Transcription et traitement manuel de la parole spontanée pour sa reconnaissance automatique

Thierry Bazillon

Université du Maine

Transcription et traitement manuel de la parole spontanée pour sa reconnaissance automatique

THÈSE

présentée et soutenue publiquement le 4 février 2011

pour l'obtention du

Doctorat de l'Université du Maine
(spécialité sciences du langage)

par

Thierry BAZILLON

Composition du jury

Présidente :	Mme Martine Adda-Decker	Directrice de Recherche	LPP, CNRS, Paris 3
Rapporteurs :	M. Jean-Yves Antoine	Professeur	LI, Université de Tours
	M. Henri-José Deulofeu	Professeur	LIF, Université d'Aix-Marseille
Examinateurs :	M. Daniel Luzzati	Professeur	LIUM, Université du Maine
	M. Yannick Estève	Professeur	LIUM, Université du Maine

Laboratoire d'Informatique de l'Université du Maine

Remerciements

J'ai toujours pensé que la présente page serait, au moment où je l'écrirais, un bien curieux paradoxe. Bien qu'étant *a priori* la seule de ce manuscrit à être dépourvue d'une quelconque connotation scientifique, sa rédaction demande pourtant une rigueur de tous les instants, alors même que d'aucuns penseraient que cet exercice n'est que fantaisie et gaudriole...

Que nenni ! Je sais que le moindre oubli, la moindre omission ou la plus infime maladresse me seront irrémédiablement et éternellement reprochés, bien plus que quelque médiocre résultat que j'aurais pu obtenir lors de mes travaux de recherche présentés ci-après. C'est donc avec une minutie de tous les instants que je vais désormais remercier les personnes qui, d'une façon ou d'une autre, m'ont permis de voir la fin d'une aventure longue de quatre années...

Et dans cette optique, comment ne pas commencer par Daniel Luzzati ? Témoin de mes tout premiers pas à l'Université du Maine, il a ensuite jalonné mes pérégrinations estudiantines avec une patience qui, parfois, a suscité ma propre admiration (ce qui n'est pas peu dire). De la conjugaison du présent de l'indicatif aux modèles de Markov cachés, notre route a connu quelques escapades gauloises pendant lesquelles il a bien voulu m'accompagner. Plus encore, il est sans cesse venu me chercher lorsque je préférais contempler le ciel pendant des heures plutôt que de marcher sur le chemin de la soutenance. Grâce à lui, je suis désormais arrivé au terme de ce périple : qu'il en soit ici sincèrement et chaleureusement remercié. Sans son aide, son soutien et ses conseils, le ciel n'aurait certes plus aucun secret pour moi... mais *quid* de la théorie du fenêtrage syntaxique ?

Je tenais à associer à ce premier paragraphe de remerciements Yannick Estève. Plus qu'un co-directeur de thèse, il fut tout d'abord la personne qui m'a permis d'intégrer le Laboratoire d'Informatique de l'Université du Maine. Pendant deux ans, j'ai ainsi vécu l'épique époque d'EPAC, qui fut le point de départ et la raison d'être de ma thèse. Sans me connaître, Yannick m'a fait confiance et m'a placé dans de très bonnes conditions pour réaliser mes travaux de recherche. L'histoire retiendra aussi que, sans le savoir, il m'a également placé dans de très bonnes conditions pour faire de la musique avec lui à nos heures (pas si) perdues. Je me souviens l'avoir entendu dire un jour : « La thèse, c'est un état d'esprit ». Peuchère, comme il avait raisong...

Je tenais ensuite à remercier les membres de mon jury. Tout d'abord Martine Adda-Decker, pour avoir accepté d'être la présidente dudit jury. J'ai souvent puisé dans ses travaux tout au long de la rédaction de cette thèse, et c'est un honneur pour moi qu'elle l'ait lue et jugée de façon positive.

Ensuite, Jean-Yves Antoine et Henri-José Deulofeu, qui ont accepté d'être les rapporteurs de mes travaux. Leurs remarques et la façon dont ils les ont exprimées m'ont permis de mieux cerner certaines limites de ce manuscrit, d'en corriger quelques aspects et de continuer à en approfondir d'autres.

Lors de mes deux années et demie passées au LIUM, de janvier 2007 à mai 2009, j'ai été amené à rencontrer nombre de personnes qui ont compté pour moi tout au long de la rédaction de cette thèse. En premier lieu, je voulais citer Julie Mauclair et Arnauld Séjourné, qui furent des compagnons de bureau éphémères mais ô combien agréables à côtoyer. Je crois savoir que leurs aventures dans le monde de la recherche continuent de plus belle, et je leur souhaite qu'il en soit ainsi pendant encore de longues années.

Comment ne pas maintenant évoquer ceux qui furent, durant deux ans, bien plus que de simples collègues ? Je mesure ainsi la chance que j'ai eue de partager tant de grands moments (scientifiques avant tout, bien sûr) avec Richard Dufour, Vincent Jousse et Antoine Laurent. Je reste convaincu que sans ce trio, mes souvenirs manceaux ne seraient pas aussi rieurs et nostalgiques à la fois. Avoir été à deux doigts de fonder le plus grand groupe de rock français de la décennie restera certes comme une déception, mais que l'histoire fut belle ! Il n'est pas donné à tout le monde d'être

enthousiaste en venant travailler le matin (ou en fin de matinée). Je l'étais, ne serait-ce qu'à l'idée de partager une nouvelle journée avec ceux que l'on appelle désormais « les hommes du 248 ».

Bien sûr, je n'oublierai pas les nombreuses autres personnes du LIUM avec qui j'ai eu des discussions tantôt scientifiques, tantôt moins scientifiques, mais qui m'ont toujours beaucoup apporté : l'équipe Parole bien sûr (Paul, Sylvain, Bruno, Teva et Carole) ; les ingénieurs Étienne et Bruno pour leurs multiples sauvetages et dépannages ; Mathilde (et Vincent) pour leur bonne humeur, les soirées Football ou Rock Band et tant d'autres choses ; Martine, pour nous avoir supportés en tant que voisins sans jamais porter plainte pour tapage diurne ; et tous ceux que j'oublie, mais qui ne m'en voudront pas ou si peu.

Je ne peux également passer sous silence les 16 mois passés au Laboratoire d'Informatique d'Avignon, et que je dois avant tout à Frédéric Béchet. Merci à lui de m'avoir fait confiance une première fois, pour ce qui allait être une bien belle aventure dans le Vaucluse. Aventure que j'ai partagée avec nombre de gens que je tiens ici à citer : Christophe, Nathalie, Juliette, Pierre C., Mickaël, Raphaël, Florian V. et Florian P., Nico, Benjamin, Corinne, Driss... Un merci tout particulier à Georges et Fabrice pour avoir pris le relais de Fred, et ainsi fait en sorte que je ne refasse pas mes valises trop vite...

Pour en finir avec mon Tour de France des laboratoires de parole, je tiens enfin à remercier ceux qui, aujourd'hui encore, sont mes collègues : les membres de l'équipe TALEP du Laboratoire d'informatique Fondamentale de Marseille. Frédéric Béchet remporte donc ici une seconde médaille, puisque c'est à nouveau grâce à lui que j'ai pu intégrer cette équipe, alors que j'apportais les dernières retouches à mon manuscrit. J'ai été (et je suis toujours) très heureux d'y rencontrer Alexis, Benoît, Joseph, Seljä, Ghasem, Jeff, Firas et Mélanie. J'espère pouvoir continuer à m'épanouir à leurs côtés, au milieu des lutins et de Mario.

Avant de conclure, je tenais aussi et surtout à remercier ma famille, qui a suivi mes travaux durant ces quatre années de recherche, et qui a été largement présente lors de ma soutenance. J'espère que celle-ci aura d'ailleurs permis de répondre aux nombreuses questions qu'elle se posait sur ma condition de recherche, et que je pourrais finalement résumer ainsi : « Mais ton travail, c'est quoi exactement ? »...

Enfin, je tenais particulièrement à dédier cette thèse à tous les membres du personnel de l'Hôpital Laënnec, à Nantes, que j'ai eu l'occasion de rencontrer. Ces travaux, ainsi que toutes les années d'études supérieures qui les précèdent, sont aussi les leurs. Sans eux, sans leur attention, sans leurs soins et sans leur extrême compétence, rien de tout cela n'aurait été possible. Qu'ils sachent que, plus que toute autre personne, je les remercie. Du fond du cœur.

Table des matières

Index des figures

Index des tableaux

Introduction

En mars 2010, le site Internet Rue89 a publié une enquête riche d'enseignements à propos du sous-titrage pour sourds et malentendants à la télévision[1]. Cette pratique, qui est désormais rendue obligatoire par une loi sur le handicap datant de 2005, peut s'opérer de différentes façons. Certaines chaînes de télévision, afin d'en réduire les coûts, utilisent un système de reconnaissance de la parole qui travaille en temps réel, même si le programme diffusé est quant à lui enregistré. Cependant, en cas d'erreur, il est très difficile de procéder à des corrections puisque l'affichage des sous-titres est quasi instantané.

