La parole spontanée : transcription et traitement

Thierry Bazillon[*], Vincent Jousse[*], Frédéric Béchet[**], Yannick Estève[*], Georges Linarès[**], Daniel Luzzati[*]

LIUM[*] – Université du Maine, Le Mans
LIA[**] – Université d'Avignon et des Pays de Vaucluse, Avignon

thierry.bazillon@lium.univ-lemans.fr

RÉSUMÉ. *Cet article traite de la parole spontanée, la définissant d'abord sous différents angles et spécificités, puis en envisageant sa transcription de façon diachronique et synchronique. Enfin, par le biais de différentes expériences, réalisées notamment dans le cadre du projet EPAC, nous avons identifié les principaux problèmes que la parole spontanée posait aux systèmes de reconnaissance automatique, et proposons des optimisations en vue de les résoudre.*

ABSTRACT. *This paper deals with spontaneous speech, considering first its specificities, and then its transcription – both diachronically and synchronically. The paper continues by listing the main problems spontaneous speech causes to automatic speech recognition systems, which were identified through several experiments. It ends by suggesting some optimizations to help solve these problems.*

MOTS-CLÉS : *parole spontanée, transcription manuelle, transcription automatique, spécificités de l'oral, système de reconnaissance automatique de la parole.*

KEYWORDS: *spontaneous speech, manual transcription, automatic transcription, oral specificities, automatic speech recognition systems.*

1. Introduction

D'un point de vue énonciatif, la parole spontanée peut se définir comme un « énoncé conçu et perçu dans le fil de son énonciation » (Luzzati, 2004), c'est-à-dire un énoncé produit pour un interlocuteur réel par un énonciateur qui improvise ; cela implique que les corrections ne peuvent se traduire que par un prolongement du message. La parole préparée (celle qu'emploient les journalistes présentant les informations radiophoniques ou télévisées) est une parole produite pour un interlocuteur plus ou moins fictif, par un énonciateur qui en possède la maîtrise, qui est capable de produire des énoncés qui n'ont plus à être repris ou corrigés, ou qui est capable de le masquer. De ce point de vue, on comprend qu'on puisse parler également de parole conversationnelle, non préméditée ou co-construite.

D'un point de vue morphosyntaxique, la parole spontanée se caractérise par deux phénomènes saillants : on y trouve un grand nombre de disfluences (Adda-Decker *et al.*, 2004) et le fenêtrage syntaxique y est particulier. Les fenêtres de cohérence syntaxique (Luzzati, 2004) y sont courtes (empan moyen inférieur à huit « mots »), elles ne sont pas nécessairement conjointes, et elles sont superposables. À l'inverse, la parole préparée tend vers l'écrit, avec des fenêtres de cohérence syntaxique parfois très longues (l'hypotaxe y est importante), conjointes et sans interjection.

D'un point de vue phonologique, la parole spontanée se caractérise par deux phénomènes importants : la disparition des schwas (ou e muet, caduc, central...) et les phénomènes d'assimilation qui en découlent. À titre d'exemple, un mot comme « cheval », suite à la disparition du schwa et à une assimilation, se prononce désormais [Sfal][1], de façon généralement inconsciente pour les locuteurs, mais patente pour un système de reconnaissance automatique de la parole.

C'est pourquoi nous nous proposons, dans un premier temps, de confronter linguistiquement la parole spontanée à la parole préparée, afin d'en faire ressortir les principales spécificités. Puis, après avoir proposé un état des lieux des corpus disponibles, nous nous intéresserons au traitement de la parole spontanée par le biais de diverses expériences, qui ont pour but d'en optimiser la détection et la transcription à l'aide d'un système de reconnaissance automatique de la parole (RAP). Les résultats ainsi obtenus nous permettront notamment d'envisager une typologie des erreurs commises par les systèmes de RAP, et de proposer quelques pistes en vue d'améliorer leurs performances.

1. Les transcriptions entre crochets qui figurent dans cet article sont écrites en alphabet SAMPA (http://www.phon.ucl.ac.uk/home/sampa/).

2. Parole spontanée *vs.* parole préparée

Huit critères (notamment morphosyntaxiques) permettent de caractériser la parole dite « spontanée », c'est-à-dire une parole altérée, variable en débit et en fluidité.

2.1. *Élisions du schwa et assimilations*

Abordons tout d'abord l'élision du schwa et les assimilations qui souvent en résultent. La réalisation (ou non) du schwa est en elle-même un problème complexe sur lequel nombre de linguistes se sont penchés. Celle-ci revêt une importance particulière pour la parole spontanée car elle induit souvent des assimilations portant sur des morphèmes ou structures parmi les plus fréquentes (pronom + verbe, de + nom, notamment).

En premier lieu, « je » + consonne sourde, qui devient [S] : des formes telles que « j'pense » ou « j'crois » deviennent respectivement « ch'pense » et « ch'crois ». Les mêmes arguments s'appliquent également à « de », et dans des proportions presque identiques : des données extraites du corpus ESTER nous ont permis de constater que des expressions comme « pas d'problème » ou « pas d'chance » reviennent à de nombreuses reprises dans la langue parlée et, toujours par effet d'assimilation, sont prononcées « pas t'problème » et « pas t'chance ».

À un degré moindre, on retrouve également l'élision du schwa avec « te », « se » ou « que » + consonne sonore (« on s'donne », « tu t'demandes » ou « qu'vous », prononcés « on z'donne », « tu d'demandes » ou « g'vous »).

Enfin, le cas des formes « le », « me » et « ne » est un peu particulier : les nasales « m » et « n » ne varient pas au contact d'une consonne sourde ou sonore lorsque le « e » est élidé (« je m'fâche », « je n'crois pas »). Quant au « l », le fait que cette lettre soit une consonne liquide fait que par nature, elle se combine facilement avec d'autres consonnes ; ainsi, que ce soit au contact d'une sourde (« l'problème ») ou d'une sonore (« je l'vois bien »), sa prononciation n'est pas modifiée.

2.2. *Autres élisions*

D'autres monosyllabes, comme « tu », « il(s) », « elle » et « vous », sont eux aussi souvent élidés : « t'as », « i' vient », « i' savaient pas », « e' va », voire « 'pouvez pas vous traîner » ou « zêt sûr », avec aspiration de « vous » par le verbe. Tout fonctionne en somme comme si aux personnes 1, 2, 4 et 5, le sujet était marqué par une enclise consonantique droite, et comme si aux personnes 3 et 6 demeurait surtout une opposition masculin/féminin (i/è).

Outre ces cas spécifiques, la principale élision rencontrée concerne la vibrante « r » dans les mots à finale en « -bre », « -cre », « -dre », « -tre » ou « -vre ». Cela est particulièrement flagrant lorsque le mot suivant commence par une consonne : en effet, l'immense majorité des locuteurs, dans un contexte spontané, ne dira jamais « à quatre pattes », mais plutôt « à quat'e pattes », séquence beaucoup plus simple à articuler dans un discours à débit relativement rapide. Dans le corpus ESTER (Galliano, 2005), on trouve ainsi : « novemb'e », « convainc'e », « descend'e », « peut-êt'e », « surviv'e »...

« é » et « è » disparaissent parfois dans « c'était », « c'est-à-dire », « déjà » ou « écoutez », qui deviennent « s'tait », « c't-à-dire », « d'jà » ou « 'coutez ». Parfois, cela fait apparaître un schwa qui, accentué, passe du [@] au [9] : il en va ainsi du démonstratif « cette », parfois prononcé « c't**e** »

« l », dans deux cas précis, peut également être élidée : « plus » ou « je lui » sont parfois réduits à « p'us' » ou « j'ui' » (notons que dans ce cas, le schwa de « je » est lui aussi élidé).

Pour terminer, nous mentionnerons quelques cas d'élisions isolés : « puis », « parce que » et « enfin » deviennent très souvent « p'is », « pac'e que » et « 'fin », avec un sens sans doute différent d'un emploi sans élision. L'expression « tout à l'heure » se transforme quelquefois en « t't à l'heure ». Enfin, certains mots commençant par « at- » ont tendance à voir cette séquence initiale disparaître : « attention » deviendra « 'tention » et « attendez », « 'tendez ».

2.3. *Troncations*

La troncation[2] est un autre phénomène spécifique de la parole spontanée : c'est un mot que le locuteur commence à prononcer puis, pour diverses raisons (principalement le bégaiement ou l'hésitation), ne finit pas. Dans certains cas, le mot tronqué est ensuite complètement prononcé. Cela donne des séquences telles que celles-ci (la troncation est symbolisée ci-dessous par l'emploi de parenthèses) :

« des idées ré() révolutionnaires »
« il était aussi passionné d'avia() d'aviation »
« et la première ém() émission »
« elle ne sera pas premier secrétai() euh secrétaire »
« et ça rebaisse régulière() régulièrement »

2. Certains linguistes, et notamment Berthille Pallaud, préfèrent dans le cas présent utiliser le terme « amorce ». Nous renvoyons par ailleurs le lecteur à ses travaux (Pallaud, 2004, 2006), très complets, pour une analyse détaillée de ce phénomène.

Cependant, il arrive que le mot tronqué ne soit pas repris ensuite : soit le locuteur poursuit alors son énoncé comme s'il n'y avait pas eu troncation (1), soit il le reprend partiellement (2 et 3) ou en totalité, créant ainsi une anacoluthe (4 et 5).

