TALN 2011, Montpellier, 27 juin – 1er juillet 2011

Qui êtes-vous ? Catégoriser les questions pour déterminer le rôle des locuteurs dans des conversations orales *

Thierry Bazillon[1], Benjamin Maza[2], Mickael Rouvier[2], Frederic Bechet[1], Alexis Nasr[1]
(1) Aix Marseille Université, LIF-CNRS, Marseille, France
(2) Université d'Avignon, LIA-CERI, Avignon, France

Résumé. La fouille de données orales est un domaine de recherche visant à caractériser un flux audio contenant de la parole d'un ou plusieurs locuteurs, à l'aide de descripteurs liés à la forme et au contenu du signal. Outre la transcription automatique en mots des paroles prononcées, des informations sur le type de flux audio traité ainsi que sur le rôle et l'identité des locuteurs sont également cruciales pour permettre des requêtes complexes telles que : « chercher des débats sur le thème X », « trouver toutes les interviews de Y », etc. Dans ce cadre, et en traitant des conversations enregistrées lors d'émissions de radio ou de télévision, nous étudions la manière dont les locuteurs expriment des questions dans les conversations, en partant de l'intuition initiale que la forme des questions posées est une signature du rôle du locuteur dans la conversation (présentateur, invité, auditeur, etc.). En proposant une classification du type des questions et en utilisant ces informations en complément des descripteurs généralement utilisés dans la littérature pour classer les locuteurs par rôle, nous espérons améliorer l'étape de classification, et valider par la même occasion notre intuition initiale.

Abstract. Speech Data Mining is an area of research dedicated to characterize audio streams containing speech of one or more speakers, using descriptors related to the form and the content of the speech signal. Besides the automatic word transcription process, information about the type of audio stream and the role and identity of speakers is also crucial to allow complex queries such as : " seek debates on X ,"" find all the interviews of Y", etc. In this framework we present a study done on broadcast conversations on how speakers express questions in conversations, starting with the initial intuition that the form of the questions uttered is a signature of the role of the speakers in the conversation (anchor, guest, expert, etc.). By classifying these questions thanks to a set of labels and using this information in addition to the commonly used descriptors to classify users' role in broadcast conversations, we want to improve the role classification accuracy and validate our initial intuition.

Mots-clés : Fouille de données orales, Traitement Automatique de la Parole, Annotation de corpus oraux, Classification en rôles de locuteurs.

Keywords: Speech data mining, Automatic Speech Processing, Speech Corpus Annotation, Speaker role classification.

*. Ce travail a été effectué dans le cadre du projet ANR DECODA (2009 CORD 005 01) http://decoda.univ-avignon.fr

1 Introduction

La fouille de données orales est un domaine de recherche visant à caractériser un flux audio contenant de la parole d'un ou plusieurs locuteurs à l'aide de descripteurs liés à la forme et au contenu du signal. Parmi ces descripteurs, le plus important est bien évidemment la transcription automatique en mots des paroles prononcées. Dans le cas de la parole multi-locuteurs, telle qu'on peut en trouver dans les émissions de radio ou de télévision, ces descripteurs peuvent aussi porter sur l'identité ou le rôle du locuteur, mais aussi sur le type de conversations enregistrées : débat, interview, journal d'information, etc. Ces informations peuvent permettre de répondre à des requêtes complexes telles que : "chercher des débats sur le thème X", "trouver toutes les interviews de Y", mais aussi d'aider le processus de transcription automatique en choisissant des modèles adaptés au type d'émissions considéré.

L'étude présentée dans cet article a été réalisée sur le corpus EPAC contenant la transcription et l'annotation d'une centaine d'heures de parole essentiellement spontanée [1]. Il est notamment constitué d'émissions ou de débats radiophoniques tels que *Le Téléphone sonne*, *Quartiers d'Été*, *Sous les étoiles exactement*, *Culture vive* ou *Les Matins de France Culture*. Toutes ces émissions ont été segmentées en locuteurs et transcrites manuellement. En complément de ces annotations, d'autres annotations sur les rôles des locuteurs et les formes interrogatives utilisées ont été ajoutées sur une partie du corpus. Nous avons choisi de nous intéresser spécifiquement au domaine de l'interrogation car il nous semble que la structure même des questions, à l'intérieur d'un discours, est représentative d'un certain type d'oral. Des travaux comme ceux de (Garcia-Fernandez & Lailler, 2008), tendant à définir une "morphosyntaxe de l'interrogation", vont déjà en ce sens. Pour notre part, nous essaierons de voir si les questions peuvent être associées à des classes de locuteurs spécifiques, dans un contexte d'oral radiophonique précis. Un arbre de catégorisation des questions a donc été créé *a priori*, et une vingtaine d'heures de données issues de l'émission *Le Téléphone sonne* ont été étiquetées selon cet arbre.

Nous présentons dans la section 2 une étude descriptive des formes interrogatives et leur répartition en différents types ou catégories. Nous tentons en 3 d'établir un lien entre les types de questions et les rôles des locuteurs qui les ont posées. L'apport de la prise en compte du type des questions posées pour la segmentation automatique en rôles de locuteurs est présentée dans la section 4. Enfin la détection et la classification automatique des questions et leur intégration dans le classifieur en rôle de locuteurs est présentée dans la section 5.