La véritable question qui se pose est alors la suivante : quelles sont la fréquence et l'ampleur de ces erreurs ? Exemples et images à l'appui, le site Rue89 donne des réponses éloquentes. Confusions homonymiques, approximations, noms propres méconnaissables... Que peut comprendre une personne sourde lorsqu'elle lit des sous-titres avec de tels défauts, en nombre conséquent qui plus est ?

Cependant, il convient d'admettre que cet exemple isolé doit s'inscrire dans une problématique beaucoup plus large, et par là même beaucoup plus complexe. La reconnaissance automatique de la parole est en effet un défi qui passionne la communauté scientifique depuis de nombreuses années. Mais le mythe d'une machine qui, à défaut de les comprendre, retranscrirait parfaitement les propos d'un être humain s'est peu à peu fissuré. Ou plus exactement, il a perdu de son aura. Aujourd'hui, des résultats tutoyant le sans-faute sont par exemple envisageables avec des propos journalistiques : en dépit des inévitables problèmes d'homonymies ou de noms propres que nous évoquions précédemment, ce genre de parole est très formaté et convient parfaitement aux systèmes de reconnaissance de la parole (SRAP) actuels. En effet, ces derniers sont essentiellement entraînés pour être efficaces sur une langue naturelle dans ce qu'elle a de plus académique, c'est-à-dire telle qu'elle est présentée dans les manuels de grammaire.

Le véritable problème est de savoir si cette langue des manuels de grammaire est

1 http://www.rue89.com/2010/03/28/le-massacre-des-sous-titrage-pour-les-sourds-144363

également celle qu'utilise tout locuteur en situation d'énonciation. Depuis un recensement fondateur fait par l'équipe de Georges Gougenheim dans les années 60 [Gougenheim *et al.*, 1964], les études se sont multipliées pour répondre par la négative. Où apparaissent « ben » et « euh » dans les grammaires officielles ? Un locuteur s'exprimant naturellement prend-il vraiment le soin de structurer ses négations avec le discordanciel « ne » ? Comment sont réellement prononcées des expressions comme « c'est-à-dire » ou « parce que » ?

Pourtant, en dépit de cette distinction évidente entre deux modèles aux grammaires finalement bien différentes, il existe peu de données permettant de l'exploiter. En effet, il faut pour cela « capturer » cette langue naturelle, celle qui n'existe finalement pas ou peu dans les livres (Signalons toutefois la série du *Petit Nicolas* de Sempé et Goscinny, ou encore *Zazie dans le métro* de Raymond Queneau). Les progrès réalisés dans le domaine de la technologie audiovisuelle permet certes aujourd'hui de saisir « au vol » des conversations naturelles, avec une qualité tout à fait honorable. Cependant, outre différentes précautions juridiques plutôt lourdes qu'il faut prendre [Baude, 2006], c'est surtout l'exploitation des données récoltées qui demeure complexe. Il faut en effet retranscrire par écrit les enregistrements récoltés, en tâchant d'être le plus fidèle possible à ce que l'on entend. La transcription est déjà en soi une interprétation, qu'il faut donc essayer de rendre aussi « neutre » que possible. Une telle démarche est très coûteuse en temps, et peu de projets français se sont jusqu'à présent aventurés dans ce domaine.

C'est notamment pour pallier cette carence que le projet EPAC (Exploration de masse de documents audio pour l'extraction et le traitement de la PArole Conversationnelle) [Estève *et al.*, 2010] a été élaboré, à l'occasion d'un appel à projets de l'Agence Nationale de la Recherche en 2006. Moyennant la participation conjointe de quatre laboratoires de recherche (le LIUM[2], le LIA[3], l'IRIT[4] et le LI[5]), EPAC a débuté en mars 2007 et s'achèvera à la fin de l'année 2010. L'un des principaux buts de ce projet était donc d'effectuer la transcription d'une centaine d'heures d'enregistrements radiophoniques, essentiellement « conversationnels ». Nous nous attarderons longuement sur ce terme dans notre dernier chapitre mais il convient d'en donner une définition, même succincte, dès à présent.

La parole « conversationnelle », que nous qualifierons également de « spontanée » dans

2 Laboratoire d'Informatique de l'Université du Maine : http://www-lium.univ-lemans.fr/

3 Laboratoire d'Informatique d'Avignon : http://lia.univ-avignon.fr/

4 Institut de Recherche en Informatique de Toulouse : http://www.irit.fr/

5 Laboratoire d'Informatique de Tours : http://li.univ-tours.fr

notre étude, s'oppose par essence à la parole « préparée », à savoir celle employée par les journalistes ou toute autre personne ayant un support textuel lorsqu'elle s'exprime (prompteur, notes...). Ainsi, par le jeu des oppositions, la parole spontanée est donc une parole non préméditée, que l'énonciateur construit au fur et à mesure qu'il l'exprime. En conséquence, il n'est pas rare d'y voir apparaître ce que nous appellerons ici des disfluences [Adda *et al.*, 2005 ; Adda-Decker *et al.*, 2004] : répétitions, troncations, faux départs, etc. Ces disfluences sont en fait tout ce qui vient parasiter l'énoncé en lui-même, et qui bien souvent témoigne d'une hésitation de la part du locuteur. Sauf situation exceptionnelle, ces phénomènes n'apparaissent pas dans la parole préparée. D'une part parce qu'elle est bien souvent maniée par des professionnels, dont la communication est le métier, et qu'ils sont familiers de ce genre d'exercices ; d'autre part parce que le support textuel dont ils bénéficient (presque) systématiquement ramène bien souvent leur exercice d'énonciation à une simple lecture.

On comprend donc bien ici que la parole préparée pose moins de problèmes aux SRAP que la parole spontanée, notamment à cause de ces disfluences qui modifient régulièrement le rythme ou la prononciation « traditionnelle » d'un discours. Par exemple, l'élision fréquente du « e » muet dans la parole spontanée provoque de nombreux phénomènes d'assimilation entre consonnes sourdes et sonores, ce qui a tendance à tromper les SRAP.

Prenons en exemple l'expression « pas de problème ». Une prononciation naturelle, sans tenir compte d'une quelconque accentuation régionale, élidera fréquemment le « e » du mot « de ». Ce qui devrait en théorie donner la séquence phonétique suivante : « padpRoblEm ». Mais comme un « d » (consonne sonore) et un « p » (consonne sourde) se trouvent ainsi confrontés, l'appareil phonatoire va naturellement modifier la prononciation du « d » pour le faire glisser vers un « t ». La prononciation exacte sera donc « patpRoblEm », d'où une confusion avec la séquence « patte problème » par exemple. La succession de ces deux mots paraît certes absurde en tant que tel, mais nous verrons qu'il est très difficile pour un SRAP de mesurer ce non-sens.

Notre rôle au sein du projet EPAC a donc été, en tant qu'ingénieur linguiste, de transcrire et d'annoter le corpus de cent heures évoqué précédemment. L'enjeu était double : permettre au SRAP du LIUM (LIUM RT) d'améliorer ses performances sur la parole spontanée, et mettre à la disposition de la communauté parole un corpus conséquent, dans un domaine jusqu'alors peu fourni.

Pour ce faire, nous nous sommes appuyés sur TRANSCRIBER, un logiciel qui permet

d'aligner facilement du texte avec un signal audio. Cette thèse a donc été également l'occasion d'acquérir des compétences spécifiques à cet outil, mais également d'être confronté à la plupart des autres logiciels d'aide à la transcription existants (PRAAT, WINPITCHPRO, TRANSANA, EXMARaLDA, ANVIL, etc.). Leur diversité de formats et de codage des données a entraîné une réflexion sur l'interopérabilité des données : une transcription réalisé avec PRAAT peut-elle être utilisée ensuite avec ANVIL ? Est-ce que les balises spécifiques de TRANSCRIBER peuvent être converties d'une quelconque façon sous WINPITCHPRO ? Notre étude montrera que la réponse à ces questions est très souvent négative, la faute à un manque d'uniformité patent.

Dans cette optique, les recommandations proposées par le consortium de la TEI (Text Encoding Initiative) apparaissent salutaires. Souhaitant proposer une méthode efficace pour coder de façon numérique toutes sortes de données, la TEI rédige depuis vingt ans des *Guidelines*[6] basées sur le langage de balisage XML, gage de pérennité et d'interopérabilité. Mais c'est seulement depuis quelques années que la TEI tend à se répandre et à se démocratiser, grâce à une organisation de plus en plus structurée et étoffée. Nous verrons toutefois qu'elle n'est pas encore véritablement intégrée au domaine de la parole, et notamment que certains logiciels de transcription ont des formats de sortie qui en sont très éloignés.

Notre travail de thèse se décomposera donc en cinq axes : tout d'abord, nous détaillerons notre parcours de recherche en insistant longuement sur les tenants et les aboutissants du projet EPAC, et plus précisément sur notre expérience de la transcription à travers le sous-projet « Annotation et évaluation ».

Ensuite, le deuxième chapitre sera consacré de façon plus générale à la transcription de la parole. Nous tenterons de voir comment cette pratique a pu évoluer au fil des avancées technologiques (avec notamment l'apport des systèmes de reconnaissance automatique de la parole), et comment elle se matérialise concrètement aujourd'hui. Nous procéderons notamment à un inventaire aussi complet que possible des corpus de parole accessibles aujourd'hui.