(1) « alors auj() le starsystem s'est emparé de la télé »
(2) « c'est t() vraiment une lettre très émouvante »
(3) « c'est-à-dire que pou() sur un repas que vous vendez sept à huit euros »
(4) « et il y a un truc s() il y a quelque chose de suspect »
(5) « oui et c'est un k() ah oui oui et c'est un canadien »

2.4. *Faux départs*

L'anacoluthe nous amène à parler du faux départ, phénomène assez proche des deux derniers exemples que nous venons de mentionner, mais qui s'en distingue en désignant une interruption à l'intérieur d'un énoncé, et non à l'intérieur d'un mot. La conséquence est toutefois la même : l'apparition d'une rupture de syntaxe puisque le locuteur commence un énoncé qu'il ne finit pas pour y adjoindre un second :

« ça a été lu et c'est () on a la photo »
« il y a dix mille () mais c'était mal »
« j'ai () on a essayé de récupérer tous les éléments »

Il arrive en outre que l'on rencontre des « semi-faux départs », où le second énoncé est en fait le complément d'une partie du premier. Le locuteur corrige son propos initial, mais sans produire une phrase complète, ayant toujours à l'esprit le premier fragment prononcé :

« je voulais vous dire aussi () passer un gros coup de gueule »

2.5. *Répétitions*

La répétition d'un même mot ou d'une même séquence de mots est aussi un signe patent d'un discours spontané. À nouveau, bégaiement et hésitations en sont les deux principaux moteurs. La répétition est parfois étroitement liée à la troncation (6) ; de même, elle peut parfois jouer sur deux mots très proches (7) :

(6) « la l() la lettre de Guy Môquet »
(7) « là c'était le le la le la l'accusation la plus grave »

2.6. *Fenêtrage syntaxique*

Toujours sur le plan morphosyntaxique, la parole spontanée se caractérise également par un phénomène remarquable : les fenêtres de cohérence syntaxique y sont courtes, pas nécessairement conjointes, et superposables. En voici un exemple, tiré du corpus « café »[3], ainsi que la représentation qui lui succède :

> « le défaut qu'ils ont ils ont une chambre pour eux pour payer moins cher et ils prennent un copain ou deux et alors voilà mais les bains qui c'est qui les paye ils payent pour un bain ils payent pas pour trois ah »

(a)	fenêtre normale	[-----]
(b)	fenêtre interrompue	[----<
(c)	fenêtre non initiée	>----]
(d)	fenêtres de bafouillage	[---[---[---
(e)	fenêtres de recherche lexicale	---]---]---]
(f)	fenêtres avec mise en commun ou **apo koïnou** :	
	(f1) mise en commun du segment central	[1a---[1b----1a]---1b]
	(f2) mise en commun du segment gauche	[1---[----1a]---1b]
	(f3) mise en commun du segment droit	[1a---[1b ----]---1]

[1*le défaut qu'ils ont****1]*** */* ***[2****ils ont une chambre* ***[****pour eux****2a]*** */ pour payer moins cher****2b]*** *et //* ***[3****ils prennent un copain ou deux****3]*** *et alors voilà / mais* ***[4****les bains qui c'est qui les paye****4]*** ***[5****ils payent pour un bain****5]*** ***[6*** *ils payent pas pour trois****6]*** */ ah*

Figure 1. *Représentation de l'exemple ci-dessus à partir de la théorie du fenêtrage syntaxique*

2.7. *Morphèmes spécifiques*

D'un point de vue lexical, la parole spontanée se caractérise par l'emploi de morphèmes typiquement oraux tels que « euh » ou « ben » (et ses dérivés) (Luzzati, 1982). Extrêmement nombreux dans les corpus que nous avons constitués, leur rôle est pourtant parfois opaque. Si « euh » indique majoritairement l'hésitation (et est à ce titre souvent employé de façon répétée), les emplois de « ben » sont, quant à eux, beaucoup plus difficiles à cerner : forme oralisée de « bien », conjonction de coordination, adverbe...

> « banalisons le pain comme n'importe quel euh objet »
> « et évidemment l'état euh euh à gauche ou à droite »
> « ben c'est gentil mais »
> « eh ben ton installation est est impeccable »
> « qui euh ben qui va s'avérer être un un apprenti euh formidable »

3. http://www.loria.fr/projets/asila/corpus_en_ligne.html

Notons que ces morphèmes et les analyses s'y rattachant ne sont pas spécifiques à la langue française : l'anglais, par exemple, possède avec la forme « *well* » une expression sémantiquement proche de nos « euh » et « ben » français dans certaines de ses acceptions. Deborah Schiffrin s'y est intéressée dans une étude sur les « *discourse markers* » (Schiffrin, 2001), terminologie plus globale que celle que nous employons ici et qui, outre « *well* », recouvre également des formes comme « *and* » ou « *y'know* ». De même, on pourra lire avec intérêt les travaux de Gisèle Chevalier (Chevalier, 2000), qui propose une étude des emplois de « *well* » en acadien du sud-est du Nouveau-Brunswick, une variante du français.

2.8. *Phénomènes prosodiques*

Enfin, nous terminerons cette étude des spécificités de la parole spontanée en évoquant quelques objets ayant trait à la prosodie : tout d'abord, le mélisme (Caelen-Haumont, 2002a), désignant dans notre champ d'études un allongement syllabique en fin de mot. Très caractéristique de l'oral, où il se veut bien souvent être la marque d'une hésitation, ce phénomène s'est révélé particulièrement efficace lors d'une expérience interne pour détecter des zones de parole spontanée dans de gros corpus audio (Jousse, 2008).

Ensuite, les pauses, et plus précisément leur durée et leur fréquence, sont un autre aspect remarquable de la parole spontanée. En effet, si l'on observe un corpus de parole préparée, et notamment journalistique, on s'aperçoit que les pauses dans le flux de parole y sont généralement peu nombreuses, relativement brèves, et bien souvent liées à la respiration et/ou à la déglutition, plus qu'à un phénomène d'hésitation, par exemple. À l'inverse, les pauses dans un cadre spontané (interviews, par exemple) sont en général beaucoup plus longues et nombreuses : d'une part parce que les locuteurs ne bénéficient pas d'un canevas (prompteur, notes) pour tisser leurs propos, et qu'ils les conçoivent donc au fur et à mesure, ce qui demande des périodes de réflexion, d'autre part parce qu'à l'inverse d'un journaliste, dont le métier sous-entend une réelle aisance pour s'exprimer, les intervenants lors d'interviews ou de témoignages ne sont pas toujours familiarisés avec ces exercices ; il en résulte souvent de longs « blancs », témoins de leurs hésitations.

Le débit phonémique obéit également à cette dualité : dans le cas d'un journal d'information, il varie généralement peu. Lors d'un entretien, et pour les raisons que nous venons d'évoquer, il arrive que le locuteur peine à enchaîner ses propos, s'attarde, puis soudain accélère son flux de parole, au gré de ses idées ou de son état émotionnel.

Enfin, pour clore cette analyse prosodique, nous aborderons l'intonation. Le projet EPAC, qui est présenté en 4.1., nous a permis d'envisager ce phénomène de manière concrète : en effet, l'un des objectifs d'EPAC est de fournir la transcription

annotée d'environ cent heures de parole, majoritairement spontanée. Or, toutes les transcriptions que nous réalisons dans cette optique sont ponctuées et, naturellement, cette ponctuation se base entre autres sur l'intonation. Il est ainsi apparu clairement que l'annotateur éprouvait beaucoup plus de difficultés à ponctuer des propos spontanés que des propos préparés. Certes, la rigueur syntactico-sémantique des propos journalistiques y est pour beaucoup, mais les intonations marquées apportent également au transcripteur des indications non négligeables. Or, celles-ci sont beaucoup moins transparentes lorsqu'il s'agit de parole spontanée, tout simplement parce que, comme nous le disions plus haut, la parole est alors élaborée au fur et à mesure qu'elle est énoncée. Ainsi, le locuteur ne sait parfois pas où le segment de parole qu'il a commencé se terminera, et ne peut donc y adjoindre une quelconque marque intonative. Ou encore, il arrive qu'il le fasse, par exemple en adoptant une intonation descendante pour indiquer la fin d'un énoncé, puis se ravise ensuite et complète celui-ci. Il est alors bien délicat de déterminer ce qui doit régir la décision de l'annotateur : l'intonation, ou la structure phrastique ?

3. Les corpus

3.1. *Historique*

Se lancer dans l'établissement d'un corpus de parole spontanée sous-entend des possibilités d'enregistrement et de traitement postenregistrement importantes. Comment en effet envisager d'analyser un objet dont on ne saurait avoir une quelconque trace ? À cet égard, l'historique des corpus qui nous intéressent est étroitement lié aux avancées technologiques du XX^{e} siècle : Queneau considérait dans *Bâtons, chiffres et lettres* que « l'usage du magnétophone a provoqué en linguistique une révolution assez comparable à celle du microscope avec Swammerdam », et, bien que quelques travaux précurseurs sur le sujet n'aient pu bénéficier d'un tel support, il est incontestable que le fait de pouvoir « capturer » l'oral en a radicalement modifié la perception.

