sein de cette classe. La décision globale consiste ensuite à

attribuer, pour une classe, le nom complet dont le score est maximal, sans prise en compte des informations des autres classes.

6.4. *Évaluation du système avec transcriptions manuelles*

6.4.1. *Évaluation du système*

Dans les expériences, il est supposé que le système connaît tous les noms complets susceptibles d'être des locuteurs. Le système de décision utilise cette connaissance pour rejeter les noms complets ne correspondant pas à des locuteurs recherchés. Dans la section 6.4.2, nous présenterons des résultats avec et sans cette connaissance *a priori*. Cette liste est constituée de 1 008 noms complets de locuteurs apparaissant dans les corpus d'apprentissage, de développement et d'évaluation. Cette connaissance est uniquement introduite dans le système de décision, bien que cela soit envisageable de l'introduire aussi au niveau du système automatique de transcription enrichie.

La comparaison entre le système de référence et le système proposé est effectuée sur des transcriptions et segmentations manuelles. C'est-à-dire qu'il n'y a pas d'erreur de segmentation et de classification en locuteurs : toutes les frontières sont justes et tous les tours de parole appartiennent à la bonne classe. De même, la transcription en mots est sans erreur et tous les noms complets de locuteurs sont correctement transcrits. En revanche, la détection des entités nommées est faite en utilisant Nemesis ; elle comporte donc des erreurs.

Comme le montre le tableau 4, le système actuel a une précision plus faible d'environ trois points en absolu, mais un rappel meilleur de plus de douze points. Il est difficile de dire quel est le meilleur système à partir de ces deux valeurs. Le calcul d'un taux d'erreur en durée ($ErrDur$) permet de clarifier la situation : le système proposé obtient dix points d'erreur de moins en absolu que le système de référence. En ce qui concerne le nombre de locuteurs identifiés, le nouveau système étiquette correctement deux fois plus de locuteurs que le système de référence. En effet, si l'on prend en compte le nombre de locuteurs, le taux d'erreur ($ErrLoc$) est d'environ 20 % pour le nouveau système contre 40 % pour celui de référence.

En conclusion, le système proposé obtient de meilleurs taux d'erreur qu'ils soient mesurés en durée ou en nombre de locuteurs.

6.4.2. *Influence de la connaissance* a priori *des noms de locuteurs*

Jusqu'ici, les résultats donnés utilisent une liste de locuteurs potentiels lors du processus de décision et d'évaluation. Le tableau 5 présente les résultats avec et sans connaissance *a priori* sur les noms complets cibles de l'application d'identification nommée. À noter que les systèmes proposés dans (Tranter, 2006 ; Chengyuan *et al.*, 2007) utilisent aussi cette connaissance. La liste contient l'ensemble des participants aux émissions des corpus d'ESTER. Ces personnes sont principalement des personnes publiques comme des journalistes, des politiciens, des artistes ou des sportifs. Cette

	En durée				En nb de locuteurs
Système	**Rappel**	**Précision**	**F-mesure**	**ErrDur**	**ErrLoc**
Référence	70,7 %	92,6 %	0,80	26,6 %	37,4 %
Proposé	83,2 %	89,7 %	0,86	16,6 %	19,5 %

Tableau 4. *Comparaison du système proposé et du système de référence sur le corpus d'évaluation ESTER 1 phase II*

———

Les résultats sont donnés en utilisant la transcription enrichie de référence.
Rappel, ***précision*** *et* ***f-mesure*** *calculés en en durée.*
ErrDur *: taux d'erreur en durée.*
ErrLoc *: taux d'erreur en nombre de locuteurs.*

population est identifiable : leurs noms et prénoms sont bien connus, ils sont présents dans plusieurs émissions, et ils correspondent aux locuteurs principaux en terme de temps de parole.

Dans le § 6.1, nous avons pu noter que seulement 26,5 % des locuteurs du corpus d'évaluation sont communs aux locuteurs du corpus d'apprentissage. Ceci s'explique par un éloignement temporel important des enregistrements. Les données d'apprentissage les plus récentes ont été enregistrées en juillet 2003 pour RTM, les autres radios, représentant la majorité des données, ont été enregistrées entre 1998 et 2000. Les données d'évaluation ont été enregistrées entre octobre et décembre 2004, à plus d'un an des données RTM et à plus de quatre ans des autres radios. En utilisant uniquement les noms complets des locuteurs présents dans le corpus d'apprentissage, l'évaluation porterait sur un nombre très faible de candidats (56 locuteurs sur les 213 présents dans le corpus d'évaluation). Nous avons choisi de nous placer dans le cadre où tous les locuteurs cibles sont connus.