Les différents outils qui permettent de transcrire de la parole et surtout de la synchroniser avec un flux audio – ce qui les différencie d'un traitement de texte classique – seront ensuite passés en revue. Plus communément appelés « logiciels d'aide à la

6 Ensemble de recommandations

transcription », ils tendent à se multiplier depuis quelques années. Nous proposerons donc une analyse détaillée d'une dizaine d'entre eux, parmi les plus répandus à l'heure actuelle.

Ce banc d'essai nous amènera à traiter la question des formats et du codage des données, que nous évoquions précédemment : la TEI est-elle le moyen d'uniformiser des données transcrites avec différents logiciels, et peut-elle ainsi permettre une utilisation optimale de celles-ci ?

Enfin, le dernier chapitre sera consacré à cette parole que nous avons qualifiée de « spontanée ». Par le biais de différents définitions, angles d'approches et expériences, nous tenterons de voir tout ce que recouvre ce terme, et s'il est finalement bien approprié aux situations qui lui sont souvent attribuées.

Chapitre 1 : Parcours de recherche

Le point de départ de notre thèse était donc le projet EPAC, que nous détaillons dans le chapitre à venir. Plus précisément, au sein même de ce projet EPAC, la section « Annotation et évaluation » comportait un objectif principal : la constitution d'un corpus de parole spontanée d'une centaine d'heures, issues des données audio non transcrites de la campagne d'évaluation ESTER [Galliano *et al.*, 2005]. Pour ce faire, nous avons occupé un poste d'ingénieur d'études à mi-temps sur 24 mois, qui a permis de financer nos travaux de recherche tout en en générant la matière première.

Nous présenterons donc dans ce premier chapitre le projet EPAC, de façon générale puis en nous intéressant aux différents sous-projets qui le composent. Puis, nous évoquerons deux autres projets de recherche ayant quelques similitudes avec EPAC : VARILING et RHAPSODIE.

Nous détaillerons ensuite le travail spécifique de transcription qui a été effectué à l'intérieur d'EPAC, dans le sous-projet intitulé « Annotation et évaluation » : corpus réalisé, méthodes de transcription, logiciel(s) utilisés, difficultés rencontrées...

Enfin, nous évoquerons l'avenir du corpus réalisé entre mars 2007 et mars 2009. De par sa richesse de données mais aussi de métadonnées, le corpus EPAC peut se prêter à toute sorte d'expériences et de recherches (morpho-syntaxiques, lexicales, sociologiques). Nous en avons déjà réalisées quelques-unes dont nous présenterons ici les résultats, mais bien d'autres restent encore à mener.

1.1 Le projet EPAC : présentation générale

Sélectionné par l'ANR dans le cadre de l'appel à projets 2006 du programme Masse de Données - Connaissances Ambiantes (MDCA), le projet EPAC (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, coordinateur). 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 2000 heures d'enregistrements radiophoniques, dont 1800 proviennent de la campagne ESTER.

Celle-ci s'est déroulée de janvier 2003 à janvier 2005 et avait pour but d'évaluer les systèmes d'indexation automatique d'émissions radiophoniques en français. Menée conjointement par l'AFCP, la DGA, la CTA ainsi que les organismes ELDA et ELRA, la campagne ESTER faisait partie du projet technolangue EVALDA, financé par le Ministère de la Recherche. L'un des objectifs principaux d'ESTER était de produire des corpus qui soient accessibles à l'ensemble de la communauté parole.

Pour réaliser l'ensemble des tâches d'évaluation (transcription, mais également segmentation et extraction d'informations), plus de 2000 heures de données audio ont été mises à la disposition des participants (dont le LIUM). Ces données provenaient de radios comme France Culture, France Info ou France Inter. Concernant la transcription, seule une centaine d'heures a été traitée, ce qui laissait à disposition une masse de données conséquente pour des projets ultérieurs. C'est donc sur ces quelques 2000 heures non transcrites que s'est appuyé le projet EPAC pour mener à bien sa tâche de transcription.

Au sein de ce vaste corpus, la parole conversationnelle / spontanée 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 journaux d'informations, 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 nature 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. C'est principalement sur ce sous-projet que nous nous attarderons, non sans avoir préalablement présenté et défini l'ensemble des tâches du projet EPAC.

1.1.1 Sous-projet 1 : Management du projet

Ce premier sous-projet est celui qui assure la coordination générale et le bon déroulement d'EPAC. Les travaux à réaliser sont ainsi plutôt de l'ordre de la communication et de la cohésion entre les quatre partenaires du projet : création d'un site WEB (http://epac.univ-lemans.fr) et d'une mailing list, gestion des réunions de travail (téléphoniques ou sur site), rédaction des rapports d'avancement du projet... Bien que dépourvues d'enjeu spécifique à proprement parler, ces tâches permettent de jalonner les 36 mois du projet.

1.1.2 Sous-projet 2 : Extraction de caractéristiques acoustiques de bas niveau

Ce sous-projet a pour but de segmenter des fichiers audio en termes d'événements acoustiques : parole, musique, locuteurs, langues, etc. Dans ce type de tâche, il est possible d'envisager deux points de vue. Le premier consiste à traiter les documents sonores concernés (issus du corpus ESTER) sans connaissance *a priori*, en ne tenant compte d'aucune informations ayant trait à leur contenu. Mais une seconde optique est envisageable, qui consiste à utiliser des métadonnées (tranches horaires couvertes, afin d'identifier les programmes concernés ; annotations s'il y en a...) pour faciliter le processus d'indexation.

Ce sous-projet est traité en trois temps :

- segmentation des documents audio en événements acoustiques primaires
- mesure de l'impact que peuvent avoir des métadonnées comme les grilles de programme, l'identité des locuteurs, les résumés d'émissions...
- utilisation et combinaison de différentes stratégies d'extraction, afin d'améliorer les

performances de cette indexation d'événements acoustiques.

1.1.3 Sous projet 3 : Reconnaissance automatique de la parole conversationnelle

Les systèmes de reconnaissance automatique de la parole du LIUM et du LIA ont obtenu de bons résultats lors de la campagne d'évaluation ESTER. C'est pourquoi ils sont utilisés pour transcrire automatique les 1700 heures de cette campagne qui ont été fournies sans aucune annotation manuelle.

Les enjeux de cette tâche sont multiples. Le premier, et le plus important d'entre eux, est de fournir des informations lexicales concernant la prononciation des mots, mais également au sujet des fillers (hésitations, morphèmes spécifiques, bruits, disfluences...). Cela est notamment possible grâce aux résultats fournis par le sous-projet 2, qui alimentent également les sous-projets 4, 5, 6 et 7. Concernant ce dernier, le but est principalement d'accélérer le processus de transcription, en le faisant passer d'une méthode manuelle (très coûteuse en temps) à une démarche assistée, grâce au système de reconnaissance automatique du LIUM, LIUM RT.

Le deuxième enjeu consiste donc précisément à améliorer les performance des SRAP du LIUM et du LIA, en en combinant les sorties. En utilisant des mesures de confiance[7] fiables, procéder de la sorte permet en outre d'espérer une amélioration spécifique à chacun des deux systèmes : si l'un obtient de faibles mesures de confiance, son score peut être compensé par celui de l'autre, pour peu que ses propres mesures de confiance soient élevées.

Enfin, le troisième objectif de ce sous-projet est d'adapter les systèmes de reconnaissance à la parole conversationnelle. En effet, et nous le montrerons au cours de cette étude, les SRAP sont beaucoup moins à l'aise avec ce genre de donnée qu'avec de la parole préparée, notamment parce que la parole conversationnelle contient souvent de nombreuses disfluences. Qui plus est, les données peuvent y être bruitées, la vitesse d'élocution très rapide et les segments de parole superposée fréquents. Enfin, les élisions et/ou assimilations qui caractérisent la prononciation du français parlé jouent elles aussi un rôle prépondérant dans les performances des SRAP, peu familiarisés avec ce type de parole. En effet, les données disponibles sont en faible quantité, et l'apprentissage des systèmes est donc difficile à réaliser.

Grâce à la constitution d'un corpus de 100 heures de données majoritairement

7 Estimation de la fiabilité d'une hypothèse de reconnaissance [Mauclair, 2006]

spontanées, il y a de bonnes raisons d'espérer que le taux d'erreur mots[8] puisse chuter. L'objectif est donc de parvenir à une réduction de 25% (en relatif) par rapport au chiffre initial.

1.1.4 Sous-projet 4 : Identification nommée du locuteur

Une fois les données audio segmentées en événements acoustiques (sous-projet 2) puis transcrites orthographiquement, il convient de s'intéresser à l'identification nommée[9] du locuteur. En effet, à l'heure actuelle, les SRAP sont seulement capables d'attribuer un identifiant anonyme à chaque tour de parole. Il s'agit donc ici d'être en mesure, sans connaissance *a priori*, de proposer une méthode qui soit en mesure d'associer à chaque locuteur un nom et/ou un prénom. Les travaux antérieurs dans ce domaine ont souvent recours à une banque de voix préalablement enregistrées, qui permet d'entraîner les modèles acoustiques. Ainsi, s'ils sont confrontés à un locuteur dont la voix se trouve dans la base de données, ils peuvent espérer être en mesure de l'identifier nommément par assimilation acoustique. Toutefois, ce processus peut être source d'erreurs car les caractéristique acoustiques de deux voix sont parfois très proches.