Et pourtant, avant même que ne naisse cette invention, Damourette et Pichon (Damourette et Pichon, 1911-1927), en s'appuyant sur des conversations recueillies auprès d'un médecin, d'une institutrice, etc., avaient dessiné les premiers contours morphosyntaxiques d'une « langue orale » dont la communauté linguistique avait encore pourtant du mal à admettre l'existence. Puis il y a eu Bally (Bally, 1929) et surtout Frei (Frei, 1929) qui s'est attaché à analyser les lettres non parvenues aux soldats de la Grande Guerre. Ces lettres étaient rédigées par des familles souvent peu familières de l'écriture, et le style employé était en conséquence très oralisé.

Mais ce sont véritablement Gougenheim et ses collaborateurs (Gougenheim et al., 1964) qui, s'appuyant sur deux cent soixante-quinze enregistrements sonores, ont révélé par effet de bord la véritable teneur de l'oral spontané : souhaitant avant

tout proposer un équivalent du « *basic English* » (Ogden, 1932), et ainsi favoriser l'apprentissage du français, ils ont effectué une étude quantitative du nombre d'occurrences des formes. Il apparut alors que des mots comme « on », « hein » ou « ben » étaient parmi les plus utilisés de la langue française, ce qu'aucune grammaire de l'époque n'envisageait. Aujourd'hui, certaines d'entre elles se refusent encore à considérer seulement leur existence.

À l'époque où Damourette et Pichon entreprirent leurs recherches, les micro-ordinateurs n'existaient évidemment pas, et les machines à écrire en étaient à leurs balbutiements. Les transcriptions étaient donc réalisées « à la volée », ce qui, on l'imagine aujourd'hui, devait s'avérer fort peu confortable. Les ordinateurs ont certes changé la donne, offrant la possibilité d'avoir recours au traitement de texte, et ainsi de corriger, modifier et surtout sauvegarder sans peine ses travaux. Un grand pas a ensuite été franchi lorsqu'il est devenu possible, d'une part de transférer les données sonores vers un ordinateur (et d'assurer ainsi leur pérennité), et d'autre part d'aligner le signal audio avec le texte de la transcription : il était alors possible, en quelques secondes, d'écouter n'importe quelle partie de l'enregistrement, et de voir apparaître à l'écran la transcription qui en avait été faite. Cette synchronisation offre, entre autres, la possibilité de réécouter très facilement un extrait pour voir si les propos transcrits y correspondent, et ainsi de corriger rapidement une erreur ou une interprétation. Les logiciels d'aide à la transcription qui proposent cette fonctionnalité sont aujourd'hui très répandus, et nous allons nous arrêter sur quelques-uns des plus utilisés à l'heure actuelle.

3.2. *Les logiciels d'aide à la transcription*

Il existe principalement trois logiciels utilisés pour la transcription orthographique d'un fichier son : TRANSCRIBER (Barras *et al.*, 1998), PRAAT[4] et WINPITCHPRO (Martin, 2003). Moins répandus pour des raisons diverses (outils payants, ergonomie discutable...), CLAN, EXMARALDA ou encore TRANSANA n'en méritent pas moins d'être cités ici, chacun offrant des possibilités intéressantes. Sur le fond, bien qu'aucun ne soit réellement optimisé pour transcrire de la parole spontanée à grande échelle, leur interface globale offre cependant la possibilité d'en gérer quelques aspects.

TRANSCRIBER, logiciel avec une interface et des fonctionnalités simplifiées, est optimisé pour la transcription et l'annotation de gros corpus, mais ne propose que quatre niveaux d'annotation (texte, locuteurs, thème, bruits de fond éventuels) et aucune possibilité analytique. Malgré cela, la gestion des locuteurs est très satisfaisante, puisque l'on peut indiquer pour chacun d'entre eux des informations telles que leur sexe, le degré de spontanéité, le canal d'expression... Par ailleurs, un nombre important de balises est intégré pour représenter les événements sonores

4. www.fon.hum.uva.nl/praat/

(bruit, respiration, toux, reniflement...), les prononciations particulières ou encore des particularités lexicales. Le gros inconvénient de TRANSCRIBER concerne la parole superposée qui, nous le verrons, est traitée de façon trop simplifiée pour pouvoir rendre compte de ce phénomène majeur dans la langue parlée.

Pour ce genre de données, PRAAT s'avère nettement plus efficace, puisqu'il offre un grand nombre de tires indépendantes les unes des autres. Il est possible d'en assigner une à chaque locuteur, et ainsi de transcrire indépendamment leurs propos, tout en les alignant avec le signal. Le fichier de sortie correspondant (« textgrid ») offre la possibilité d'organiser la transcription suivant l'échelle temporelle ou bien par locuteur, ce qui se révèle fort pratique, pour des recherches lexicales, par exemple. On peut toutefois regretter qu'il ne soit pas au format XML, standard aujourd'hui incontestable pour assurer l'échange et la compatibilité des données. Autres aspects dommageables, ce logiciel présente une interface assez austère, et n'offre qu'une gestion minimale des locuteurs : hormis leur nom, rien ne peut être indiqué dans l'espace qui leur est attribué. À vrai dire PRAAT, bien moins efficace que TRANSCRIBER pour le traitement de gros corpus, est généralement privilégié pour des tâches spécifiques, notamment l'analyse de la prosodie, domaine dans lequel il se révèle très complet grâce à la possibilité d'intégrer à ce logiciel de nombreux modules complémentaires.

WINPITCHPRO est pour sa part plus difficile d'accès, moins par son interface (plutôt intuitive) que par la richesse de ses fonctionnalités. Certes moins habile que TRANSCRIBER pour gérer les fichiers audio de grande taille, il permet des analyses très fines (quatre-vingt-seize niveaux d'annotation sont disponibles, soit autant de possibilités de codage) et à plusieurs niveaux : prosodie, phonologie... Par ailleurs, il traite les fichiers audio et vidéo, ce qui le distingue des deux outils précités et permet une synchronisation entre l'image, le signal et la transcription. Il est malgré tout regrettable que cet outil ne fonctionne que sous Windows, et qu'il ne soit pas *open source*.

3.3. *Corpus disponibles et conventions existantes*

Aujourd'hui, de plus en plus de corpus de français parlé se créent au sein de divers laboratoires et groupes de recherche. Les plus importants à l'heure actuelle – plus de quatre cent mille mots – sont la base de données CLAPI (Lyon, équipe ICAR)[5], le corpus CRFP (Aix-en-Provence, équipe DELIC)[6], le projet ASILA[7], ainsi que deux initiatives belges, VALIBEL[8] et ELICOP[9]. À ces derniers viennent se joindre ceux issus de la communauté « parole » (ESTER par exemple), ainsi que

5. http://clapi.univ-lyon2.fr
6. sites.univ-provence.fr/~veronis/pdf/2004-presentation-crfp.pdf
7. www.loria.fr/projets/asila
8. www.uclouvain.be/valibel
9. bach.arts.kuleuven.be/elicop

divers autres apports tels que le corpus PFC (phonologie du français contemporain) (Durand *et al.*, 2002) ou les projets RHAPSODIE et VARILING[10]. Tous n'ont pas été constitués avec le même objectif : ESTER se place dans le domaine de la reconnaissance de la parole, le CRFP s'intéresse à la morphosyntaxe, VARILING s'inscrit dans une perspective sociologique, RHAPSODIE traite de la prosodie, tandis que PFC considère les aspects phonétiques et phonologiques de la langue.

Le projet PFC s'intéresse à la prononciation du français sous trois angles différents : géographique, social et stylistique. Son objectif principal est de constituer, grâce aux enquêtes très précises de chercheurs et d'étudiants, un important corpus représentatif du français parlé dans le monde. Une telle entreprise laisse entrevoir de nombreuses perspectives, notamment dans le domaine de l'enseignement du français : celui-ci pourrait évoluer pour être plus proche d'une langue de « terrain » si une dynamique de corpus à grande échelle tels que PFC venait à se créer.

En plus de ces corpus, un important travail de coordination a été mis en place ces dernières années pour que ceux-ci ne soient pas dispersés dans les différents laboratoires qui les ont collectés. Le CRDO[11] (Centre de ressources pour la description de l'oral) en est un exemple patent, puisque son « archive ouverte » regroupe actuellement de nombreux corpus dans différentes langues et dialectes, dont certains sont consultables librement. À une moindre échelle, le projet ASILA[12] met également à disposition une dizaine de corpus aux thématiques variées. De même, Paul Cappeau et Magali Seijido ont réalisé pour la DGLFLF un « inventaire des corpus oraux »[13], qui passe en revue un grand nombre de corpus de langues française et étrangères, fournissant pour la plupart d'entre eux des informations sur le type de données enregistrées, la transcription, la disponibilité, etc.

Cette disponibilité est d'ailleurs un sujet qui fait débat : en effet tous les corpus existants sont loin de proposer la même accessibilité, quand certains restent même fermés à toute personne n'ayant pas participé à leur élaboration – autant dire à tout le monde. Ainsi, ASILA apparaît aujourd'hui bien seul dans la catégorie du « libre-service » : si CLAPI tend à s'en rapprocher en proposant au téléchargement certaines transcriptions ainsi que leurs fichiers audio correspondants, ELICOP et VALIBEL ne sont « que » consultables en ligne, tandis que de gros corpus comme ESTER sont disponibles en échange d'une somme d'environ 300 €. Enfin, le CRFP (corpus de référence du français parlé), probablement le plus riche en parole spontanée, n'est à ce jour pas accessible en dehors du laboratoire qui l'héberge.