2 Un corpus étiqueté en questions

À l'intérieur du corpus EPAC, l'émission *Le Téléphone sonne* représente près de 20 heures de données, réparties en 32 émissions. Dans chacune d'entre elles, chaque question a été annotée manuellement grâce à un système de balises spécifiques. Les questions ont été catégorisées selon différentes catégories et sous-catégories linguistiques. Tout d'abord, une distinction de premier niveau a été établie entre les questions directes (*comment vas-tu ?*) et les questions indirectes (*je voudrais savoir comment tu vas ?*). Ensuite, à l'intérieur de ces deux ensembles, trois sous-catégories ont été isolées : les interrogations totales, les interrogations partielles et les interrogations alternatives (*vous voulez une réponse précise ou pas ?*). Enfin, un troisième et dernier niveau d'analyse concernant les marqueurs interrogatifs a été pris en compte. Ces marqueurs peuvent être :

– des pronoms interrogatifs (*qui*, *que*)
– des adverbes interrogatifs (*quand*, *comment*, *pourquoi*)
– des déterminants interrogatifs (*quel*, *quelle*)
– des structures complexes (*qu'est-ce que*, *qu'est-ce qui*, *à qui*, *depuis quand*...)
– la forme *est-ce que*
– l'adverbe *si* (je voudrais savoir si vos intervenants sont d'accord ?)
– l'inversion du sujet
– l'intonation seule (*tu viens ?*)

La nomenclature ci-dessus est certes un peu différente de celle proposée notamment par (Stolcke *et al.*, 2000), mais les principales distinctions y sont préservées. Ainsi, notre critère « intonation seule » correspond aux questions « déclaratives » de Stolcke, et ce qu'il nomme les *wh- questions* est ici transposé en pronoms, adverbes

1. nous renvoyons le lecteur à (Estève *et al.*, 2010) pour une description détaillée des tâches de transcription et d'annotation du corpus EPAC

et déterminants interrogatifs. Enfin, les questions avec la locution *est-ce que* ou l'inversion sujet-verbe (soit les interrogations dites « totales ») correspondent aux « yes-no-question » de la terminologie de Stolcke.

Type de questions			Nombre d'occurrences	Fréquence (%)
directe	totale	est-ce que	214	13,76
		intonation	178	11,45
		inversion	154	9,9
	partielle	intonation	404	25,98
		adverbe	198	12,73
		complexe	117	7,52
		pronom	88	5,66
		déterminant	76	4,89
	alternative	inversion	10	0,64
		intonation	5	0,32
		est-ce que	1	0,06
indirecte	totale	si	46	2,96
	partielle	adverbe	26	1,67
		complexe	17	1,09
		déterminant	17	1,09
		groupe nominal	3	0,19
		pronom	1	0,06

TABLE 1 – Répartition des questions par type

Sur les 48 catégories possibles, 17 sont attestées dans le corpus du *Téléphone sonne*. La table 1 présente ces 17 catégories avec leur fréquence absolue et relative. Cette table met en relief plusieurs éléments : en premier lieu, les questions directes sont beaucoup plus nombreuses que les questions indirectes. Cet écart s'explique par le fait que *Le Téléphone sonne* est avant tout un débat, donc avec des propos essentiellement spontanés. En conséquence, les locuteurs posent la plupart de leurs questions de la façon la plus directe qui soit, c'est-à-dire avec un nombre réduit de mots. En conséquence, ils ont très peu recours aux structures telles que *je voudrais savoir si...* ou *je me demandais si....* À la place de *je voudrais savoir si vos invités sont d'accord avec ça*, on trouvera ainsi beaucoup plus souvent la forme oralisée *vos invités sont d'accord avec ça ?*.

Ce dernier exemple nous amène à une autre observation, qui peut être faite à la lecture de la table 1 : les questions avec l'intonation pour seul marqueur interrogatif sont les plus représentées (plus de 37% en cumulant interrogations totales et partielles). L'explication est directement liée à ce que nous écrivions plus haut au sujet du type de parole utilisé dans *Le Téléphone sonne*. En effet, la parole spontanée est souvent synonyme de structures interrogatives réduites, notamment dans les débats. Pour être plus en phase avec des situations discursives souvent mouvementées (parole simultanée, locuteurs qui se coupent la parole, temps d'antenne réduit), les locuteurs se doivent d'être les plus concis possibles. En conséquence, en plus d'utiliser essentiellement des structures interrogatives directes, ils font aussi abstraction des marqueurs interrogatifs traditionnels (*Est-ce que vous y croyez ?* deviendra *vous y croyez ?*, par exemple). Il en va de même pour les constructions syntaxiques avec inversion du sujet (*voulez-vous poser une autre question ?*), qui sont assez peu nombreuses dans notre relevé. Bien qu'elles ne rallongent pas à proprement parler la longueur d'une question (aucun mot supplémentaire n'est nécessaire pour passer de *vous avez une autre question ?* à *avez-vous une autre question ?*), elles ne correspondent pas au modèle canonique français sujet + verbe + complément.

3 Un corpus étiqueté en rôles

Le rôle d'un locuteur correspond à son statut et à sa fonction dans une émission donnée : présentateur principal, invité, consultant, journaliste hors studio, etc. Identifier le rôle de chaque locuteur est une étape indispensable à la compréhension d'une émission. Le format de chaque émission définit le nombre et les rôles de chaque intervenant. Pour cette étude, nous avons privilégié une segmentation en rôles relativement générique. Par exemple, les éventuels changements de rôle d'un locuteur à l'intérieur d'une même émission n'ont pas été pris en compte. Cette distinction, pertinente dans le cadre d'un découpage en actes de dialogue, ne nous a pas semblé nécessaire ici. Ainsi, dans l'émission *Le Téléphone sonne*, nous avons donc identifié 4 rôles de locuteurs :

1. le présentateur : c'est l'animateur de l'émission, il a comme rôle de distribuer la parole aux différents intervenants en animant le débat ;
2. les experts : ce sont les invités, sur le plateau ou par téléphone, ils ont pour rôle à la fois de répondre aux questions des auditeurs mais aussi de débattre entre eux sur le sujet du jour ;
3. les auditeurs : toujours au téléphone, ils sont sélectionnés avant l'émission et sont appelés pour qu'ils puissent poser leur questions, sans pour autant participer au débat ;
4. le rapporteur : il a pour rôle de lire les questions écrites des auditeurs ; ce rôle est généralement assumé par le présentateur.