En terme de durée, le taux d'erreur $ErrDur$ augmente d'environ douze points si le système n'utilise pas la liste de noms complets cibles pour filtrer les décisions (*cf.* tableau 5). Les taux d'erreur sont respectivement de 16,7 % et 28,9 % pour le système utilisant la liste de locuteurs et pour le système ne l'utilisant pas. En terme de nombre de locuteurs, la différence est d'environ huit points en absolu en défaveur du système n'utilisant pas la liste. La liste contient 1 008 noms complets, dont 213 présents dans le corpus d'évaluation. Sur les 1 514 occurrences de noms complets détectés, 655 noms complets sont éliminés car ils ne font pas partie de la liste. Lorsque l'on n'utilise pas la liste des locuteurs, des noms de locuteurs qui auraient du être considérés comme « autre » se retrouvent propagés au sein des classes, augmentant ainsi les sources d'erreur. Ces noms éliminés sont majoritairement des personnes citées dans le discours mais ne parlant pas dans l'enregistrement. Quelques erreurs dues à un mauvais étiquetage de Nemesis sont aussi évitées.

	En durée				En nb de locuteurs
Noms complets	**Rappel**	**Précision**	**F-mesure**	**ErrDur**	**ErrLoc**
Connus	83,2 %	89,7 %	0,86	16,7 %	19,5 %
Inconnus	73,61 %	74,27 %	0,74	28,9 %	27,4 %

Tableau 5. *Résultats avec et sans connaissance* a priori *sur les noms complets, évaluation faite sur le corpus d'évaluation ESTER 1 phase II*

———

Les résultats sont donnés en utilisant la transcription enrichie de référence.
***Noms complets connus** : le système de décision connaît les noms complets des locuteurs potentiels.*
***Noms complets inconnus** : le système de décision ne connaît pas les noms complets des locuteurs potentiels.*
***Rappel**, **précision** et **f-mesure** calculés en durée.*
***ErrDur** : taux d'erreur en durée.*
***ErrLoc** : taux d'erreur en nombre de locuteurs.*

6.5. *Vers un système entièrement automatique*

6.5.1. *Évaluation du système avec transcriptions automatiques*

Les résultats présentés dans le tableau 6 correspondent aux expériences utilisant des segmentations et classifications en locuteurs automatiques ou manuelles ainsi que des transcriptions automatiques ou manuelles. Le système de référence et le système proposé ont dans tous les cas été évalués avec une détection des entités nommées automatique.

On constate que plus le système tend vers un système entièrement automatique, plus les performances se dégradent. Pour le système proposé le taux d'erreur en durée $ErrDur$ passe de 16,7 % à 75,2 %, étant ainsi multiplié par plus de 4,5.

Le système proposé est tributaire des erreurs de la transcription enrichie. Cette dégradation des performances provient autant des erreurs de segmentation et classification en locuteurs que des erreurs de transcription. Concernant la classification et la segmentation automatique en locuteurs, 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 qui est de 11,5 %. L'impact des erreurs de transcription est étudié dans la section suivante 6.5.2.

6.5.2. *Influence de la qualité de la transcription*

Les résultats précédents montrent une dégradation des performances lorsque la transcription est automatique. Cette dernière obtient un taux d'erreur sur les mots de 20,5 %. Le tableau 7 compare les résultats d'identification nommée obtenus en uti-

		En durée				En nb de locuteurs
Trans.	**Seg/Class.**	**R**	**P**	**F**	**ErrDur**	**ErrLoc**
Système présenté						
M	**M**	83,2 %	89,7 %	0,86	16,7 %	19,5 %
M	**A**	38,0 %	58,2 %	0,46	58,3 %	-
A	**M**	31,0 %	58,3 %	0,40	62.8 %	70,0 %
A	**A**	18,4 %	42,1 %	0,26	75,2 %	-
Système de référence						
M	**M**	75,7 %	95,3 %	0,84	22,5 %	33,1 %
M	**A**	27,8 %	71,6 %	0,40	66,0 %	-
A	**M**	28,1 %	76,9 %	0,41	63,9 %	74,0 %
A	**A**	15,0 %	75,7 %	0,25	77,34 %	-

Tableau 6. *Système proposé avec 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.*

lisant 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 $ErrDur$ et $ErrLoc$ 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.

	En durée				En nb de locuteurs
Transcription	**Rappel**	**Précision**	**F-mesure**	**ErrDur**	**ErrLoc**
LIUM	31,0 %	58,3 %	0.40	62,8 %	70,0 %
LIMSI	41,0 %	65,1 %	0.50	53,8 %	59,5 %

Tableau 7. *Comparaison des résultats avec deux systèmes de transcription différents sur le corpus d'évaluation ESTER 1 phase II*

———

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

7. Conclusion

La méthode d'identification des locuteurs proposée dans cet article vise à extraire les identités des locuteurs des transcriptions. L'identification est réalisée à l'aide d'un arbre de classification sémantique qui attribue les prénoms et patronymes détectés dans la transcription aux locuteurs s'exprimant dans l'enregistrement. Des travaux antérieurs ont montré que l'utilisation d'un arbre de classification était une solution robuste, aussi bien pour une utilisation sur des transcriptions manuelles que sur des transcriptions automatiques. Dans cet article, nous proposons un nouveau système de décision original qui améliore les performances. Le choix des identités des locuteurs est reporté en fin de processus où tous les noms complets candidats sont mis en concurrence.