L'un des problèmes qui se posent lorsque l'on entreprend d'effectuer une transcription est celui des conventions d'annotation à adopter : outre le texte lui-même, que veut-on représenter à l'écran, et surtout comment souhaite-t-on le faire ?

10. www.univ-orleans.fr/eslo/spip.php?rubrique13
11. http://crdo.risc.cnrs.fr/exist/crdo/
12. http://www.loria.fr/projets/asila/
13. www.culture.gouv.fr/culture/dglf/recherche/corpus_parole/Inventaire.pdf

TRANSCRIBER, par exemple, a son propre « manuel du transcripteur », indépendant du manuel d'utilisation, et qui passe en revue de nombreux aspects de la langue orale, en en proposant à chaque fois un codage[14]. Heureuse initiative qui permet aux utilisateurs de réaliser rapidement des transcriptions complètes et unifiées, d'autant que ni PRAAT ni WINPITCHPRO ne proposent ce genre de documentation.

Dans la pratique, des conventions diverses sont nées au fil des projets ou des groupes de recherche qui ont vu le jour. Des initiatives telles que le *Linguistic Data Consortium*[15] ou la TEI[16] proposent également des conventions pour la transcription de la parole. Ainsi, si parfois ces codages se recoupent, il arrive souvent que les possibilités de représentation soient très nombreuses.

Conventions	**Codage proposé**
TRANSCRIBER	des idées ré() révolutionnaires
LDC	des idées [ré-] * révolutionnaires
DELIC / ICOR	des idées ré- révolutionnaires
PFC / VALIBEL	des idées ré/ révolutionnaires
TEI	des idées <del type="truncation">ré</del> révolutionnaires

Tableau 1. Codages proposés pour la troncation

Comme on le voit ci-dessus, pour un seul phénomène (parmi bien d'autres), il existe au moins cinq codages différents. Il serait pourtant indispensable de s'orienter vers un codage unifié, ne serait-ce que pour permettre un échange, une compatibilité et une lecture des données plus simples. La TEI (*Text Encoding Initiative*), ensemble de recommandations pour coder des informations avec une nomenclature prédéfinie afin de pouvoir les échanger facilement ensuite, est une base difficilement contestable. Un chapitre y est consacré à la transcription de la parole. Très complet, il passe en revue tous les principaux phénomènes conversationnels et propose des solutions très intéressantes, pour la parole interactive et superposée notamment. La prosodie y est également considérée sous de nombreux aspects (vitesse d'élocution, volume sonore, intonation, rythme, qualité de la voix...). Cependant, ce format reste assez difficile à appréhender car, s'il permet de coder de très nombreux paramètres, il n'existe malheureusement pas d'interface qui les représente de façon intuitive à l'écran.

14. http://trans.sourceforge.net/en/transguidFR.php
15. www.ldc.upenn.edu/Projects/MDE/Guidelines/SimpleMDE_V6.2.pdf
16. www.tei-c.org

4. Traitement de la parole spontanée

4.1. *Le projet EPAC*

Sélectionné par l'ANR[17] dans le cadre de l'appel à projets 2006 du programme Masse de Données – Connaissances Ambiantes (MDCA), le projet EPAC[18] (Exploration de masse de documents audio pour l'extraction et le traitement de la parole conversationnelle) concerne quatre laboratoires : l'IRIT (Toulouse), le LI (Tours), le LIA (Avignon) et le LIUM (Le Mans). Il a pour but de proposer des méthodes d'extraction d'information et de structuration de documents audio, en mettant l'accent sur le traitement de la parole conversationnelle. Le corpus mis à disposition pour ce projet est constitué d'environ deux mille heures d'enregistrements radiophoniques, dont mille huit cents proviennent de la campagne ESTER. La parole conversationnelle/spontanée y occupe une place à première vue modeste que nous avons estimée, d'après une évaluation interne, à environ 30 %. Cependant, cette proportion doit être rapportée à la nature des données : ESTER comporte une bonne part de *broadcast news*, c'est-à-dire un mode d'expression fortement contraint, tant du point de vue du contenu que de la forme, la parole étant monopolisée par des professionnels. Ainsi, les enregistrements de France Info, qui représentent près de 40 % du corpus, ne contiennent par essence pas ou très peu de parole spontanée. Autrement dit, si l'on ne tient pas compte de cette radio, le chiffre passe de 30 à 50 %, ce qui est finalement beaucoup, et suppose que la parole spontanée ne se réduit pas aux interviews hors grande écoute de personnes de milieu modeste.

L'un des sous-projets d'EPAC, intitulé « annotation et évaluation », a précisément pour objectif de définir quels enregistrements doivent être considérés comme étant conversationnels, pour ensuite en transcrire une centaine d'heures et ainsi fournir les données nécessaires à l'entraînement, au développement et à l'amélioration de systèmes de reconnaissance automatique de la parole.

Ayant préalablement défini ce que nous entendions par « parole conversationnelle », avec toutes les spécificités que cela sous-entend, nous allons à présent nous attacher à présenter et détailler la tâche de transcription elle-même.

En premier lieu, il nous faut préciser que ces transcriptions sont réalisées à l'aide du logiciel TRANSCRIBER, qui présente en la circonstance de multiples avantages :

– il offre la possibilité d'aligner le signal audio avec la transcription, ce qui permet d'écouter ou de réécouter très facilement n'importe quelle partie de l'enregistrement, tout en en visualisant la transcription ;

17. http://www.agence-nationale-recherche.fr
18. http://epac.univ-lemans.fr

– la gestion des locuteurs est très complète, puisqu'il est possible de spécifier pour chacun d'entre eux, en plus de leur identité, des éléments tels que le type de parole, la qualité de l'enregistrement, le canal utilisé... ;

– ce programme est gratuit, *open source*, ergonomique, simple d'accès et sait traiter de nombreux formats en entrée comme en sortie ;

– il peut gérer des fichiers audio de plusieurs heures, bien qu'il ne soit pas forcément optimisé pour la parole conversationnelle, problème sur lequel nous reviendrons ultérieurement ;

– des balises sont disponibles pour représenter des éléments sonores (bruits divers, jingles, inspirations...), lexicaux ou encore des prononciations particulières ;

– le format des fichiers de transcription (.trs) permet de faire de l'apprentissage sur les systèmes de reconnaissance, ce qui est une condition *sine qua non* dans le cadre du projet EPAC.

Pour les besoins du projet, nous avons principalement transcrit des données radiophoniques (France Culture, RFI, France Inter, RMC), mais également quelques émissions de la chaîne de télévision France 5 susceptibles de contenir de la parole spontanée, eu égard à leur forme même (débats, interviews, rencontres). Nous avons ainsi constitué un corpus d'environ quatre heures avec des transcriptions de *Arrêt sur images*, *C dans l'air*, *C'est notre affaire*, *Madame monsieur bonsoir* et *Ripostes*. Par rapport aux émissions radiophoniques, la principale différence, et pour ainsi dire la principale difficulté, est d'identifier les locuteurs qui prennent la parole. En effet, la radio ne bénéficiant pas du support de l'image, le nom des intervenants est systématiquement précisé, pour que l'auditeur ne perde pas le fil de l'émission. À la télévision, l'image rend les choses totalement différentes : en général le nom du locuteur s'affiche à l'écran lors de sa première intervention, et ensuite la caméra aura toujours tendance à se cadrer systématiquement sur la personne en train de s'exprimer. Seulement, TRANSCRIBER ne gérant que très sommairement l'usage de la vidéo, il nous a fallu régulièrement faire des allers-retours entre l'interface du logiciel et le fichier multimédia pour attribuer correctement les locuteurs, ce qui représente une perte de temps conséquente.

Par ailleurs, des émissions comme *Ripostes*, où le nombre d'intervenants simultanés peut parfois s'élever à quatre ou cinq, sont très difficiles à traiter, sans même poser la question de l'efficacité du logiciel utilisé : dans de tels cas, l'annotateur humain ne peut percevoir distinctement chaque flux de parole, et, quand bien même il y parviendrait, ce serait au prix d'un très long travail d'écoute.

4.2. *Le concept de « qualité d'élocution »*

Pour tenter de contourner l'ambiguïté que suppose la distinction parole préparée/parole spontanée (lorsqu'un locuteur habitué à s'exprimer parle spontanément, son propos s'apparente à de la parole préparée), nous avons réalisé

une étude en considérant cette fois-ci la qualité d'élocution, concept que nous allons expliciter dans les lignes qui suivent. Un corpus radiophonique d'environ onze heures a été choisi, comprenant des extraits de France Inter, France Info, Radio Classique, RFI et France Culture. Les fichiers ont été segmentés de façon automatique, et le texte a été, quant à lui, supprimé. Deux annotateurs étaient chargés de noter chaque segment de parole suivant une échelle numérique allant de 1 à 9. La note 1 était celle attribuée à un segment sans aucune disfluence, avec une élocution parfaitement claire ; 9 indiquait un segment inaudible tant les hésitations, répétitions, faux départs, etc., étaient nombreux – ce cas extrême n'a jamais été rencontré au cours de l'expérience. Une note globale était ensuite attribuée à chaque tour de parole, pour éviter d'accorder la même importance à un segment très bref, donc potentiellement moins susceptible de comporter des disfluences, et à un segment long de plusieurs secondes.