Nous présentons dans les lignes suivantes une analyse du type des questions exprimées par les différents rôles de locuteurs. C'est cette étude qui nous a conduit à utiliser les questions comme indices ou marqueurs du rôle des locuteurs, comme nous le verrons dans la partie expérimentale de cet article.

3.1 Les questions comme marqueurs des rôles des locuteurs

La table 2 présente les premiers résultats de l'annotation des rôles des locuteurs, couplée avec celle des questions. Il apparaît ainsi que c'est le présentateur qui pose plus de la moitié des questions lors d'une émission comme *Le Téléphone sonne*. Cette prédominance tient essentiellement à sa fonction de *médiateur* que nous avons évoquée précédemment, et que nous expliciterons plus précisément avec d'autres chiffres dans le paragraphe suivant.

Rôle	Nombre de questions	Fréquence (%)
Présentateur	791	50,87
Expert	323	20,77
Auditeur	307	19,74
Rapporteur	134	8,62
TOTAL	1555	100

TABLE 2 – Répartition du nombre de questions en fonction du rôle des locuteurs

Il est également intéressant de constater que les auditeurs, pourtant supposés être au cœur du programme, posent en moyenne moins de questions que les experts présents en studio lors de chaque émission. Toutefois, ces chiffres doivent être nuancés par quelques précisions. Tout d'abord, les questions de la catégorie *rapporteur* peuvent être associées à celles des auditeurs, dans la mesure où ce sont eux qui les rédigent puis les envoient au standard de l'émission. Elles sont certes lues par le présentateur, mais il n'en est jamais l'auteur. Ainsi, on peut considérer que près de 30% des questions de notre corpus sont, directement ou non, posées par des auditeurs.

Ensuite, beaucoup de questions de la catégorie *expert* sont des interrogations n'attendant pas véritablement de réponses. Sans être rhétoriques (au sens linguistique du terme), elles permettent plutôt au locuteur d'étayer sa réponse en s'appuyant plus ou moins explicitement sur la question qui lui a été posée :

Quand tout cela va cesser ? Ça madame je n'en sais rien, mais ce qui sûr...
Alors est-ce qu'il doit démissionner ? Oui, bien sûr, puisque la situation actuelle du pays...

Enfin, et c'est souvent le cas dans les émissions dites de débat, le temps de parole alloué à chaque participant est loin d'être équitablement réparti. Les auditeurs, n'étant pas présents physiquement sur le plateau, sont les premiers à être coupés, interrompus ou parfois même exclus du débat faute de temps. Ils ne posent d'ailleurs qu'une seule voire deux questions, jamais plus. À l'inverse, les experts bénéficient d'une liberté d'expression très large. Celle-ci leur permet certes de répondre aux auditeurs, mais aussi d'étayer le débat en se posant mutuellement de nombreuses questions :

Monsieur XX, vous pensez vraiment que ça va changer les choses ?
Comment pouvez-vous en être sûrs, madame YY ?

Afin d'envisager une analyse plus fine, la table 3 indique, pour chaque type de questions recensées dans notre corpus, leur répartition selon le rôle des locuteurs. Ces résultats sont notamment l'occasion de revenir sur la prépondérance du rôle du présentateur, que nous évoquions précédemment. Comme on le voit, celle-ci est majoritairement due aux questions ayant l'intonation pour seul marqueur interrogatif. En effet, en tant que responsable

des débats qu'il instaure, le présentateur du *Téléphone sonne* est celui qui distribue la parole à ses invités et à ses auditeurs. Pour ce faire, il recourt abondamment à des questions elliptiques tantôt partielles (*sur la guerre en Irak, votre sentiment ?*), tantôt totales (*vous êtes d'accord avec ça, monsieur Gorce ?*). Ces deux catégories réunies représentent ainsi près de 70% des questions assimilées au rôle de présentateur. Les 30% restants concernent principalement des questions qui sont posées au début de chaque émission, lors d'un monologue servant à introduire le sujet du jour. Manifestement écrites, elles sont toujours très structurées sur le plan grammatical, que ce soit avec la locution *est-ce que*, l'inversion du sujet ou des pronoms ou adverbes interrogatifs.

Type question			Présentateurs (%)	Experts (%)	Auditeurs (%)	Rapporteurs (%)
directe	totale	est-ce que (214)	28,5	34,58	33,64	3,27
		intonation (178)	80,9	16,3	2,81	0
		inversion (154)	33,12	7,14	21,43	38,31
	partielle	intonation (404)	98,51	1,24	0,25	0
		adverbe (198)	18,69	43,94	21,72	15,66
		complexe (117)	21,37	48,72	19,66	10,26
		pronom (88)	34,1	35,23	19,32	11,36
		déterminant (76)	44,74	19,74	19,74	15,79
	alternative	inversion (10)	20	70	0	10
		intonation (5)	80	20	0	0
		est-ce que (1)	100	0	0	0
	global (1445)		54,46	21,94	14,46	9,13
indirecte	totale	si (46)	4,35	2,18	93,48	0
	partielle	adverbe (26)	0	3,85	88,46	7,69
		complexe (17)	5,88	11,76	82,35	0
		déterminant (17)	0	11,76	88,24	0
		groupe nominal (3)	0	0	100	0
		pronom (1)	100	0	0	0
	global (110)		3,64	5,45	89,1	1,82