Des méthodes linguistiques sont également envisageables. En effet, lors d'une prise de parole, il est fréquent qu'un locuteur se présente – souvent en fin de reportage -, ou alors qu'il soit présenté par le locuteur antérieur (« Marc Crépin, Washington, France Info » ; « À Washington, nous retrouvons Marc Crépin »). Cette information peut théoriquement être utilisée pour associer automatiquement le nom fourni au tour de parole correspondant. Cependant, il faut cette fois disposer non plus de données audio, mais également de la transcription orthographique segmentée en tours de parole.

Ce traitement linguistique a obtenu des résultats intéressants lors d'études préalables : le LIUM est parvenu, en le combinant avec la méthode acoustique, à un taux d'identification de 70% des locuteurs sur un journal d'informations. L'objectif de ce sous-projet est donc de travailler à l'amélioration de ces performances, afin de proposer un système fiable et automatisée d'identification nommée des locuteurs.

8 Unité de mesure indiquant le pourcentage de mots mal reconnus par un SRAP par rapport à une référence

9 Processus consistant à détecter automatiquement l'identité d'un locuteur à l'intérieur d'un flux de parole

1.1.5 Sous-projet 5 : Traitement du langage naturel

Comme nous l'avons dit, le projet EPAC a pour but d'extraire des informations au sein d'un corpus audio. De façon plus précise, il convient ici d'analyser linguistiquement les transcriptions automatiques fournis par LIUM RT, le système de reconnaissance automatique du LIUM. Cette tâche est complexe dans la mesure où elle concerne de la parole spontanée. Il faut donc s'attendre à des données avec des taux d'erreur mots élevés.

Ce traitement s'articule autour de quatre axes :

- La détection d'entités nommées, qui est un élément majeur dans le domaine de l'extraction d'informations. Repérer les noms propres, notamment, permettra de faciliter le processus d'identification nommée que nous venons d'évoquer.

- L'analyse syntaxique de surface, c'est-à-dire procéder à un découpage en unités syntaxiques minimales (groupe nominal, groupe verbal...). Par extension, celui-ci fournit également des informations sémantiques pertinentes, qui permettent d'autres traitements *a posteriori*. Par ailleurs, eu égard à la parole spontanée, ce genre de découpage est essentiel puisque fortement influencé par les disfluences. Il offre donc la possibilité de mesurer les difficultés que pose la parole spontanée lorsque l'on souhaite extraire des informations fines, et met en évidence quelques-unes de ses manifestations (ruptures de syntaxe, faux départs...)

- La classification de la parole conversationnelle est quant à elle l'occasion d'étiqueter les données transcrites sous des unités thématiques cohérentes. En premier lieu, suivant les velléités énonciatives des locuteurs (assertions, interrogations, réponses...), mais également en fonction du genre d'émissions concerné : débats, interviews, entretiens, témoignages, etc.

- La détection d'opinion se rapproche quelque peu de la classification ci-dessus, si ce n'est qu'il s'agit d'évaluer la position du locuteur par rapport au thème global de l'interview ou du débat auquel il participe. Y est-il favorable, opposé ou nuancé ? A-t-il un jugement objectif face à la situation qu'on lui présente ? Cette tâche, combinée à la précédente, doit permettre d'effectuer des requêtes sur les données étiquetées, avec un niveau de granularité relativement important.

1.1.6 Sous-projet 6 : Structuration et agrégation

Les recherches à mener dans ce sous-projet portent sur deux grandes thématiques. La première consiste à établir la structure d'un document audio, pour notamment mettre en évidence les zones de parole conversationnelle. À ce titre, des éléments comme les jingles, les

pages de publicité ou les remerciements en fin d'émissions sont autant d'indices pertinents, qui peuvent être détectés de façon automatique, afin de fournir une macro-segmentation. Concernant la parole conversationnelle, sa détection sert d'autres sous-projets : avant qu'un débat ou une interview ne commence réellement, il est souvent fait mention du nom des intervenants (SP 4) et du thème abordé (SP 5, classification de la parole conversationnelle).

Cette structuration doit également permettre de comparer ou de classer des documents audio grâce aux indices recueillis. Suivant la granularité et le critère choisi, on peut par exemple associer différents documents selon qu'ils partagent la même thématique, les mêmes locuteurs, etc. De même, il doit être possible, grâce aux jingles par exemple, de « fouiller » les données afin de repérer et regrouper les bulletins météorologiques ou les chroniques boursières.

La seconde thématique inhérente à ce sous-projet concerne le regroupement de documents audio présentant des informations temporelles similaires. Sans connaissance *a priori*, elle consiste à utiliser des indices acoustiques (locuteurs, applaudissements, etc.) pour détecter les rôles des intervenants (présentateur, invités) et les différentes séquences d'une émission radiophonique.

1.1.7 Sous-projet 7 : Annotation et évaluation

Dernier axe d'EPAC, ce sous-projet a principalement trait à la transcription et à l'annotation de données audio. Plus précisément, il convient dans un premier temps de déterminer, au sein de la partie non-transcrite du corpus ESTER (1700 heures), les zones contenant de la parole conversationnelle. Ce corpus étant avant tout un corpus journalistique, celles-ci ne sont pas majoritaires, et une telle exploration peut donc demander du temps. Ensuite, il s'agit d'assurer la transcription des données récoltées, à hauteur d'une centaine d'heures. Pour ce faire, un processus de transcription assistée est mis à l'épreuve. Le transcripteur bénéficie des sorties automatiques générées par LIUM RT, ce qui fait glisser sa tâche de la transcription vers la correction. Cet apport doit permettre un gain de temps, qui dépend toutefois des données auxquelles LIUM RT est confronté. Sur de la parole très conversationnelle, ses performances sont beaucoup moins bonnes que sur un flux journalistique par exemple. L'un des objectifs de ce sous-projet est donc de mesurer le gain de temps pouvant être obtenu en recourant à la transcription assistée. Avant le lancement du projet EPAC, on estimait que pour une heure de données audio, 50 heures de traitement

manuel étaient nécessaires pour une granularité d'annotation moyenne (noms des locuteurs, thèmes des émissions, informations acoustiques...). Parallèlement, le LIUM avait mené des évaluations internes laissant à penser qu'un processus assisté permettait d'aller jusqu'à cinq fois plus vite. Nous verrons à la fin de notre deuxième chapitre où se situent finalement ces chiffres par rapport à la réalité d'une expérience scientifique.

Ce facteur temps est sans doute l'élément essentiel de la bonne tenue de ce sous-projet. L'avantage principal du corpus EPAC est que les données audio sont déjà recueillies, ce qui permet de s'affranchir d'une phase de collecte particulièrement longue et coûteuse. Pour autant, il faut en premier lieu définir précisément ce que referme l'appellation « parole conversationnelle ». Se limite-t-elle à un cadre interactif spécifique, à un genre d'émissions bien précis, à une élocution particulière ? Des expériences sur ces réflexions seront menées au cours de cette étude, afin de cerner plus précisément ce problème.

Par ailleurs, transcrire de la parole conversationnelle dans l'optique du projet EPAC suppose des conventions d'annotation particulières, afin de mettre en lumières certaines spécificités de la langue parlée : troncations, faux départs, répétitions, élisions... Les recommandations d'ESTER[10] sont, en grande partie, la base de ce travail d'annotation, moyennant quelques ajouts ou modifications selon les besoins du projet. Il faut être conscient que cette couche d'informations supplémentaires signifie un traitement plus long. Cependant, elle s'avère indispensable en vue des interactions avec les autres sous-projets, et également dans l'optique de travaux futurs. A ce titre, le corpus constitué sera distribué par ELRA sous des conditions financières avantageuses.

1.1.8 Résultats et conséquences attendus

Une fois arrivé à son terme, c'est-à-dire à la fin de l'année 2010, le projet EPAC devrait avoir permis plusieurs avancées significatives dans le domaine de la parole, et ce pour plusieurs raisons :

- Tout d'abord, les outils développés durant le projet seront mis à disposition sous licence open-source. L'exemple d'ALIZE[11], plate-forme consacrée à la reconnaissance du locuteur développée au LIA, est la première étape de cette distribution. Les autres outils viendront compléter cette démarche au fur et à mesure de l'avancement du projet.

10 http://www.afcp-parole.org/ester/docs/Conventions_Transcription_ESTER2_v01.pdf

11 http://mistral.univ-avignon.fr/

- Ensuite, dans le cadre du sous-projet 7, le corpus de parole conversationnelle d'environ 100 heures a déjà été constitué. Il vient donc compléter celui de taille équivalente qui avait été réalisé lors de la campagne d'évaluation ESTER. Lorsque ses modalités de distribution seront finalisées, la communauté parole française aura ainsi à sa disposition (et le terme n'est pas anodin, au vu des problèmes de disponibilité des données existantes) un corpus de parole de taille conséquente. Qui plus est, EPAC se focalisant sur la parole conversationnelle, l'apport de ce nouveau corpus est double puisque les 100 heures déjà disponibles ont essentiellement pour cible de la parole préparée.