Une partie de ces onze heures a été évaluée conjointement par les deux annotateurs, pour qu'ils soient sûrs de bien utiliser les mêmes critères de notation. Ensuite le coefficient Kappa (Cohen, 1960) a été calculé pour valider le processus. Un score de 0,852 a été obtenu, sachant que les scores dépassant 0,81 sont considérés comme étant excellents. Cela prouve que malgré la relative subjectivité du concept de « qualité d'élocution », les deux annotateurs étaient d'accord pour déterminer ce qui était de la parole de bonne qualité et ce qui n'en était pas.

Puis, afin de vérifier si cette « qualité d'élocution » allait de pair avec le taux d'erreur mot, nous avons ensuite mis en parallèle les sorties automatiques générées par LIUM RT, le système de reconnaissance automatique du LIUM (Estève *et al.*, 2004 ; Deléglise *et al.*, 2005) avec les transcriptions manuelles de référence. Le taux d'erreur mot a été mesuré sur chaque ensemble de segments ayant obtenu la même note. Voici les résultats que nous avons obtenus :

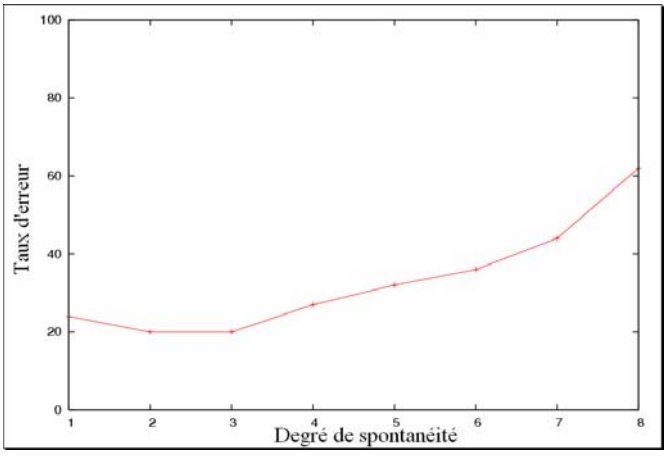

Figure 2. Taux d'erreur de LIUM RT en fonction de la qualité d'élocution

Comme l'indique ce graphique, une corrélation entre qualité d'élocution et taux d'erreur mot existe. Si l'on excepte les segments auxquels ont été adjointes les notes 2 et 3, légèrement mieux reconnus que ceux à qui a été attribuée la note 1, les autres suivent une courbe ascendante, dépassant la barre des 60 % d'erreurs sur les segments notés 8. À bien y regarder, il est en fait possible de dégager trois grandes catégories :

– de 1 à 3, la parole est de bonne voire très bonne qualité, avec peu ou pas de disfluences, et le taux d'erreur se situe aux alentours de 20 % ;

– de 4 à 6, la parole est moins fluide, comprend quelques hésitations, et le taux d'erreur va jusqu'à approcher les 40 % ;

- les segments notés 7 et surtout 8 tendent vers une parole difficilement compréhensible, riche en hésitations, bégaiements..., d'où les performances moindres de LIUM RT, qui oscillent entre 45 et 60 %.

4.3. *Détection automatique de la parole spontanée*

Afin de pouvoir mettre en place des stratégies spécifiques pour la transcription automatique de la parole spontanée, mais aussi pour extraire des zones de parole spontanée à partir de masses de données audio, il est nécessaire de développer des outils adéquats.

Nous avons donc étudié plusieurs critères linguistiques, qui ont comme particularité importante de pouvoir être fournis par un système de transcription automatique, afin de caractériser la parole spontanée par rapport à la parole préparée.

Parmi ces critères, la prosodie occupe une place importante : en effet, des travaux antérieurs ont mis en évidence des liens entre prosodie et parole spontanée. Comme proposé dans (Shriberg, 1999), nous avons ainsi utilisé la durée des voyelles et des mélismes (*cf.* 1.9). Nous avons pris en compte, pour un segment de parole donné, la durée moyenne, la variance et l'écart type de ces unités pour mesurer la dispersion des durées autour de leur moyenne. De plus, nous avons intégré, comme proposé dans (Caelen-Haumont, 2002b) qui montre la corrélation entre débit de parole et état d'émotivité d'un locuteur, des informations concernant le débit phonétique sous deux formes : la variance du débit phonétique pour chaque mot, et la moyenne des débits phonétiques pour l'ensemble du segment de parole visé.

Par ailleurs, un grand nombre de travaux ont permis de décrire les disfluences (pauses pleines, hésitations, répétitions, faux départs, etc.) au niveau acoustique (Shriberg, 1999) ou lexical (Siu et Ostendorf, 1999). Nous avons retenu deux critères pour les représenter qui peuvent être fournis par un système de transcription automatique de la parole :

– les pauses pleines : le dictionnaire du système de transcription automatique contient plusieurs symboles qui représentent les « euh », « ben », « hum »... Le nombre d'occurrences de ces symboles dans un segment de parole est un premier critère retenu ;

– répétitions et faux départs : nous avons réduit ce critère au nombre d'unigrammes et de bigrammes qui sont répétés dans un segment de parole. Ce nombre est le second critère retenu.

EXPERIENCES — Nous avons utilisé le corpus annoté manuellement en niveau de fluidité présenté en section 3.2 pour mettre en place un apprentissage automatique des seuils pertinents pour chacun des critères prosodiques et linguistiques retenus, et pour les utiliser ensemble pour détecter la parole spontanée.

Les segments de ce corpus ont été classés en deux catégories : parole spontanée (note subjective comprise en 4 et 10), et parole préparée (note subjective comprise entre 1 et 3). Cette utilisation conjointe des critères retenus pour distinguer parole spontanée *vs* parole préparée a été effectuée en utilisant l'outil BoosTexter fondé sur l'algorithme de classification AdaBoost (Schapire et Singer, 2000).

Pour mener à bien nos expériences, nous avons eu recours à la technique du *leave-one-out* qui consiste à extraire une partie du corpus disponible, à effectuer un apprentissage sur la partie restante, et à tester le système sur la partie extraite. Afin de pallier la petite taille des données, ce processus est réitéré plusieurs fois en retirant d'autres parties indépendantes et en réinjectant les données retirées précédemment. Les tableaux suivants montrent les résultats obtenus en terme de précision, rappel et f-mesure (qui combine précision et rappel : ici sont présentés les résultats maximisant la f-mesure) sur les deux types de parole. La première colonne montre les résultats obtenus en utilisant les critères linguistiques sur les phrases de référence transcrites manuellement, la deuxième les résultats obtenus par ces critères sur les phrases proposées par le système de reconnaissance de la parole (SRAP), la troisième les résultats obtenus à l'aide des critères acoustiques fournis par le SRAP et la dernière les résultats obtenus par le système combinant tous les critères extraits automatiquement.

Parole spontanée				
Caractéristiques	ling(réf)	ling(rap)	acou(rap)	ling+acou(rap)
Précision	70,8	62,7	63,9	66,5
Rappel	62,1	56,7	56,3	61,5
F-mesure	66,2	59,5	59,9	63,9

Parole préparée				
Caractéristiques	ling(réf)	ling(rap)	acou(rap)	ling+acou(rap)
Précision	79,2	75,8	75,9	78,3
Rappel	84,9	80,1	81,2	81,7
F-mesure	81,9	77,9	78,5	80,0

Tableau 2 et 3 : *Performances en détection (précision/rappel)*

Nous constatons que la combinaison des critères extraits automatiquement à l'aide d'un SRAP autorise des résultats comparables aux résultats obtenus en étudiant uniquement les transcriptions de référence à l'aide des critères linguistiques, malgré un taux d'erreur sur les mots avoisinant globalement 32 %. Ces résultats, quoique encourageants, nécessitent d'être encore améliorés pour être exploités avec précision.

4.4. *Transcription automatique de la parole spontanée*

Comme nous avons pu le voir, la parole spontanée propose des spécificités qui la distinguent de la parole préparée. Ces spécificités, malheureusement, entraînent une baisse des performances des systèmes de reconnaissance automatique de la parole en terme de taux d'erreur sur les mots.

Une grande partie de nos travaux actuels, dans le cadre du projet EPAC qui a débuté en 2007, vise justement à améliorer ces performances.

Dans cette optique, nous avons voulu mesurer l'apport de données différentes de celles utilisées dans nos expériences qui sont principalement des données issues du corpus ESTER.

Pour cela, nous avons utilisé le corpus PFC (Durand et al., 2002). Comme vu précédemment, ce corpus contient un grand nombre de transcriptions de parole conversationnelle : nous avons utilisé 26 000 phrases contenant 285 000 occurrences de mots pour construire un modèle de langage probabiliste de parole spontanée.

En particulier, les thèmes abordés dans ces transcriptions par les locuteurs sont très éloignés des thèmes rencontrés dans l'application visée.

RÉSULTATS D'EXPÉRIENCE —: Nous avons découpé le corpus de 11 heures d'audio utilisé dans les expériences précédentes en deux corpus : un corpus de développement (sept fichiers pour un peu moins de 7 heures de parole) et un corpus de test (quatre fichiers pour une durée totale de 4 h 15). Nous avons classé les segments en trois catégories : parole préparée (niveau 1 d'annotation subjective de

fluidité de l'énonciation, contenant 13 493 occurrences de mots), parole légèrement spontanée (niveaux 2-3-4, contenant 12 218 occurrences de mots) et parole fortement spontanée (niveaux supérieurs ou égaux à 5, contenant 19 292 occurrences de mots).