TABLE 3 – Répartition du type de questions en fonction du rôle des locuteurs

Les experts, outre le fait d'occuper la deuxième place au niveau du nombre global de questions posées (table 2), utilisent également de nombreuses structures interrogatives différentes. S'il n'y a pas une catégorie aussi dominante que chez les présentateurs, on notera toutefois qu'ils utilisent beaucoup les questions directes à base d'adverbes, de pronoms, de *est-ce que* et de structures complexes (pronoms ou déterminants : *qu'est-ce que*, *qu'est-ce qui*, *duquel*, *lequel*, etc.). Cela témoigne d'un certain soin quant à la formulation et l'expression, puisque les questions basées sur l'intonation sont ici beaucoup moins employées. Cela est d'autant plus remarquable que dans un cadre énonciatif spontané, les structures les plus simples sont souvent les plus utilisées (*vous pensez vraiment que...* en lieu et place de *pensez-vous vraiment que...* ou *est-ce que vous pensez vraiment que...*). Mais il ne faut pas oublier que nous sommes ici dans le cadre d'une émission radiophonique écoutée, jugée et aussi soumise à des directives éditoriales précises. Les personnes qui y participent n'apportent donc sans doute pas le même soin à leurs propos dans la vie quotidienne, ni même peut-être lorsqu'elles s'expriment dans d'autres medias.

C'est sans doute ce qui explique que le constat soit sensiblement identique du côté des auditeurs, où l'on constate que les questions sans marqueur interrogatif grammatical sont moins fréquentes encore. À l'inverse, la structure *est-ce que* est fortement utilisée par les auditeurs, de même que les constructions indirectes (quel que soit leur marqueur interrogatif). Les incises telles que *j'aimerais savoir*, *j'aurais voulu savoir* ou *je voulais savoir* sont en effet presque exclusivement employées par les auditeurs, sans doute parce qu'elles permettent une sorte de transition entre l'invitation à la prise de parole du présentateur (*Posez votre question*, *nous vous écoutons*) et la question à proprement parler, qu'il serait assez brutal de formuler au style direct. Ainsi, en réponse aux deux exemples ci-dessus, on trouvera beaucoup plus souvent des formes comme *je voulais savoir si la droite avait une chance* plutôt que *est-ce que la droite a une chance ?*. De leur côté, les questions commençant par la locution *est-ce que* sont souvent précédées d'un témoignage de l'auditeur, plus ou moins long, mais qui lui permet de contextualiser sa demande (*j'ai lu dans un journal que [...] Est-ce que vos experts sont d'accord avec ça ?*). De façon plus générale, et pour en revenir à notre idée de départ, les questions posées par les auditeurs sont elles aussi particulièrement soignées sur le plan de la syntaxe, d'une part parce qu'elles sont présélectionnées par le standard du *Téléphone sonne*, et d'autre part parce que beaucoup de personnes les écrivent avant de les lire à l'antenne, craignant de les oublier ou de mal les formuler sinon.

4 Segmentation automatique en rôles de locuteurs

La segmentation automatique en rôles de locuteurs est une tâche relativement nouvelle qui vient en complément des tâches de segmentation en locuteurs effectuées en préalable de tout processus de transcription automatique de parole. En effet, la parole étant un flux, il convient de la segmenter en *tours de parole* correspondant à chaque locuteur de la conversation à transcrire, puis en segments (correspondant généralement à des groupes de souffle) sur lesquels les processus de transcription automatique sont appliqués. La segmentation en rôle correspond à une tâche d'étiquetage des tours de parole attribués aux différents locuteurs en fonction d'une liste de rôles possibles dans le document sonore à traiter.

Une des première études sur le sujet a été publiée en 2000 (Barzilay *et al.*, 2000) et de nombreuses études récentes ont popularisé cette tâche dans la communauté du traitement automatique de la parole (Bigot *et al.*, 2010; Hutchinson *et al.*, 2010; Yaman *et al.*, 2010; Damnati & Charlet, 2011). Les approches diffèrent de par le nombre de rôles considérés (de 3 à 6), le type d'émissions (débats, reportages, interviews, etc.) et le niveau de segmentation utilisé pour l'évaluation : les segments, les tours de parole ou bien directement les locuteurs. Différents types de paramètres sont utilisés dans la phase de classification : paramètres acoustiques et prosodiques (Bigot *et al.*, 2010) ; paramètres lexicaux (Barzilay *et al.*, 2000; Hutchinson *et al.*, 2010) ou encore combinaison des deux (Damnati & Charlet, 2011).

Les études se distinguent également par le degré de supervision nécessaire à la production des paramètres utilisés dans l'étape de classification : depuis une supervision complète en utilisant les segmentations et transcriptions manuelles des émissions comme dans (Yaman *et al.*, 2010) ; ou bien sans supervision aucune avec des processus automatiques de segmentation et de transcription comme dans (Damnati & Charlet, 2011).

Par rapport aux études précédentes, nous nous proposons ici d'apporter un nouveau type de paramètres afin de caractériser les rôles de locuteurs, basé sur la catégorisation des questions présentée dans le paragraphe précédent. Le but de cette étude est de valider l'intérêt de ces paramètres grâce à des expériences contrastives que nous allons effectuer sur le corpus EPAC précédemment décrit. Les expériences sont faites sur les segmentations et les transcriptions manuelles du corpus EPAC afin de démontrer l'intérêt de notre approche indépendamment des erreurs faites durant les étapes de transcription par un système automatique. Néanmoins il nous restera dans une prochaine étude à valider ces résultats en montrant qu'ils restent pertinents même en présence d'erreurs de segmentation et de transcription.

4.1 Un classifieur pour la segmentation en rôle

Nous utilisons ici une méthode de classification supervisée basée sur un algorithme de combinaison de classifieurs simples (méthode de *boosting* dans la terminologie de l'apprentissage automatique). Ce type de classifieur discriminant a donné des résultats comparables aux approches basées sur les machines à vecteur de support (SVM) sur un grand nombre de tâches de classification tout en apportant un certain nombre d'avantages : d'une part une grande liberté est donnée dans la définition des classifieurs simples, ce qui permet de prendre en compte très facilement des paramètres hétérogènes tels que des symboles, des séquences de symboles ou bien encore des valeurs numériques ; d'autre part il est possible de connaître facilement quels classifieurs simples ont été choisis comme étant les plus discriminants pour la classification durant la phase d'apprentissage, et quel est leur poids dans le modèle final.