- Outre ces données transcrites manuellement (ou semi-automatiquement), la partie restante du corpus non-transcrit ESTER a été transcrite automatiquement, ce qui représente un ensemble de 1700 heures. Celui-ci va donc constituer une banque de données précieuse pour les chercheurs, et sa mise à disposition dépendra des mêmes conditions que celle du corpus transcrit manuellement.

- De façon plus générale, le projet EPAC doit donner une réelle impulsion à la communauté « parole » francophone, à la suite de la campagne ESTER. La deuxième phase de celle-ci, en novembre 2008, a déjà été l'occasion de tester et de confronter les premiers résultats obtenus.

- Enfin, il est à souhaiter que le projet permette d'établir des ponts entre informatique et linguistique. En effet, le faible nombre de corpus de parole conversationnelle est non seulement un handicap pour les systèmes de reconnaissance automatique de la parole, mais également pour les théories linguistiques qui manquent de matière sur laquelle s'appuyer. En conséquence, les deux communautés ont un intérêt commun à la réussite du projet EPAC. La constitution du corpus de 100 heures de parole spontanée est un premier pas qui va dans ce sens, mais il ne doit pas être sans suite.

Cependant, quelques autres équipes scientifiques se sont elles aussi engagées dans cette voie de la complémentarité. Des projets de grande ampleur tels que VARILING ou RHAPSODIE ont ainsi vu le jour, proposant à leur tour une collaboration entre linguistique et informatique. Des corpus de parole conséquents devraient ainsi émerger de ces deux projets actuellement en cours, que nous allons maintenant présenter en détails.

1.2 Autres projets « corpus » similaires

1.2.1 VARILING

VARILING est un projet coordonné par le CORAL (Centre Orléanais de Recherche en anthropologie et Linguistique), et auquel participent les laboratoires MoDyCo (Modèles Dynamiques Corpus), LACITO (Langues et Civilisation à Tradition Orale) ainsi que le LI (Laboratoire d'Informatique de Tours). Son objectif principal est d'accroître l'expertise en français sur la variation de la langue et son analyse, et ce à plusieurs niveaux : reconnaissance des données, étiquetage, traitements...

VARILING présente plusieurs points communs avec EPAC : tout d'abord, il est lui aussi financé par l'Agence Nationale de la Recherche. Ensuite, l'un de ses buts est également de fournir à la communauté parole un corpus de parole spontanée. Cela étant, l'ampleur est toute autre puisqu'il est question avec VARILING d'apporter des données à hauteur de 700 heures, soit environ 10 millions de mots.

Pour parvenir à de tels résultats, le projet s'appuie majoritairement sur les deux corpus ESLO (Enquêtes Socio-Linguistiques à Orléans). Le premier, ESLO 1, a été constitué en 1968 par des universitaires britanniques. Le but était alors d'enregistrer puis de transcrire le français en tant que langue « orale », en vue de son enseignement dans le système éducatif anglais. Plus de 300 heures d'enregistrements (4,5 millions de mots) ont ainsi été collectées, réparties entre interviews en face-à-face, conversations téléphoniques, réunions, repas de famille, etc. Les locuteurs et locutrices choisis sont volontairement représentatifs de toutes les catégories socio-professionnelles, et sont originaires de différentes régions. Toutefois, ce corpus a priori immense demeurait très incomplet lorsqu'il a été récupéré par les membres de VARILING : les transcriptions, réalisées sur de simples feuilles de papier, se sont révélées incomplètes et approximatives. Quant aux documents sonores, il a fallu travailler à partir des bandes magnétiques d'époque pour « nettoyer » et numériser l'ensemble. Tout cela sous-entend un travail de traitement considérable, d'autant que toutes les transcriptions doivent être reprises intégralement, pour les faire passer du format papier au format informatique. L'enjeu est cependant de taille, puisque ce corpus représente un témoignage linguistique unique, datant tout de même de plus de quarante ans...

Parallèlement au travail sur ESLO 1, le CORAL a également mis en place ESLO 2. Il s'agit en fait de créer un corpus « miroir » du premier, presque un demi-siècle plus tard. Il est prévu de collecter environ 400 heures d'enregistrements, ce qui permettrait à l'ensemble des données ESLO d'atteindre la barre des 10 millions de mots. Ce chiffre est notamment comparable aux données orales que propose le BNC (British National Corpus), une référence en la matière.

Par ailleurs, le projet VARILING prévoit également de rapprocher les deux corpus ESLO d'autres corpus du français, comme celui de Lormont par exemple. Lormont est une communauté urbaine au nord-est de Bordeaux, où a été menée une enquête socio-linguistique entre 1980 et 1981. Celle-ci portait notamment sur l'élision ou non du schwa en fin de polysyllabe, phénomène plutôt prégnant dans cette région. VARILING envisage de reproduire la même enquête aujourd'hui, de façon à mener une enquête en diachronie permettant de mesurer l'évolution de ce phénomène.

Ce projet est très prometteur en ce qui concerne la disponibilité et l'accessibilité des corpus ESLO 1 et ESLO 2, une fois ceux-ci terminées. Ils devraient être mis en ligne sur le site Web du projet en juin 2010, avec une interface permettant une utilisation « intelligente » des données (recherches statistiques, lexicales, morpho-syntaxiques...).

1.2.2 RHAPSODIE

RHAPSODIE est un autre projet relatif à la constitution de corpus oraux, bénéficiant également du support de l'Agence Nationale de la Recherche. Coordonné par Anne Lacheret-Dujour, celui-ci est porté par les laboratoires MoDyCo, IRCAM, LATTICE, ERSS ainsi que le LPL. RHAPSODIE a pour objectif premier de créer un « corpus de référence de français parlé, échantillonné en différents genres discursifs et doté d'annotations prosodiques et syntaxiques ». La notion de prosodie est très importante ici, car peu de corpus français bénéficient d'informations précises sur ce domaine. Le corpus attendu est cependant d'une taille assez modeste (environ 30 heures) si on le compare à EPAC ou aux ESLO, d'autant que les annotations syntaxiques et prosodiques ne portent que sur un total de 12 heures. Cela étant, de nombreuses démarches sont effectuées au sein de RHAPSODIE pour le rendre accessible et exploitable facilement, ce qui est loin d'être une pratique répandue. Ainsi les recommandations du guide d'Olivier Baude [Baude, 2006] sont-elles suivies à la lettre concernant les domaines déontologiques et juridiques, de même que la TEI et XML sont

utilisés pour le formatage des données.

Outre les transcriptions et les annotations, RHAPSODIE prévoit également la mise au point d'outils de fouille de données, permettant d'interroger le corpus *via* un système de requêtes. De même, des algorithmes d'étiquetage et de segmentation sont implémentés dans le corpus et peuvent facilement être réutilisables.

Enfin, RHAPSODIE s'appuie sur une équipe nombreuse (plus de quarante membres participants sont référencés sur le site du projet) et bénéficie de plusieurs partenaires : l'ATILF, CORAL, le CRDO de Paris, l'INRIA...

1.3 La transcription et l'annotation du corpus EPAC

Le corpus EPAC a été réalisé entre mars 2007 et mars 2009. Il a nécessité un apprentissage spécifique de la tâche de transcription ainsi que la maîtrise de TRANSCRIBER, l'outil informatique utilisé dans le cadre de ce travail. Avant même de commencer la transcription, certains problèmes ont dû être résolus, tandis que d'autres se sont posés en cours de route. Nous tenterons donc ici de synthétiser l'ensemble de ces questions, en souhaitant notamment que ce paragraphe puissent servir de base de travail à d'éventuels néo-transcripteurs.

1.3.1 Introduction

L'étape liminaire de ce travail était la maîtrise de TRANSCRIBER, le logiciel retenu pour réaliser les transcriptions. Cet apprentissage a été facilité par les qualité d'ergonomie et de prise en main de cet outil, que nous aurons l'occasion d'évoquer en 3.1. Les premières transcriptions réalisées nous ont permis notamment de paramétrer de nombreux raccourcis clavier, synonymes de précieux gain de temps par la suite. De même, la gestion des balises et de l'alignement temporel a demandé un léger temps d'adaptation pour en profiter pleinement.

Une fois le logiciel maîtrisé, il convenait de mesurer sans tarder la qualité du travail effectué, et de s'assurer que la transcription était conforme à l'exigence du projet. Ainsi, après quelques semaines de pratique, l'expérience suivante a été menée : transcrire un fichier audio *déjà* transcrit par la DGA avec TRANSCRIBER, et comparer les deux productions. Les différences ont été relativement peu nombreuses, ce qui a permis de valider le processus de transcription et de commencer réellement la construction du corpus.

1.3.2 Corpus de transcription

Comme l'indique explicitement l'intitulé du projet EPAC, le corpus en question devait être un corpus de parole spontanée (ou conversationnelle). Il convenait donc de rechercher, au sein des données audio non transcrites de la campagne ESTER, les extraits susceptibles d'être pertinents pour notre travail. Quatre stations de radio étaient concernées : France Inter, France Info, France Culture et RFI. Il est apparu presque naturel de rejeter d'office France Info, puisque le principe même de cette radio est de diffuser de l'information (donc de la parole

préparée), et ce de manière quasi permanente. Il nous faut néanmoins reconnaître que l'écoute attentive de cette radio a permis de détecter, sporadiquement, quelques extraits de parole spontanés (interview au sein d'un journal par exemple). Toutefois, ils étaient trop rares et trop peu nombreux pour qu'il soit judicieux de les prendre en compte. Le simple fait de les repérer dans la masse d'information émise demandait déjà beaucoup de temps.