Deux modèles de langage trigrammes ont été comparés :

– le modèle de langage de référence (*base*), qui correspond au modèle de langage utilisé lors de la campagne ESTER (Deléglise, 2005) estimé à partir de transcriptions manuelles de 90 heures d'enregistrements radiophoniques et de dix-sept années d'articles du journal *Le Monde* ;

– un modèle de langage (*base+pfc*) qui combine linéairement le modèle de référence avec le modèle de langage estimé sur les 26 000 phrases du corpus PFC citées plus haut, le coefficient d'interpolation ayant été optimisé sur le corpus de développement.

En utilisant la mesure de perplexité généralement employée pour estimer la pertinence d'un modèle de langage probabiliste vis-à-vis d'un corpus textuel, le tableau suivant présente les résultats obtenus par ces deux modèles sur les différentes classes de spontanéité du corpus de test.

Modèle de langage	**Parole préparée**	**Légèrement spontanée**	**Fortement spontanée**
Base	156	193	203
Base+pfc	171	184	164 **(– 19,21 %)**

Tableau 4. *Perplexité des modèles de langage obtenue sur les segments du corpus de test en fonction de leur classe de spontanéité*

Comme on peut le voir, le corpus PFC n'apporte rien pour la parole préparée, et dégrade même les résultats en terme de perplexité (plus la valeur perplexité est basse, plus le modèle est pertinent) : ceci s'explique certainement par la non-concordance des thèmes abordés dans le corpus PFC et le corpus de test utilisé ici. En revanche, il est intéressant de remarquer que le corpus PFC permet d'obtenir un modèle de langage plus pertinent pour les segments de parole spontanée, et plus particulièrement pour les segments de parole fortement spontanée pour lesquels la perplexité diminue de presque 20 %, ce qui est une réduction très significative.

Malheureusement, lorsque l'on mène des expériences de reconnaissance de la parole, ces gains ne sont pas retrouvés, malgré une baisse non significative du taux d'erreur global. Le tableau suivant montre les résultats en terme de taux d'erreur sur les mots obtenu par LIUM RT sur les enregistrements audio correspondant aux segments précédents.

Modèle de langage	Parole préparée	Légèrement spontanée	Fortement spontanée	Global
Base	21,4	31,3	41,2	**32,6**
Base+pfc	21,7	31,8	40,1	**32,4**

Tableau 5. *Taux d'erreur sur les mots en fonction des modèles de langage obtenu sur les segments du corpus de test en fonction de leur classe de spontanéité*

Nous pensons que les apports visibles du corpus PFC sur la modélisation du langage ne se répercutent pas dans les performances du SRAP en raison de phénomènes extérieurs à cette modélisation qui seraient mal appréhendés par le système. En particulier, les dictionnaires de mots phonétisés que nous utilisons sont trop rigides et ne prennent pas en compte les spécificités de la parole spontanée.

Dans le cadre du projet EPAC, nous travaillons actuellement à mieux modéliser ces prononciations.

En revanche, ces résultats confirment que la parole spontanée doit être traitée de façon spécifique et qu'il est donc utile de développer des outils de détection automatique de la parole spontanée afin de choisir les meilleures stratégies pour la reconnaissance de la parole.

4.5. *Transcription manuelle* vs *transcription assistée : quel(s) gain(s) ?*

Dans le but de quantifier le gain de temps qui pouvait être obtenu grâce à l'utilisation d'un système de reconnaissance automatique par rapport à une transcription réalisée entièrement à la main, nous avons mené l'expérience suivante : vingt-quatre segments d'environ 10 minutes ont été sélectionnés parmi les données non transcrites du corpus ESTER : douze ont été considérés comme étant de la parole spontanée (débats ou interviews), et douze comme de la parole préparée (informations). Sur chacun de ces fichiers, une transcription manuelle et une transcription assistée ont été effectuées par le même transcripteur (suffisamment longtemps après pour que la seconde transcription ne soit plus influencée par la mémoire de la première).

Cette transcription comportait trois niveaux :

– la segmentation en tours de parole et la transcription ;

– l'assignation des locuteurs ;

– la vérification orthographique.

Pour chacune de ces étapes, un chronométrage à la minute a été effectué. Voici les principaux résultats que nous avons obtenus :

	Parole préparée	**Parole spontanée**
Transcription manuelle	17 h 36	19 h 33
Transcription assistée	8 h 31	15 h 44

Tableau 6. *Durée totale de la transcription (durées respectives des corpus : 2 h 08 et 2 h 10)*

Le tableau 6 montre que la transcription assistée induit un important gain de temps, surtout pour la parole préparée. Pour ce type de données, le temps nécessaire à la transcription est approximativement deux fois moins important lorsque le transcripteur est assisté. Lorsqu'il s'agit de parole spontanée, ce bénéfice est bien moindre.

	Parole préparée	**Parole spontanée**
Transcription manuelle	8,26	9,05
Transcription assistée	4,00	7,29

Tableau 7. *Rapport entre la durée totale de la transcription et la durée totale des fichiers*

Étant donné que les fichiers spontanés et préparés représentent peu ou prou la même durée, le rapport entre celle-ci et le temps total nécessaire à la transcription (tableau 7) est un élément qu'il est pertinent de prendre en compte : si l'on considère un segment de parole préparée de 10 minutes, le transcripteur aura besoin d'environ 40 minutes pour transcrire le texte, assigner les locuteurs et vérifier l'orthographe, s'il s'appuie sur un fichier de transcription généré automatiquement. Si l'on réalise les mêmes tâches sur le même fichier, mais cette fois de façon entièrement manuelle, environ 83 minutes seront nécessaires, soit un temps de travail plus que doublé.

La même expérience, mais cette fois avec un fichier de parole spontanée, montre qu'une transcription assistée demande 73 minutes de travail, chiffre qui est presque le double de celui obtenu dans les mêmes conditions avec la parole préparée. À l'inverse, la transcription manuelle (90 minutes) n'est cette fois pas beaucoup plus coûteuse en temps que la transcription assistée. Ainsi, s'il est indéniable qu'une transcription assistée est synonyme de gain de temps, ce dernier est beaucoup plus important lorsqu'il s'agit de parole préparée.

	Parole préparée	**Parole spontanée**
Transcription manuelle	13 h 36	16 h 15
Transcription assistée	5 h 06	12 h 41

Tableau 8. *Transcription du texte et segmentation*

C'est lors de la tâche de transcription du texte (tableau 8) que le gain le plus intéressant a été obtenu : sur de la parole préparée, une transcription manuelle nécessite environ 2,67 fois plus de temps qu'une transcription assistée (5 h 06 *vs* 13 h 36). Ce chiffre est très significatif, notamment s'il est comparé à celui obtenu avec la parole spontanée : pour une durée sensiblement équivalente, il chute à 1,28. Cet écart met en exergue le fait que les systèmes de reconnaissance automatique de la parole éprouvent des difficultés à traiter la parole spontanée, obligeant le transcripteur à effectuer par la suite beaucoup de corrections manuelles.

	Parole préparée	**Parole spontanée**
Transcription manuelle	1 h 17	2 h 13
Transcription assistée	1 h 17	2 h 13

Tableau 9. *Assignation des locuteurs*

En ce qui concerne l'assignation des locuteurs (tableau 9), il est surtout important de retenir que cette tâche demande presque deux fois plus de temps quand la parole est spontanée. Cela peut s'expliquer relativement facilement : les nombreux tours de parole de la parole spontanée contraignent le transcripteur à devoir leur assigner un locuteur chacun, et ce même s'il n'y en a que deux dans un fichier. À l'inverse un segment de parole préparée contient souvent de nombreux locuteurs (journalistes, reporters, interviewés, speakers...), mais beaucoup moins de tours de parole dans la mesure où ceux-ci sont beaucoup plus longs. De plus, dans un segment spontané se trouve parfois de la parole superposée, et lorsque trois locuteurs ou plus sont susceptibles de prendre la parole, cela peut être long et difficile de déterminer qui parle réellement.

	Parole préparée	**Parole spontanée**
Transcription manuelle	2 h 43	1 h 05
Transcription assistée	2 h 08	0 h 51

Tableau 10. *Correction orthographique*

Le minutage de la correction orthographique (tableau 10) a permis d'observer un phénomène remarquable : si la différence spécifique entre transcription manuelle et assistée n'est certes pas très significative, celle entre parole préparée et spontanée l'est beaucoup plus. La raison en est fort simple : les segments de parole préparée contiennent essentiellement de l'information radiophonique ; or ce genre de données s'avère très riche en noms propres (reporters, interviewés, personnalités, villes...), dont les orthographes exactes ne peuvent être systématiquement connues de l'annotateur. Les rechercher peut donc être une tâche assez longue, notamment dans le cas de noms étrangers. Inversement, les fichiers de parole spontanée étant des interviews ou des débats, on y trouve très peu de noms propres car les thèmes abordés ne nécessitent en général qu'un faible emploi de ces entités nommées.