Dans toutes nos expériences l'implémentation ICSIBOOST (Favre *et al.*, 2007) de l'algorithme AdaBoost a été choisie comme méthode d'apprentissage et de classification. Ce classifieur est appliqué à chaque tour de parole de notre corpus en utilisant la segmentation en locuteur de référence produite manuellement. Il calcule un score pour chaque rôle possible pour chaque segment. En choisissant l'hypothèse ayant reçu le score maximum, nous obtenons une classification en rôle des différents tours de parole des locuteurs, chaque tour étant classé indépendamment des autres tours de parole du même locuteur.

Nous avons testé 3 types de paramètres pour cette phase de classification :

– la durée du tour de parole : ce paramètre est relatif à la structure de la conversation, il est pertinent car les différents rôles sont souvent caractérisés par des temps de parole très différents ;

– les 2-gram de mots : les choix lexicaux sont bien évidemment des paramètres majeurs dans l'attribution des rôles aux locuteurs ; nous considérons ici toutes les séquences de 2-gram de mots ;

– les labels des questions présentes dans le tour de parole ; ces labels représentent à la fois le nombre de questions se trouvant dans le tour, mais aussi leurs caractéristiques (directe/indirecte, totale/partielle, type).

Il est à noter que nous n'utilisons pas ici d'optimisation globale de la segmentation en rôles sur toute l'émission, ni d'informations connues *a priori* sur la structure de cette émission. notre but est d'effectuer une expérience contrastive de classification en rôle, sans pour autant chercher à obtenir les meilleurs taux possibles de classification sur ce corpus. En effet on peut grandement améliorer les résultats en utilisant des connaissances *a priori* sur le format de l'émission telles que : *les auditeurs sont toujours au téléphone* ; *le présentateur est toujours celui qui parle en premier* ; *après un auditeur il y a toujours une reprise de la parole du présentateur* ; etc. Ces informations relatives à la structure connue de l'émission *Le Téléphone Sonne* sont ignorées dans nos expériences.

4.2 Protocole expérimental et premiers résultats

Etant donné le nombre limité d'émissions *Le Téléphone Sonne* dans le corpus EPAC, nous avons utilisé un protocole expérimental basé sur la validation croisée par la méthode du *Leave-One-Out*. Ce protocole consiste, sur un jeu C de n exemples à classer, à retirer un exemple e du jeu d'exemples, à apprendre un classifieur B sur l'ensemble $C - \{e\}$, puis à tester B sur l'exemple e pour obtenir l'hypothèse e' que l'on ajoute à l'ensemble C', initalement vide. À l'issue de n itérations de cet algorithme, l'ensemble C' contient tous les exemples de C avec les hypothèses prédites par les n classifieurs. En comparant les hypothèses prédites dans C' aux hypothèses de référence de C nous obtenons une estimation de la qualité du processus de classification sur l'ensemble du corpus, sans le problème du biais de la sélection de corpus séparés pour l'apprentissage et le test.

Nous avons adapté le principe du *Leave-One-Out* à notre corpus d'émissions de la manière suivante :

– le corpus C contient 32 enregistrements de l'émission *Le Téléphone Sonne* : $C = \{e_1, e_2, \ldots, e_{32}\}$;
– à chaque itération i on sélectionne l'émission e_i comme étant le corpus de test T_i, l'émission e_{i+1} comme étant le corpus de développement D_i et les 30 émissions restantes $A_i = C - \{e_i, e_{i+1}\}$ constituent le corpus d'apprentissage A_i ;
– un classifieur B_i est entraîné sur les tours de parole du corpus A_i ; le nombre d'itérations de l'algorithme de boosting est choisi sur D_i et enfin les tours de parole du corpus T_i sont étiquetés automatiquement par B_i et rangés dans T'_i ;
– à l'issue des 32 itérations, le corpus $C' = \bigcup_{i=1}^{32} T'_i$ contient toutes les hypothèses de classification en rôles des tours de parole du corpus C.

En comparant les annotations manuelles de C et celles automatiques de C', nous pouvons évaluer la qualité de nos prédictions selon plusieurs métriques soit au niveau des tours de parole, soit au niveau des locuteurs. Étant donné que la répartition en rôles n'est pas uniforme (le nombre de tours de parole du présentateur est bien supérieur à celui des auditeurs ; inversement il y a bien plus d'auditeurs différents que de présentateurs), les métriques utilisées sont la précision, le rappel, la F-mesure [2] pour chaque type de rôles en complément de l'erreur totale de classification.

Notre première série d'expériences vise à conforter notre hypothèse initiale concernant la pertinence de la clas-