Ainsi, la constitution de notre corpus s'est limitée aux trois autres radios : France Culture, France Inter et RFI. Il restait à détecter les zones contenant la plus grande proportion de parole spontanée. Pour ce faire, la seule méthode finalement fiable s'est révélée être celle de l'écoute minutieuse. Les procédures de détection automatique que nous avons pu éprouver se sont avérées trop imprécises pour nous fournir le niveau de fiabilité souhaité. Par ailleurs, la richesse des données audio disponibles (environ 2000 heures) assurait la présence itérative de certaines plages horaires intéressantes. De nombreuses occurrences d'émissions telles que « Le téléphone sonne », « Culture vive » ou « Sous les étoiles exactement », particulièrement susceptibles de contenir de la parole spontanée, ont ainsi retenu notre attention. Une fois la plage horaire détectée, il ne restait plus qu'à chercher les fichiers dans lesquels elle apparaissait de nouveau. Cette méthode certes empirique a permis, de façon générale, de ne perdre que peu de temps dans l'exploration des données. En effet, les émissions précitées durent environ 40 minutes et les propos tenus y sont essentiellement spontanées, ce qui a permis de façonner notre corpus de façon efficace et rapide.

Ayant été réalisées avec le logiciel TRANSCRIBER, les transcriptions sont disponibles sous la forme de fichiers .trs. Parfois, les fichiers audio ont été transcrits dans leur intégralité, mais cela reste plutôt marginal car il était bien rare qu'une plage horaire longue d'une ou de plusieurs heures ne contienne que de la parole spontanée. Ainsi, nombre de nos fichiers TRANSCRIBER contiennent une voire deux parties étiquetées « nontrans », correspondant au début et/ou à la fin du signal audio, lorsque les données qui nous semblaient pertinentes se trouvaient « encerclées » par de la publicité, de la musique ou surtout de la parole trop peu « spontanée ».

De la même manière, au sein même d'un fichier volumineux où plusieurs débats, émissions, etc. nous semblaient intéressantes à transcrire, il est arrivé que des parties « intermédiaires » assez peu pertinentes pour nos travaux viennent se greffer. Dans ce cas, plusieurs sections « nontrans » pouvaient parfois jalonner notre fichier de transcription. Toutefois, il est important de préciser que ce type d'annotation contient un ancrage temporel,

et qu'il permet ainsi à la transcription de toujours rester alignée avec le signal audio, même si celle-ci contient plusieurs passages « blancs ».

Cet alignement a d'ailleurs été une de nos priorités, afin de permettre à nos transcriptions d'être utilisées sans problème par toute personne possédant ou acquérant les données sonores correspondantes. Pour ce faire, il a fallu éviter de découper les fichiers audio, même si nous n'en avons exploité qu'une petite partie. Plus exactement, il a fallu éviter de modifier le T0 des fichiers audio. En effet, retrancher la fin d'un fichier son ne posait guère de problèmes, dans la mesure où TRANSCRIBER aligne le début de la transcription avec le début du signal audio qu'on lui adjoint. Mais, à l'opposé, réaliser une transcription en ayant amputé le signal audio, ne serait-ce que de quelques segments publicitaires précédant une émission, rend très délicat son utilisation avec le fichier audio original. Il faut, dans pareil cas, réaligner manuellement tous les segments transcrits, un à un, ce qui est une perte de temps considérable. Tous les fichiers du corpus EPAC sont donc alignés temporellement avec les fichiers audio « bruts » du corpus ESTER.

Toujours dans un souci de lisibilité et d'utilisation optimales, nous avons également documenté chacune des transcriptions réalisées : durée du fichier, informations sur les données transcrites, locuteurs, et parfois indications sur la période où la transcription a été réalisée.

1.3.3 Difficultés liminaires

Transcrire de la parole spontanée présente de nombreuses difficultés. La première que nous mentionnerons est également la première à laquelle se trouve nécessairement confronté tout transcripteur, à savoir la perception auditive. En effet, si un journal d'informations ne nécessite bien souvent qu'une écoute par segment de parole pour pouvoir être transcrit sans erreur, il n'en va pas de même pour un débat où plusieurs invités se coupent la parole, hésitent, bégayent, bafouillent... Il s'agit alors de discerner ce qui est effectivement prononcé, par qui et à quel moment. Les écoutes multiples deviennent donc indispensables pour peu que l'on souhaite une transcription rigoureuse. Toutefois, cela va de pair avec une perte de temps dès lors qu'on se trouve face à des masse de données à transcrire, comme nous le montrerons dans le deuxième chapitre.

Outre cette difficulté majeure, la transcription fait rapidement apparaître un autre souci saillant : l'orthographe. En premier lieu, celle de la langue française n'est pas des plus

« intuitives », et nombreux sont les cas où même un annotateur aguerri peut être amené à hésiter : problèmes d'homonymies (ces/ses, été/était...), accords de participes passés de verbes pronominaux, pluriels incertains (pas de problème/problèmes), etc. Même si l'hésitation peut parfois ne durer que quelques secondes, il suffit de ramener ce laps de temps au nombre d'hésitations que le transcripteur peut avoir dans un corpus de cent heures...

Qui plus est, et cette fois sans même avoir à parler de spécificités du français, un autre versant de cet aspect orthographique est tout aussi important, sinon plus : les noms propres. Bien que plus fréquents dans un cadre préparé, comme l'une de nos expériences l'a montré (voir en 2.2), ils n'en demeurent pas moins un élément prégnant de certains débats ou interviews, pour peu que le sujet y soit propice. Une joute verbale politique sera rapidement sujette à l'évocation de présidents, ministres, députés... et parfois dans des pays peu familiers à l'annotateur francophone, ce qui peut poser parfois de vrais problèmes, ne serait-ce que pour identifier clairement la personne dont il est question. S'ajoutent à cela les noms de journalistes ou de techniciens mentionnés à la fin d'une émission, et qui demeurent bien souvent assez confidentiels, donc délicats à identifier.

La question de la ponctuation se pose elle aussi rapidement. Les transcriptions doivent-elles être ponctuées et, si oui, selon quels principes ? Il est évident que le moindre point ou la moindre virgule ne peut être qu'un artifice, puisque le texte transcrit lui-même n'est déjà qu'une *représentation* du signal audio correspondant. Qui plus est, quels critères détermineraient le choix d'un point plutôt que d'une virgule ou d'un point-virgule ? L'intonation, la durée d'un silence, la syntaxe du propos ? Par ailleurs, il faut également prendre en compte les valeurs interrogatives (?) ou exclamatives (!), les propos rapportés (« »), qui relèvent d'une interprétation et non d'une simple observation... L'hypothèse d'une absence totale de ponctuation est envisageable, mais la simplification radicale de cette option contraste avec la délicate lisibilité qu'elle produit. Une transcription non ponctuée s'avére rapidement pénible à lire, et ne facilite pas la compréhension globale des propos. Qui plus est, une telle décision rend caduque de potentielles analyses *a posteriori* sur des sujets comme l'interrogation par exemple.

Finalement, c'est le choix d'une ponctuation complète qui a été adopté pour le corpus EPAC, avec toutes les réserves que cela entraîne, et notamment une relative perte de temps. Toutefois, les transcriptions sont ainsi bien plus structurées, et surtout beaucoup plus « lisibles ».

1.3.4 La parole superposée

S'il est une difficulté à laquelle nous avons été confrontés tout au long de ce travail, c'est bel et bien la parole superposée. Sans même évoquer la tâche de transcription elle-même ou le logiciel utilisé, la parole superposée en elle-même pose de toutes façons problème : il faut parvenir à en distinguer les différents intervenants, ainsi que leurs propos respectifs. Il s'agit là d'un véritable travail d'écoute, souvent complexe, relativement long, et dont l'exactitude est sujette à caution. Entre les mots que l'on croit distinguer derrière un voile opaque de parole et ceux qui ont réellement été prononcés, il y a parfois un fossé important.

Recentrée dans le contexte d'une transcription, cette activité se voit complétée par le problème de l'alignement de la parole avec le signal audio. TRANSCRIBER permet cette synchronisation de façon naturelle. Toutefois, dans le cadre de la parole superposée, il n'autorise pas un alignement spécifique à chaque locuteur. Il faut en somme repérer *exactement* le mot du locuteur A sur lequel commence l'énoncé du locuteur B, ainsi que le mot sur lequel il se termine. Il n'est nullement possible de proposer deux segmentations distinctes, ce qui contraint le transcripteur à une écoute particulièrement attentive et coûteuse.