	Parole préparée	**Parole spontanée**
Transcription manuelle	16,95	35,21
Transcription assistée	15,83	34,33

Tableau 11. *Taux d'erreur mot (%)*

Les dernières observations effectuées concernent le taux d'erreur mot (tableau 11). Celui-ci a été mesuré à partir des sorties automatiques générées par le système LIUM RT, lesquelles ont été comparées aux transcriptions manuelles, puis assistées, réalisées par l'annotateur. Les moyennes indiquées ci-dessus confirment ce que nous disions précédemment : le système de reconnaissance automatique du LIUM n'est pas aussi performant sur la parole spontanée que sur la parole préparée. À titre d'exemple, le taux d'erreur mot le plus élevé que nous ayons obtenu sur de la parole préparée était de 21,8 %, alors qu'il s'est élevé à 53,4 % avec la parole spontanée. Les différences observées entre les tâches manuelles et assistées peuvent être expliquées par le fait que le transcripteur n'a pas forcément transcrit le même texte à chaque fois : il est parfois difficile de percevoir clairement des phénomènes tels que les répétitions, les faux départs ou encore la parole superposée. En conséquence les transcriptions ne seront pas toujours identiques, même lorsqu'elles sont réalisées deux fois par la même personne.

4.6. *Relevé, classement et analyse des principales erreurs des systèmes de reconnaissance automatique*

Si l'on sait aujourd'hui que les systèmes de reconnaissance automatique sont moins performants sur la parole que l'on appelle « spontanée », il n'en reste pas moins que la parole « préparée » est elle aussi source d'erreurs, bien que celles-ci soient en nombre nettement inférieur, et surtout appartiennent à des catégories bien précises. Nous allons donc tenter de proposer, exemples à l'appui, un classement et une analyse des erreurs commises par le système de reconnaissance LIUM RT.

4.6.1. *Homonymes/paronymes*

La principale difficulté éprouvée par le système de reconnaissance automatique est de traiter les phénomènes d'homonymie et de paronymie. Ceux-ci sont particulièrement importants en français, où les monosyllabes homophones sont beaucoup plus nombreux que dans d'autres langues, et où la combinaison de syncopes et d'assimilations produit une morphologie liée particulièrement ambiguë. Concernant les homonymes, nombreux sont en effet les cas où la suite de phonèmes perçue par le système est la bonne, mais sans la transcription orthographique idoine. En voici quelques exemples :

(8) « là je viens d'ouvrir » : *l'âge vient d'ouvrir*
(9) « affirment elles avoir interpellé » : *affirmaient l'avoir interpellé*
(10) « proches, hein !» : *prochain*
(11) « chevauchement de compétence » : *chevauchent Mende compétences*
(12) « sont là statiques i(l)s bougent pas » : *sont lasse Tati qui bougent pas*

Comme en atteste ce relevé, la distinction parole préparée/parole spontanée n'est pas forcément la cause des erreurs du système : dans l'exemple 12, l'élision du « l » appartient au domaine du spontané, et il est à peu près certain que la prononciation du phonème correspondant aurait évité les confusions qui résultent de son élision. Néanmoins, si l'on considère l'exemple 9, qui est issu d'un flash d'information, aucune altération phonologique n'apparaît, ce qui n'empêche pas le système de proposer une séquence, acoustiquement exacte, mais sémantiquement erronée. Ce type d'erreurs, difficilement évitable à l'heure actuelle, est donc susceptible d'apparaître quel que soit le contexte langagier dans lequel on se trouve.

Il est cependant indéniable que la langue française elle-même joue un rôle important dans l'apparition de ces confusions : contrairement à ses homologues anglaise, allemande ou espagnole, elle est d'une très grande richesse homonymique, allant des monosyllabes (*foi/fois/foie/Foix ; lait/les/lais/laie...*) aux vers holorimes (*Gal, amant de la reine, alla tour magnanime/galamment de l'arène à la tour Magne à Nîmes*). Cette singularité, qui passerait volontiers pour un charmant idiotisme, devient dans le domaine de la reconnaissance automatique de la parole un insoluble casse-tête... Par rapport aux autres langues latines notamment, elle repose sur le fait que le français a opéré au cours de son histoire une réduction syllabique massive, qui aboutit à un nombre d'autant plus considérable de monosyllabes homophones qu'on inclut les formes fléchies et la morphologie liée. Le tableau 12 met par exemple en regard les différentes graphies de la séquence [tã] et leurs traductions en italien, en l'occurrence toutes fondées sur les mêmes étymons latins. À cela s'ajoute, notamment pour les verbes, que le français marque la personne non plus à droite du verbe par une désinence (comme le latin ou l'italien) mais à gauche du verbe, par un « pronom » susceptible d'être modifié et disjoint (*eux, qui parlent, sont...*), représenté (*moi qui ai/a*), ou même réduit (*je suis, tu es, vous êtes* deviennent [Sshi ou Shi, *te* ou tE, zEt]).

Français	Italien
tant	tanto
temps	tempo
taon	tafano
tends (tendre pers1)	tendo
tends (tendre pers2)	tendi
tend (tendre pers3)	tende
t'en (je t'en parle)	te ne (te ne parlo)

Tableau 12 : *graphies [ta~] et traductions en Italien*

4.6.2. *« e » ouvert/« e » fermé*

Ensuite, et c'est là sans doute le nœud du problème, il existe en français une confusion parfois totale entre le « e » ouvert et le « e » fermé. Théoriquement, la phonétique voudrait, par exemple, que la forme verbale « j'ai » se prononçât [ZE], puisque composée de la séquence « ai ». Toutefois, nombreux sont les cas dans lesquels le son produit est un « e » fermé, ce qui donne la séquence [Ze] (j'ai mis/gémir). Cette ambivalence est notamment très délicate à gérer pour les formes de l'imparfait, parfois presque impossibles à distinguer de celles du passé composé ou de l'infinitif (*l'enfant aimait sauter dans l'eau/l'enfant aimé sautait dans l'eau)*. De même, entre autres mots-outils monosyllabiques, déterminants et pronoms sont systématiquement source de confusions (j*e l'ai/geler, les faits/l'effet, des faits/défaire*...). Enfin, cette ambivalence induit des erreurs de structure. L'ambiguïté phonologique transforme la morphologie et fait dérailler la syntaxe :

« j'ai été » : *j'étais*
« le papa c'est une » : *le pas passé une*
« c'est cool » : *s'écoule*
« traîner » : *Trénet*
« vous demandez » : *vous demandait*

De même, il arrive parfois que LIUM RT assimile certaines séquences sonores à des suites de lettres, toujours phonétiquement identiques ou très proches :

« et ça » : *SA*
« et euh » : *et E*
« j'ai j'ai » : *g g*
« c'était » : *CT*

Inversement, il se peut que le système ne reconnaisse pas un sigle et le transcrive sous forme de mots :

« MSA » : *mais ça*

Notre corpus le montre : ces erreurs ne sont pas toutes dues à l'emploi de la parole spontanée, et bon nombre d'entre elles proviennent d'extraits contenant de la parole préparée, ou s'y appliqueraient volontiers.

4.6.3. *Assimilations*

Cela dit, il est effectivement des spécificités de la parole spontanée qui sont source d'erreurs d'interprétation du logiciel de reconnaissance automatique, et notamment l'assimilation. Cette variation phonétique, entraînant la modification de la prononciation d'une consonne sourde au contact d'une consonne voisine sonore (ou l'inverse), est d'autant plus fréquente dans la parole spontanée qu'elle est très souvent provoquée par la disparition d'un schwa, caractéristique récurrente de ce type de discours. Et le système, souvent peu entraîné à ce genre de phénomène, ne sait pas toujours déduire le mot ou la séquences de mots exacts à partir de sa prononciation « assimilée », d'où un nombre important d'erreurs potentielles, tant les possibilités d'interférence entre consonnes sont nombreuses. La plus fréquente est certainement celle confrontant le « d » (qui est une consonne sonore) à une consonne sourde, dans la séquence « de + nom ou verbe », où le « e » est élidé, contraignant ainsi le son « d » à devenir « t ». Nous avons eu l'occasion de relever plusieurs occurrences lors de nos expériences :

« envie d(e) passer » : *vite passé*
« pas d(e) sanitaires » : *patte sanitaire*
« coup d(e) fil » : *coûte fils*

Dans chacun de ces trois exemples, le système LIUM RT, ne percevant ni la consonne « d » (puisqu'elle est prononcée « t ») ni la voyelle « e » (puisqu'elle est élidée), est incapable de générer la structure prépositionnelle introduite par « de ». En lieu et place de celle-ci, il propose donc une suite de mots rigoureusement exacte phonétiquement, mais incohérente contextuellement, comme il le faisait pour les autres cas d'homonymies que nous avons vus précédemment.

4.6.4. *Répétitions, faux départs, troncations*

Par ailleurs, outre l'assimilation, d'autres spécificités de la parole spontanée posent régulièrement problème aux systèmes de reconnaissance automatique : les répétitions, faux départs, troncations ou autres disfluences sont autant d' « anomalies » langagières qu'ils n'ont pas l'habitude de rencontrer. Pour les premières citées, il est intéressant de constater que LIUM RT s'est même, en de rares occasions, refusé à proposer deux occurrences consécutives du même mot, bien que la prononciation ne laissait planer aucune ambiguïté :

« faut faut faut faut » : *faut fois font fois*

Au sujet des troncations ou des faux départs, ils génèrent inévitablement de nouvelles alternatives homonymiques. Et à nouveau, le système de reconnaissance

automatique se retrouve à traiter des suites de sons qu'il va chercher à associer à des mots qui lui sont connus, et jamais à des amorces ou fins de mots, particularités qu'on ne retrouve (presque) que dans la parole spontanée. Ce qui ne manque pas, à nouveau, de créer de nouvelles confusions :

« bah s() » : *basses*
« on a des r() » : *on adhère*
« (en)fin » : *fin*

4.6.5. *Autres*

Enfin, nous mentionnerons pour terminer quelques problèmes généraux, que nous avons rencontrés dans une majorité de fichiers. Tout d'abord, et cela concerne surtout la parole spontanée, la parole superposée n'est pas correctement traitée. Naturellement, les enregistrements utilisés étant monophoniques, cette tâche est d'autant plus difficile, voire impossible à réaliser. Il n'en reste pas moins que la superposition de locuteurs fait partie intégrante de la parole spontanée, et que réussir à la traiter serait une avancée considérable dans le domaine de la reconnaissance automatique de la parole.