2. La précision, le rappel et la F-mesure sont calculées de la manière suivante :
   – Soit un échantillon $e \in C$ correspondant à un tour de parole (ou à un locuteur selon le niveau d'évaluation choisi) avec $r = ref(e)$ l'étiquette en rôle de référence contenue dans C et $r' = hyp(e)$ l'étiquette hypothèse prédites par les classifieurs B et contenue dans C'. Nous avons $r, r' \in \{$*présentateur,auditeur,expert,rapporteur*$\}$.
   – Si $r = r'$ alors *correct(r)*=*correct(r)*+1
   – Si $r \neq r'$ alors :
     – *erreur_totale*=*erreur_totale*+1
     – *suppression(r)*=*suppression(r)*+1
     – *insertion(r')*=*insertion(r')*+1
   – La mesure de précision pour l'étiquette r est : $P(r) =$ (*correct(r)* $\times$ 100) $\div$ (*correct(r)+insertion(r)*)
   – La mesure de rappel pour l'étiquette r est : $R(r) =$ (*correct(r)* $\times$ 100) $\div$ (*correct(r)+suppression(r)*)
   – La F-mesure pour l'étiquette r est : $F(r) = (P \times R \times 2) \div (P + R)$
   – La mesure d'erreur totale est définie par : $E = erreur\_totale \div |C|$

sification des questions pour caractériser les rôles des locuteurs. Pour cela nous avons effectué une expérience contrastive consistant à ajouter dans la liste des classifieurs simples utilisés par l'algorithme d'apprentissage, des paramètres liés à la présence ou non de questions dans les tours des locuteurs, puis des paramètres sur la forme et le type des questions posées. Nous obtenons 5 expériences contrastives définies de la manière suivante :

1. *durée+2-grams* : les seuls classifieurs utilisés ici sont l'absence ou la présence de bigrammes de mots dans les transcriptions des tours de parole ainsi qu'un classifieur sur la durée des tours de parole.
2. *durée+2-grams+question* : on ajoute aux classifieurs précédents un classifieur sur la présence ou l'absence de questions dans un tour de parole.
3. *durée+2-grams+question+directe/indirecte* : le label *directe/indirecte* est ajouté aux étiquettes *question*.
4. *durée+2-grams+question+directe/indirecte+totale/partielle* : même chose mais avec l'indication de portée de la question.
5. *durée+2-grams+question+directe/indirecte+totale/partielle+type* : on considère maintenant tous les types de questions, tels qu'ils sont définis dans le tableau 1.

paramètres	nb tests	durée+2-grams (1)	+question (2)	+directe/indir. (3)	+totale/part. (4)	+type (5)
F(auditeur)	*500*	66,4	67,0	66,9	65,4	**67,1**
F(expert)	*1443*	73,7	**74,6**	74,0	73,9	74,2
F(présentateur)	*1860*	81,2	81,5	81,3	81,5	**81,9**
F(rapporteur)	*163*	37,2	42,2	41,9	36,4	**57,3**
Erreur totale (*E*)	*3966*	24,7%	23,9%	24,2%	24,5%	**23,5%**

TABLE 4 – Résultats sur l'étiquetage de chaque tour de parole (annotation manuelle des questions et types de questions)

paramètres	nb tests	durée+2-grams (1)	+question (2)	+directe/indir. (3)	+totale/part. (4)	+type (5)
F(auditeur)	*220*	88,2	89,6	89,2	89,7	**90,4**
F(expert)	*118*	83,9	83,0	82,4	81,2	**84,4**
F(présentateur)	*35*	66,7	67,4	67,3	64,7	**68,8**
F(rapporteur)	*27*	45,7	45,7	36,4	36,4	**74,4**
Erreur totale (*E*)	*400*	17,8%	17,3%	18,0%	18,5%	**15,0%**

TABLE 5 – Résultats sur l'étiquetage de chaque locuteur (annotation manuelle des questions et types de questions)

Les résultats sont donnés dans le tableau 4 pour les tours de parole et dans le tableau 5 pour les locuteurs. Les résultats sur les locuteurs sont obtenus à partir de l'étiquette en rôle majoritaire de tous les tours de parole de ce même locuteur dans une émission donnée. Comme nous pouvons le voir, l'introduction du classifieur binaire *question/non question* améliore légèrement les résultats de classification en rôle, par contre l'ajout des labels *directe/indirecte* et *totale/partielle* n'améliore pas, voir dégrade les résultats.

Les meilleurs résultats sont obtenus en rajoutant le type des questions dans les paramètres de classification, ce qui conforte les analyses descriptives faites à partir de la table 3. On obtient une réduction significative de l'erreur totale, à la fois sur les tours de parole et les locuteurs, grâce à cette catégorisation. Ces résultats valident notre hypothèse initiale sur la pertinence des formes interrogatives pour caractériser les rôles des locuteurs dans des conversations.

Cependant, dans une perspective de réalisation d'un système entièrement automatique, il appartient maintenant de vérifier dans quelle mesure le type d'une question peut être déterminé automatiquement, et quel est l'impact des inévitables erreurs d'étiquetage en questions et en types de questions sur la tâche de segmentation en rôle. Comme précisé en début de paragraphe, nous utilisons dans cette étude les segmentations en locuteurs, en tours de parole ainsi que les transcriptions de référence (manuelles) de notre corpus. Nous limitons ainsi l'étude aux seules erreurs d'étiquetage en question, l'impact des erreurs de transcription et de segmentation en locuteurs est l'objet d'une étude en cours. Le paragraphe suivant présente une méthode d'étiquetage de questions dans des transcriptions de parole, utilisant à la fois des paramètres lexicaux et prosodiques. Les résultats de la segmentation en rôle utilisant ces étiquettes automatiques sont présentés dans le paragraphe 5.3.

5 Détection et classification automatique des questions

La tâche de détection automatique de questions dans des énoncés oraux a principalement été abordée dans des corpus de parole conversationnelles (Yuan & Jurafsky, 2005) et des enregistrements de réunions (Boakye *et al.*, 2009). Dans les deux cas les études se basent sur une segmentation *a priori* des énoncés, effectuée manuellement sur les transcriptions de référence. La tâche revient à une classification binaire des segments de parole : segment interrogatif ou affirmatif. Elle constitue ainsi une sous-tâche d'un étiquetage plus général des conversations en *actes de dialogue* qui consiste à segmenter un dialogue en unités discursives telles que : affirmation, question, appréciation, confirmation, négation, etc. Différentes listes d'actes de dialogue ont été proposées, comme par exemple la liste *DAMSL* (Core & Allen, 1997). Les paramètres utilisés sont principalement des indices lexicaux, prosodiques, également couplés à une analyse syntaxique dans (Boakye *et al.*, 2009).