Qui plus est, il arrive qu'au milieu de deux segments superposés se trouve un bref segment où un seul locuteur intervient. Dans pareil cas, il convient de segmenter de façon très précise pour isoler la brève séquence mono-locuteur au milieu des segments superposés. La figure 1 donnera une idée du temps qu'un tel découpage peut réclamer :

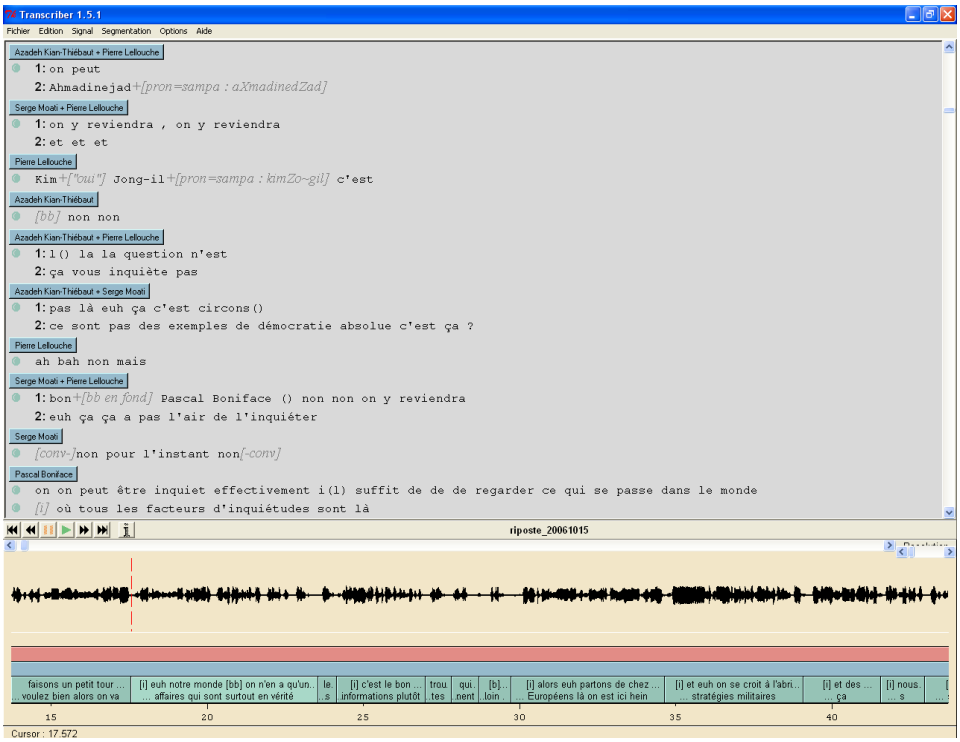


Figure 1 : Segmentation d'un dialogue sous TRANSCRIBER

Par ailleurs, TRANSCRIBER ne permettant la gestion de la parole superposée que par des moyens détournés, il n'est pas possible (à moins d'utiliser une version non officielle, modifiée par Mathieu Quignard[12], sur laquelle nous reviendrons) de gérer plus de deux flux de parole superposée. Ainsi, si trois personnes parlent simultanément, il sera impossible de reproduire leurs propos de façon alignée. Il conviendra alors d'utiliser d'autres artifice (balise, bruit de fond) pour espérer parvenir à faire apparaître la transcription orthographique de tous les intervenants :

12 http://www.loria.fr/~quignard/Transcriber/

Figure 2 : Transcription d'un dialogue à 3 participants sous TRANSCRIBER

Lorsque ces intervenants sont particulièrement nombreux et/ou tous du même sexe, un problème supplémentaire peut se poser : celui de l'identification des locuteurs. Il est parfois compliqué, voire même impossible, de déterminer avec certitude quel locuteur prend la parole, surtout lorsque sa voix est presque entièrement recouverte par celle de deux ou trois autres locuteurs. Même après de nombreuses écoutes, le doute subsiste et, bien que cela puisse paraître surprenant, il est quelquefois plus aisé d'identifier les propos d'un locuteur que le locuteur lui-même. Notons que cet aspect est surtout prégnant lors de la transcription d'émissions radiophoniques. Dans le cas de données télévisuelles, le support de l'image est un atout indéniable. En effet, le jeu des caméras permet très régulièrement d'identifier les locuteurs, même lorsqu'ils parlent en même temps. À l'inverse, si les animateurs ou journalistes de radio font régulièrement l'effort de nommer leurs invités lorsque ceux-ci s'expriment, ce souci de clarté devient rapidement impossible à mettre en place lorsque plusieurs d'entre eux n'ont de cesse de se couper la parole. Ces situations arrivent de façon relativement fréquente dans des émissions aux multiples invités, telles que *Sous les étoiles exactement* par exemple.

1.3.5 De la transcription manuelle à la transcription assistée

Lors de la première phase de notre travail, les transcriptions ont été réalisées de façon manuelle, c'est-à-dire sans le support d'une sortie automatique générée par système de reconnaissance de la parole. TRANSCRIBER était simplement couplé à un fichier son, et tout le travail de transcription, segmentation, annotation... restait à faire. Durant cette période, nous avons apporté un soin tout particulier à l'annotation des spécificités de la parole spontanée. Chaque prononciation particulière due à une élision et/ou une assimilation a été mentionnée en alphabet SAMPA à l'intérieur d'une balise. Un corpus d'une dizaine d'heures a ainsi été créé, et demeure disponible pour des requêtes spécifiques. Dans le cadre d'EPAC, une telle granularité d'annotation s'avérant trop coûteuse en temps, il a été décidé de ne pas étoffer cet aspect du corpus par la suite. Pour autant, nous nous sommes efforcé de conserver une granularité d'annotations aussi élevée que possible, de façon à pouvoir bénéficier d'informations paratextuelles pertinentes et variées.

Ainsi, tous les types de bruit ont été annotés : bruits de fond, inspirations / expirations / respirations, bruits buccaux, rires, applaudissements... De même, les plages musicales (chansons, jingles) ont également été étiquetées de façon relativement complète avec, par exemple, le titre et l'interprète dans le cas d'une chanson. Du point de vue morphosyntaxique, des critères de lisibilité nous ont amené à adopter une attitude centrée sur la normativité orhtographique : les « erreurs » grammaticales perceptibles à l'oral (« fautes » d'accord ou de genre), les liaisons « erronées » (« va-t-être ») ont aussi été mentionnées. Nous avons également indiqué toute prononciation ambiguë, notamment au sujet des noms propres ou du vocabulaire étranger, en caractères SAMPA.

Dans un second temps, la grande majorité des transcriptions ont bénéficié des sorties automatiques provenant de LIUM RT. De « transcripteur », nous sommes ainsi devenus « correcteur », mutation qui concernera de plus en plus ceux qui ont à produire ces représentations de l'oral que l'on nomme « transcription » (notamment les acteurs de projets comme VARILING, présenté en amont dans ce chapitre). La principale conséquence d'un tel changement, en l'occurrence le gain de temps, sera détaillée en section 2.2.

Cette démarche assistée comporte en outre un facteur humain non négligeable : au lieu de se trouver comme Sisyphe face à un travail solitaire en perpétuel renouvellement, le transcripteur / correcteur se trouve impliqué dans un travail collectif, impliquant des compétences et des finalités diverses. Quand bien même le gain de temps par rapport à une

transcription manuelle n'est que minime, il s'agit d'une autre forme de travail, différente de celle des transcripteurs qui ont constitué les principales bases de corpus parlé actuelles [Gougenheim *et al.*, 1964 ; Damourette et Pichon, 1911-1940]. D'une part, il s'agit d'un travail moins rébarbatif. D'autre part, ce travail réclame des compétences pluridisciplinaires, aussi bien en RAP qu'en linguistique de l'oral.

A la différence du projet VARILING, nous avons travaillé dans un tel contexte, sans faire appel à un autre laboratoire pour nous fournir les transcriptions. Cela nous a notamment amené à assimiler sur le vif le fonctionnement de systèmes qui utilisent une approche stochastique. Ainsi s'est créé au LIUM, puis au LIA, une collaboration entre linguistes et informaticiens d'une nature complètement différente de celle qui existait avant 1985, c'est-à-dire à une époque où les systèmes de RAP tentaient de fonctionner à partir de règles produites grâce à l'expertise des linguistes. Frederick Jelinek, alors responsable d'un groupe de recherche sur la parole chez IBM, disait il y a vingt ans qu'à chaque fois qu'il mettait un linguiste à la porte, les performances de son SRAP augmentaient[13] [Jelinek, 1988 et 2004]. Ce serait aujourd'hui un propos sans fondement dans la mesure où notre travail de transcripteur a permis de réduire le taux d'erreur mot de LIUM RT d'environ 8% en absolu (17,2% contre 25%) en ce qui concerne la parole spontanée [Estève *et al.*, 2010]. Cet écart illustre bien la tendance qui s'est opérée durant ces vingt-cinq dernières années en reconnaissance de la parole : si la linguistique peut apporter sa pierre à l'édifice, c'est désormais plus avec des masses de corpus que des études grammaticales (voir chapitre 2.2.3). Ce qui sous-entend qu'il existe aujourd'hui un vrai besoin de transcripteurs ou de correcteurs qualifiés car, malgré les progrès constants des SRAP, la parole spontanée pose toujours quelques problèmes que même un humain peine parfois à résoudre. Nous en ferons par ailleurs état dans notre prochain chapitre.