Les expériences ont été réalisées sur des émissions radiophoniques en français issues de la campagne d'évaluation ESTER 1 phase II. Le système obtient de très bonnes performances pour le traitement de transcriptions manuelles, en revanche les performances se dégradent fortement lorsque l'identification est réalisée à partir de transcriptions obtenues de manière automatique.

Les travaux futurs se focaliseront sur le traitement des transcriptions automatiques. Les expériences ont montré que toutes les sources d'erreurs en amont augmentaient le taux d'erreur d'identification. Bien qu'une solution soit d'améliorer les systèmes de classification et de segmentation en locuteurs, le système de transcription ainsi que le système de détection des entités nommées, nous préférerons faire collaborer plus étroitement les systèmes de transcription automatique et d'identification nommée afin de prendre en compte ces erreurs.

Remerciements

Les auteurs remercient vivement J.-L. Gauvain et les membres de l'équipe TLP pour avoir mis à leur disposition les transcriptions automatiques du LIMSI issues de la campagne d'évaluation ESTER 1.

Ces travaux sont soutenus par le région des Pays de la Loire dans le cadre du projet MILES ainsi que par l'ANR dans le cadre du projet EPAC.

8. Bibliographie

Barras C., Zhu X., Meignier S., Gauvain J., « Multi-stage speaker diarization of broadcast news », *IEEE Transactions on Audio, Speech and Language Processing*, vol. 14, n° 5, p. 1505-1512, September, 2006.

Bechet F., Nasr A., Genet F., « Tagging Unknown Proper Names Using Decision Trees », *ACL, 38th Annual Meeting of the Association for Computational Linguistics*, Hong-Kong, China, p. 77-84, October, 2000.

Canseco-Rodriguez L., Speaker Diarization in Broadcast News, PhD thesis, École doctorale sciences et sechnologies de l'information des télécommunications et des systèmes, Université Paris XI, July, 2006.

Canseco-Rodriguez L., Lamel L., Gauvain J.-L., « A comparative study using manual and automatic transcriptions for diarization », *Proc. of ASRU, Automatic Speech Recognition and Understanding*, San Juan, Porto Rico, USA, November, 2005.

Chengyuan M., Nguyen P., Mahajan M., « Finding Speaker Identities with a Conditional Maximum Entropy Model », *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, Honolulu, HI, USA, April, 2007.

Deléglise P., Estève Y., Meignier S., Merlin T., « The LIUM Speech Transcription System : a CMU Sphinx III-based System for French Broadcast News », *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September, 2005.

El Khoury E., Meignier S., Sénac C., « Segmentation et regroupement en locuteurs pour la parole conversationnelle », *Proc. of JEP, Journées d'études sur la parole*, Avignon, France, juin, 2008.

Estève Y., Meignier S., Deléglise P., Mauclair J., « Extracting true speaker identities from transcriptions », *Proc. of Interspeech, European Conference on Speech Communication and Technology*, Antwerp, Belgium, September, 2007.

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

Galliano S., Geoffrois E., Gravier G., Bonastre J.-F., Mostefa D., Choukri K., « Corpus description of the ESTER evaluation campaign for the rich transcription of French broadcast news », *LREC, Language Evaluation and Resources Conference*, Genoa, Italy, May, 2006.

Galliano S., Geoffrois E., Mostefa D., Choukri K., Bonastre J. F., Gravier G., « The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News », *Proc. of Eurospeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September, 2005.

Gauvain J.-L., Adda G., Adda-Decker M., Allauzen A., Gendner V., Lamel L., Schwenk H., « Where Are We in Transcribing French Broadcast News ? », *Proc. of Interspeech, European Conference on Speech Communication and Technology*, Lisbon, Portugal, September, 2005.

Grass T., « Typologie et traductibilité des noms propres de l'allemand vers le français », *Traitement automatique des langues*, vol. 41(3), p. 643-670, 2000.

Gravier G., Bonastre J.-F., Galliano S., Geoffrois E., « The ESTER Evaluation Campaign of Rich Transcription of French Broadcast News », *LREC, Language Evaluation and Resources Conference*, Lisbon, Portugal, May, 2004.

Jousse V., Jacquin C., Meignier S., Estève Y., Daille B., « Étude pour l'amélioration d'un système d'identification nommée du locuteur », *Proc. of JEP, Journées d'études sur la parole*, Avignon, France, juin, 2008.

Kuhn R., De Mori R., « The application of semantic classification trees to natural language understanding », *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, n° 5, p. 449-460, 1995.

Mauclair J., Meignier S., Estève Y., « Speaker diarization : about whom the speaker is talking ? », *IEEE Odyssey 2006*, San Juan, Puerto Rico, USA, June, 2006.

NIST, « The Rich Transcription Spring 2003 (RT-03S) Evaluation Plan, (Version 4, Updated 02/25/2003) », , `http://www.nist.gov/speech/tests/rt/2003-spring/docs/rt03-spring-eval-plan-v4.pdf`, February, 2003.

Tranter S. E., « Who really spoke when ? Finding speaker turns and identities in broadcast news audio », *Proc. of ICASSP, International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, Toulouse, France, p. 1013-1016, May, 2006.