Nous allons enrichir cette tâche dans nos expériences en rajoutant à cette classification binaire la classification en types de questions, en considérant les 8 types de questions suivants : *adverbe*, *complexe*, *déterminant*, *est-ce-que*, *inversion*, *pronom*, *si* et *intonation*. Les marqueurs des 7 premiers types de question sont des marqueurs "syntaxiques" dans la mesure où c'est la structure syntaxique des énoncés qui permet de les considérer comme des questions. Pour le dernier type, *intonation*, ce sont uniquement des marqueurs prosodiques qui permettent de qualifier les énoncés. Deux types de traitement ont donc été mis en oeuvre sur ces deux familles de questions : un classifieur basé sur des marqueurs syntaxiques, un classifieur basé sur des marqueurs acoustiques.

5.1 Caractérisation des questions avec marqueurs syntaxiques

Nous avons utilisé pour les 7 types de questions avec marqueurs syntaxiques la même méthodologie que pour la classification en rôle présentée dans le paragraphe 4. Cette fois chaque échantillon d'apprentissage ou de test correspond à un segment ou à une « phrase » manuellement annoté sur les transcriptions de référence du corpus. Est considérée comme phrase toute séquence de mots, à l'intérieur d'un tour de parole, séparée par un signe de ponctuation forte (point, point d'interrogation, point d'exclamation) ajouté par les annotateurs humains durant la phase de transcription manuelle [3]. Nous avons, sur les 32 émissions *Le Téléphone Sonne* de cette étude, un ensemble de 13224 segments dont 973 questions avec marqueurs syntaxiques et 562 questions *intonation*. Le classifieur ICSIBOOST a été entraîné sur ces échantillons en utilisant la méthodologie de validation croisée *Leave-One-Out* présentée dans le paragraphe 4.2. Les résultats sont présentés dans la table 6 en utilisant comme seuls paramètres des bigrammes de mots. Nous n'avons pas pour l'instant intégré d'informations relatives aux structures syntaxiques des énoncés, ce travail fait partie d'une étude en cours.

Classification	nb de segments	Précision	Rappel	F-mesure
segments interrogatifs	*995*	94,2	85,1	89,4
autres segments	*12229*	98,8	99,6	99,2
question type=adverbe	*223*	96,1	87,9	91,8
question type=complexe	*139*	79,0	67,6	72,9
question type=déterminant	*99*	87,6	78,8	83,0
question type=est-ce-que	*209*	96,7	97,6	97,1
question type=inversion	*159*	82,5	53,5	64,9
question type=pronom	*94*	80,9	58,5	67,9
question type=si	*45*	83,3	66,7	74,1
segments non interrogatifs	*12229*	98,4	99,7	99,1

TABLE 6 – Résultats sur l'étiquetage des segments en question et type de question

Comme nous pouvons le voir le taux de détection moyen des questions est satisfaisant (environ 90% de F-mesure), cependant de grandes disparités sont constatées selon le type de questions. De manière assez prévisible les questions de type *inversion* et *pronom* sont les plus difficiles à classer, ce qui justifie l'intérêt de disposer de paramètres liés à la structure syntaxique des énoncés et non pas seulement à leur lexicalisation. Cependant l'analyse syntaxique automatique de l'oral spontané est encore un domaine de recherche largement ouvert.

3. Bien évidemment tout symbole de ponctuation a été supprimé des transcriptions des segments dans toutes les expériences de classification

5.2 Caractérisation des questions avec intonation

Nous avons choisi de traiter les questions "purement" intonatives de notre corpus uniquement avec des paramètres prosodiques basés sur la courbe de fréquence fondamentale, ou $F0$ (Yuan & Jurafsky, 2005; Quang *et al.*, 2007). Ces paramètres sont obtenus directement à partir du signal de parole avec une fenêtre temporelle de 10 millisecondes. À partir de cette courbe nous proposons d'extraire un ensemble de 15 paramètres divisés en 3 classes : paramètres statistiques (6 paramètres), paramètres de trajectoire (5 paramètres) et paramètres de formes (4 paramètres). Voici une description rapide de ces paramètres qui sont calculés sur la fin de chaque phrase sur des périodes de 300 et 700 millisecondes :

– **Statistique** : nous avons 6 paramètres numériques sur la courbe de F0 : minimum, maximum, intervalle, moyenne, médiane et déviation standard de la F0 sur nos fenêtres de 300 et 700 millisecondes.
– **Trajectoire** : ces 5 paramètres décrivent si la courbe de fréquence fondamentale monte ou descend en fin de phrase.
– **Forme** : Les 4 paramètres de formes constituent l'une des originalités de cette étude. Ils consistent à modéliser la forme de la courbe de F0 grâce à une interpolation polynomiale Lagrangienne. Différents degrés de polynômes ont été testés et des résultats empiriques ont montré qu'un degré de 2 était satisfaisant pour la tâche. Nous utilisons donc les 3 paramètres a, b, c du polynôme $a * x^2 + b * x + c$ ainsi que l'erreur d'interpolation de la fonction approchée comme quatrième paramètre.