Des mots relativement brefs comme « et » ou « ou » échappent assez régulièrement à la vigilance de LIUM RT, ce qui s'explique précisément par leur brièveté et par le fait qu'ils soient souvent « aspirés » par les mots qui les précèdent ou les suivent.

Enfin, il arrive fréquemment qu'une inspiration soit interprétée par le système de reconnaissance automatique comme une occurrence de la conjonction de subordination « que ».

4.7. *Quelques pistes pour optimiser les systèmes de reconnaissance automatique*

Pour conclure, nous nous proposons de réfléchir à des hypothèses qui permettraient d'optimiser les systèmes de reconnaissance automatique de la parole. En premier lieu, les corpus eux-mêmes peuvent être un élément de réponse : bien que, comme nous l'avons vu, il n'en existe réellement pas beaucoup de disponibles, il serait peut-être bon d'aller, dans la mesure du possible, vers une unification de ceux-ci. Cela en faciliterait grandement l'échange et la disponibilité, d'autant que ces données sont trop rares pour qu'elles ne profitent pas au plus grand nombre. Malheureusement, c'est souvent difficile, non seulement en raison de transcriptions et de codages spécifiques à un projet, à un laboratoire de recherche ou à un logiciel, mais également à cause d'une diffusion parfois très confidentielle.

Une autre solution serait naturellement de créer de nouveaux corpus, centrés sur la parole spontanée. Or, nous l'avons vu, de tels projets sont très coûteux : quand bien même l'on s'affranchirait de la collecte de données en s'appuyant sur des

enregistrements déjà effectués, la tâche de transcription à elle seule est synonyme de centaines d'heures de travail pour un corpus moyen.

Enfin, nous essayons actuellement d'intégrer dans le dictionnaire de prononciation de LIUM RT les variantes de prononciations que l'on rencontre à l'oral (*parce que* = « pasque », *c'est-à-dire* = « stadir »), afin qu'elles figurent parmi les hypothèses envisageables lors du processus de reconnaissance automatique. De même, nous aimerions pouvoir donner un score de probabilité à ces variantes : il est avéré que dans une situation de parole spontanée, « c'est-à-dire » est souvent prononcé « stadir ». En conséquence, il serait intéressant de pouvoir orienter le système, lorsqu'il rencontre cette séquence sonore, vers la forme « c'est-à-dire » plutôt que vers « salir », solution qu'il propose actuellement face à pareil cas.

En d'autres termes, donner à un système de reconnaissance de la parole la capacité de détecter la parole spontanée (au même titre qu'on est capable d'identifier la langue) pourrait permettre de mettre en œuvre des modalités de reconnaissance spécifiques, intégrées dans l'apprentissage et/ou dans les bases de connaissances.

5. Conclusion

La parole spontanée, avant d'en envisager un quelconque traitement, nécessite d'être définie sous des angles divers, qui ne revêtent pas le même intérêt selon les buts poursuivis. Dans cet article, nous en avons proposé différentes définitions, en insistant sur la morphosyntaxe qui, en français, diverge beaucoup par rapport à la parole préparée et à l'écrit.

Les diverses tâches et expérimentations que nous avons menées dans le cadre de cet article ont permis des avancées significatives :

– constitution d'un corpus de transcriptions d'émissions radiophoniques, extraites du corpus ESTER, qui avoisinera à terme les 100 heures ;

– détection de la parole spontanée au sein de masses de données audio ;

– mesure de l'apport positif d'un corpus de parole conversationnelle comme le corpus PFC pour l'estimation de modèles de langage utilisés dans les systèmes de reconnaissance de la parole spontanée ;

– expérimentation du concept de « qualité d'élocution » ;

– quantification et analyse du gain de temps apporté par la transcription assistée ;

– relevé, classement et analyse des principales erreurs des systèmes de reconnaissance.

De ces travaux se dégagent des hypothèses susceptibles d'améliorer les systèmes de reconnaissance automatique, hypothèses que nous comptons tester au cours des mois à venir.

6. Bibliographie

Adda-Decker, M. *et al.*, « Une étude des disfluences pour la transcription automatique de la parole spontanée et l'amélioration des modèles de langage », JEP 2004, 19-22 avril 2004, Fès (Maroc).

Bally, C., *Traité de stylistique française*, Klincksieck, Paris, 1929.

Barras, C. *et al.*, « Transcriber: a free tool for segmenting, labeling and transcribing speech », LREC 98, Grenade (Espagne), 28-30 mai 1998, p. 1373-1376.

Bartkova, K. et Segal, N., « Détection automatique de frontières prosodiques dans la parole spontanée », JEP 2006, 12-16 juin 2006, Dinard.

Caelen-Haumont, G., « Prosodie et dialogue spontané : Valeurs et fonctions perlocutoires du mélisme », in *TIPA*, n°21, pp. 13-24, 2002a.

Caelen-Haumont, G. « Perlocutory Values and Functions of Melisms in Spontaneous Dialogue », Proceedings of the 1st International Conference on Speech Prosody, Aix en Provence, France, pp. 195–198, 2002b.

Chevallier, G., « Description lexicographique de l'emprunt *well* dans une variété de Français parlé du sud-est du Nouveau-Brunswick », in *Contacts de langue et identités culturelles : perspectives lexicographiques*, Presses de l'Université Laval (PUL), Québec, 2000.

Cohen J., « A coefficient of agreement for nominal scales », in *Educational and Psychological Measurement* 20, pp. 37-46, 1960.

Damourette J. et Pichon, E., *Des mots à la pensée, essai de grammaire de la langue française*, D'Artrey, Paris, 1911-1927.

Deléglise, P. *et al.*, « The LIUM Speech Transcription System: A CMU Sphinx III-Based System for French Broadcast News », Interspeech'2005, Lisbonne (Portugal).

Durand, J. *et al.*, « Un corpus numérisé pour la phonologie du français », in G. Williams (ed.) *La linguistique de corpus*. Rennes: Presses Universitaires de Rennes. pp. 205-217. Actes du colloque *La linguistique de corpus*, Lorient, 12-14 septembre 2002.

Durand, J. *et al.*, « Synopsis du projet PFC, La phonologie du français contemporain : Usages, Variétés et Structure », in *Bulletin PFC* n°1.

Durand, J. et Tarrier, J.-M. (2006), « PFC, corpus et systèmes de transcription », in *Cahiers de Grammaire*, n° 30, pp. 139-158.

Estève, Y., Déléglise, P. et Jacob B., « Systèmes de transcription automatique de la parole et logiciels libres », in *Traitement Automatique des Langues*, vol. 45, n° 2.

Frei, H., *La grammaire des fautes*, Slatkine, Genève, 1929.

Galliano, E. *et al.*, « The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast New », Interspeech'2005, Lisbonne (Portugal).

Gougenheim, G. *et al.*, *L'élaboration du français fondamental : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Didier, Paris, 1964.

Jousse, V. *et al.*, « Caractérisation et détection de parole spontanée dans de larges collections de documents audio », JEP 2008, 9-13 juin 2008, Avignon.

Luzzati, D., « "Ben", appui du discours », in *Le Français Moderne*, vol. 50, n°3, pp. 193-207, 1982.

Luzzati, D., « Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané », MIDL 2004, 29-30 novembre 2004, Paris.

Martin, P., « Winpitch Corpus, a software tool for alignement and analysis of large corpora », Workshop E-MELD 2003, Michigan State University (Etats-Unis), 2003.

Martin, P., « Intonation du Français : Parole spontanée et parole lue », in *Estudios de Fonética Experimental*, n°15, pp. 133-162, 2006.

Ogden, C., *Basic English: A General Introduction with Rules and Grammar*, Kegan Paul, Londres, 1932.

Pallaud, B., « Amorces de mots et répétitions : des hésitations plus que des erreurs en français parlé », JADT 2004, 10-12 mars 2004, Louvain-la-Neuve (Belgique).

Pallaud, B., « Troncations de mots, reprises et interruption syntaxique en français parlé spontané », JADT 2006, 19-21 avril 2006, Besançon.

Schapire, R. E. et Singer, Y. « BoosTexter: A boosting-based system for text categorization », Machine Learning, vol. 39, pp. 135-168, 2000.

Schiffrin, D., « Discourse markers : Language, Meaning, and Context », in *The Handbook of Discourse Analysis*, Blackwell Publisher, pp. 54-, 2001.

Shriberg, E., « Phonetic consequences of speech disfluency », Proceedings of the International Congress of Phonetics Sciences (ICPhS-99), pp. 619-622, 1999.

Siu, M.H. et Ostendorf, M., « Modeling disfluencies in conversational speech », ICSLP 1996, vol. 1, 1996.

Vaissière, J., « Utilisation de la prosodie dans les systèmes automatiques : un problème d'intégration des différentes composantes », in *Faits de Langues*, n°13, pp. 9-16, 1999.