Une fois les 15 paramètres extraits, un classifieur est entraîné, en utilisant le même protocole que décrit précédemment, pour séparer les segments "question" des segments "autre".

question/non question	Précision	Rappel	F-Mesure
Forme+Statistique	0,62	0,37	0,46
Forme+Trajectoire	0,58	0,32	0,41
Statistique+Trajectoire	0,58	0,33	0,42
Combinaison	0,58	0,41	0,48

TABLE 7 – Combinaison des paramètres prosodiques basés sur la F0 pour la classification binaire *question/non question* de segments de parole

Une évaluation de ces paramètres est donnée dans la table 7 sur la classification binaire question/non question des segments du corpus. Comme nous pouvons le voir les meilleurs résultats sont obtenus en combinant les différents paramètres avec une F-mesure d'environ 50%. Dans notre système de classification du type des questions, ce classifieur est utilisé de la manière suivante : si un segment n'est pas considéré comme une question par le classifieur basé sur les marqueurs syntaxiques mais qu'il est classé *question* par le classifieur prosodique, alors le segment reçoit l'étiquette *question intonation*.

5.3 Évaluation sur la segmentation en rôle

niveau	*tours de parole*			*locuteurs*		
paramètres	nb tests	type question (ref)	type question (aut)	nb tests	type question (ref)	type question (aut)
F(auditeur)	*500*	65,9	66,5	*220*	90,2	90,2
F(expert)	*1443*	74,8	73,8	*118*	83,6	83,3
F(présentateur)	*1860*	82,5	81,2	*35*	77,7	71,7
F(rapporteur)	*163*	56,7	52,4	*27*	68,3	57,9
Erreur totale (E)	*3966*	23,1%	24,2%	*400*	14,5%	15,8%

TABLE 8 – Résultats sur l'étiquetage en rôle des tours de parole et des locuteurs. Comparaison annotation manuelle/automatique des questions avec leurs types

La table 8 présente les résultats obtenus sur la tâche de segmentation en rôle en comparant l'utilisation des étiquettes de type de question de référence (manuelles) à celles produites automatiquement par les deux classifieurs présentés dans ce paragraphe. Comme nous pouvons le voir, même si une dégradation est constatée dans les performances à cause des erreurs de détection et de classification des questions, les résultats restent meilleurs que

ceux obtenus sans ces paramètres : -0.5% d'erreur totale pour les tours de parole et -2% d'erreur totale pour les locuteurs.

6 Conclusion

Nous avons proposé dans cette étude une analyse du type des questions exprimées dans un corpus de conversation d'émissions de radio. Nous avons montré que le typage des questions pouvait être un indicateur du rôle du locuteur à l'intérieur de la conversation. Nous avons validé cette hypothèse sur une tâche de segmentation automatique en rôle des locuteurs de notre corpus, en constatant une amélioration significative des résultats après ajout de paramètres liés au typage des questions dans le processus de classification automatique. Enfin nous avons validé la mise en pratique de ces paramètres en montrant qu'on pouvait les obtenir de manière complètement automatique au prix d'une légère dégradation des résultats. Il nous reste cependant à nous attaquer au défi que constitue la segmentation automatique en unité, ou « pseudo-phrases » de l'oral spontané. Se baser uniquement sur les pauses ou les groupes de souffle ne permet pas de segmenter de manière cohérente les énoncés, et les résultats des systèmes de segmentation automatiques en phrases basés sur la prosodie et des indices syntaxiques, s'ils obtiennent des résultats intéressants sur de la parole lue ou préparée, sont encore très insuffisants pour être utilisé directement sur des conversations spontanées. À terme, une solution pourrait consister à utiliser des méthodes et paramètres syntaxiques qui se passeraient d'une segmentation en phrases ou « pseudo-phrases », afin de ne pas être exposé aux limites que sous-entend cette tâche sur la parole spontanée.

Références

BARZILAY R., COLLINS M., HIRSCHBERG J. & WHITTAKER S. (2000). The rules behind roles : Identifying speaker role in radio broadcasts. In *Proc. of AAAI*.

BIGOT B., PINQUIER J., FERRANÉ I. & ANDRÉ-OBRECHT R. (2010). Looking for relevant features for speaker role recognition. In *Proc. of Interspeech*.

BOAKYE K., FAVRE B. & HAKKANI-TÜR D. (2009). Any Questions ? Automatic Question Detection in Meetings. In *ASRU, Merano (Italy)*.

CORE M. & ALLEN J. (1997). Coding dialogs with the DAMSL annotation scheme. In *AAAI Fall Symposium on Communicative Action in Humans and Machines*, p. 28–35 : Citeseer.

DAMNATI G. & CHARLET D. (2011). Robust speaker turn role labeling of tv broadcast news shows. In *ICASSP'2011*.

ESTÈVE Y., BAZILLON T., ANTOINE J.-Y., BÉCHET F. & FARINAS J. (2010). The EPAC corpus : manual and automatic annotations of conversational speech in French broadcast news. In *LREC*, Malta.

FAVRE B., HAKKANI-TÜR D. & CUENDET S. (2007). Icsiboost. `http://code.google.come/p/icsiboost`.

GARCIA-FERNANDEZ A. & LAILLER C. (2008). Morphosyntaxe de l'interrogation pour le système question-réponse ritel. In *RECITAL 2008*.

HUTCHINSON B., ZHANG B. & OSTENDORF M. (2010). Unsupervised broadcast conversation speaker role labeling. In *Proc. of ICASSP*.

QUANG V., BESACIER L. & CASTELLI E. (2007). Automatic question detection : prosodic-lexical features and cross-lingual experiments. In *Proc. Interspeech*, volume 2007, p. 2257–2260.

STOLCKE A., RIES K., COCCARO N., SHRIBERG E., BATES R., JURAFSKY D., TAYLOR P., MARTIN R., ESS-DYKEMA C. & METEER M. (2000). Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, **26**(3), 339–373.

YAMAN S., HAKKANI-TUR D. & TUR G. (2010). Social role discovery from spoken language using dynamic bayesian networks. In *Proc. of Interspeech*.

YUAN J. & JURAFSKY D. (2005). Detection of questions in Chinese conversational speech. In *2005 IEEE Workshop on Automatic Speech Recognition and Understanding*, p. 47–52.