Transcrire de façon assistée offre donc un confort de travail certain, mais présente également quelques inconvénients. Tout d'abord, la transcription automatique peut parfois influencer le transcripteur, influence parfois positive (identification de mots que l'oreille humaine peine à identifier, comme certains noms propres, connus du modèle de langage, mais pas forcément d'un annotateur humain), et parfois négative (face à un segment audio dont le transcripteur ne perçoit pas tout de suite le sens, il sera tenté de s'appuyer sur le résultat fourni par SRAP). Il est difficile, dans l'action, de savoir quand on est pleinement objectif et quand

13 La citation exacte est la suivante : « Whenever I fire a linguist our system performance improves »

on l'est moins. Une écoute peut aisément être « biaisée » par la représentation que le transcripteur a face à lui.

Ce problème est surtout patent dans certaines situations précises. Face à la répétition d'un même mot, la préposition « à » par exemple, il est particulièrement difficile de quantifier le nombre d'occurrences lors d'un bégaiement ou d'une hésitation. En cas de doute, le transcripteur sera tenté de se fier au nombre d'occurrences détectées par le système de reconnaissance automatique plutôt que de prendre un temps précieux en procédant à un décompte manuel, au demeurant incertain.

La transcription assistée peut également poser quelques soucis au niveau de la segmentation en tours de parole. Outre le texte, la sortie automatique générée par LIUM RT fournit également un découpage en locuteurs et en tours de parole. Si le premier ne cause aucune perte ou gain de temps particuliers, dans la mesure où il n'identifie pas nommément les locuteurs, il n'en va pas de même pour le second. Une segmentation automatique en tour de parole n'est efficace que si elle est précise ; dans le cas contraire, il faut réajuster les segments manuellement, ce qui est relativement long et surtout contraignant.

Il faut enfin mentionner les performances parfois médiocres du SRAP. Bien sûr, il s'agit de cas isolés, mais lorsque le transcripteur se retrouve face à l'un d'eux, il est légitime d'être désarçonné : de longues séquences de mots sont parfois aux antipodes du signal audio correspondant, et il est alors délicat de retrouver le moment où la situation se rétablit. Cela est particulièrement visible lorsque deux locuteurs ou plus parlent en même temps, lorsque la qualité du signal est mauvaise (canal téléphonique, conversation bruitée...) ou encore lorsqu'un locuteur ne s'exprime pas dans sa langue maternelle. En de pareilles circonstances, effacer puis retranscrire des segments complets s'avère plus long qu'un processus entièrement manuel.

1.3.6 Facteurs humains

Le corpus attendu à l'issue des deux premières années de travail devait être constitué d'environ 100 heures de transcriptions. Prise en tant que telle, la valeur de ce chiffre demeure relativement floue pour un transcripteur néophyte : il est en effet impossible de mesurer la quantité de travail que cela peut signifier. Quelques études liminaires ont avancé le chiffre de 50 heures de traitement (manuel) pour une heure de flux audio. Envisagé sous cet angle, le corpus EPAC aurait donc signifié environ 5000 heures de travail. Cependant, ce chiffre doit

être relativisé pour plusieurs raisons.

Tout d'abord, nous ne sommes resté que peu de temps sur le principe d'une transcription manuelle. D'une part, cela demandait beaucoup de temps, et d'autre part le projet EPAC prévoyait le recours à la transcription semi-automatique, notamment afin d'établir des comparaisons chiffrées entre les deux méthodes. Ainsi, comme le montrera la seconde partie du chapitre suivant, ce processus semi-automatisé a permis d'obtenir un gain de temps significatif, bien que parfois inégal.

Ensuite, un corpus d'une telle taille, et qui plus est transcrit par une seule personne, amène nécessairement à se poser des questions d'optimisation. Le logiciel TRANSCRIBER, utilisé pour réaliser l'ensemble du travail, a donc été progressivement exploité au mieux de ses possibilités, notamment avec la création de raccourcis spécifiques. Ces derniers ont ainsi permis d'annoter rapidement un certain nombre de phénomènes extra-linguistiques en vue d'études ultérieures.

Les nombreuses heures d'écoute ont également permis de se familiariser avec les particularités de la parole spontanée (parole superposée, répétitions...) et de perdre de moins en moins de temps en les traitant.

Cela étant, se lancer dans une transcription à grande échelle demande une organisation et une attention particulières. La moindre difficulté peut faire perdre de longues minutes, et il convient alors de se décider rapidement : vaut-il mieux multiplier les écoutes (et donc le temps de travail) dans l'espoir de percevoir clairement ce qui a été dit, ou bien est-il préférable de considérer le passage comme étant inaudible, et donc de ne pas le transcrire ? Certes la seconde solution tronque une partie des données, mais elle permet aussi d'avancer en ne perdant que peu de temps. La transcription étant une activité assez astreignante, les quelques secondes de paroles laissées de côté dans un premier temps permettent souvent d'être efficace plus longtemps par la suite.

1.4 Avenir du corpus EPAC

La constitution d'un corpus comme EPAC ne doit pas être une fin en soi. Une telle somme de données récoltées (environ 100 heures de transcription) n'a de sens que si elle peut être exploitée dans de nombreux domaines d'application [Cappeau, 2007]. Pour ce faire, il convient tout d'abord de les mettre à la disposition du plus grand nombre, pour éviter le syndrome des « corpus-placards » trop souvent rencontrés dans la communauté Parole francophone. Nous évoquerons donc dans un premier temps cet aspect contributif, avant de nous intéresser aux applications que permet un corpus comme EPAC. Nous en profiterons pour présenter deux études que nous avons nous-même réalisées : la morphologie orale du verbe *être*, ainsi qu'un classement des mots les plus utilisés dans une situation de parole spontanée, quarante-cinq ans après le travail fondateur de Gougenheim.

1.4.1 Enrichissement des données pour la communauté Parole

A ce titre, les transcriptions EPAC seront distribuées gratuitement aux laboratoires ayant précédemment acquis le corpus ESTER 1. Concernant les fichiers audio correspondants, ils seront fournis moyennant la facturation d'un disque dur sur lequel ils seront stockés.

Cette démarche permettra à la communauté Parole d'avoir à sa disposition une solide base de données, qui ne soit pas limitée à un nombre d'utilisateurs confidentiels. Car, nous le verrons, beaucoup de corpus de langue française s'apparentent à des « corpus fantômes », dont l'accès est verrouillé à double-tour. C'est pourquoi les membres du projet EPAC tiennent à ce que leur travail puisse profiter au plus grand nombre, et éventuellement susciter d'autres projets similaires.

En France à ce jour, il n'y a en effet que quelques initiatives isolées (ASILA notamment) qui permettent de profiter librement du travail scientifique d'autres chercheurs. Alors que certains pays comme l'Écosse mettent à disposition de tout à chacun des corpus impressionnants (nous y reviendrons longuement dans le chapitre suivant), il semble que cette notion collaborative tarde à se répandre dans notre pays. Pour autant, il existe quelques associations comme ELRA, qui ont pour vocation de gérer la distribution des ressources dans le domaine de la parole. De nombreux laboratoires ou projets (dont EPAC) s'y sont rattachés, ce qui permet d'avoir un accès – payant – à une précieuse banque de données multimédia, et

qui ne se limite pas à la langue française.

Par ailleurs, d'autres projets comme VARILING, que nous avons présenté en amont, s'inscrivent eux aussi dans cette politique de la (libre) distribution des données. Il est à souhaiter que de telles initiatives non seulement soient suivies, mais également qu'elles incitent d'autres détenteurs de corpus oraux (parfois de taille considérable) à mettre leurs données à disposition de l'ensemble des chercheurs.

1.4.2 Études lexicales, sémantiques et autres

La granularité de l'annotation du corpus EPAC offre la possibilité, à qui le souhaiterait, de mener de nombreuses expériences, notamment sur le plan sémantique ou lexical. Par le passé, de telles études ont d'ailleurs déjà été menées sur des corpus oraux, notamment concernant la place du pronom « on » ou l'importance des adverbes [Bilger et Cappeau, 2004].

Pour notre part, nous en avons déjà réalisé quelques-unes au cours de nos recherches, que nous présenterons ci-dessous. Au préalable, nous apporterons quelques précisions sur le corpus EPAC et les annotations qu'il contient, afin que les chercheurs désireux de l'exploiter sachent de quelles métadonnées ils bénéficieront.

1.4.2.1 Quelques précisions sur l'annotation du corpus EPAC

Tout d'abord, au sein même du corpus EPAC, un sous-ensemble représentant une dizaine d'heures de données a été annoté d'une façon particulière. Chaque prononciation spécifique due à une élision ou à une assimilation y a été mise en exergue de la façon suivante : la graphie standard a été préservée dans le corps du texte, tandis que la forme « oralisée » était indiquée à l'intérieur d'une balise, en alphabet SAMPA. Ce travail, qui a demandé un important travail d'écoute, permet par exemple d'observer dans quel contexte le schwa est élidé, avec quels locuteurs ou quel type de parole, etc. La recherche au sein du corpus, pour peu que l'utilisateur ait recours à TRANSCRIBER pour le visualiser, est facilitée par l'usage des balises et d'une nomenclature précise. Chaque balise indiquant une prononciation spécifique commence par l'expression *pron=sampa*. Ainsi, une simple recherche lexicale de ce terme permet d'accéder instantanément aux données voulues.

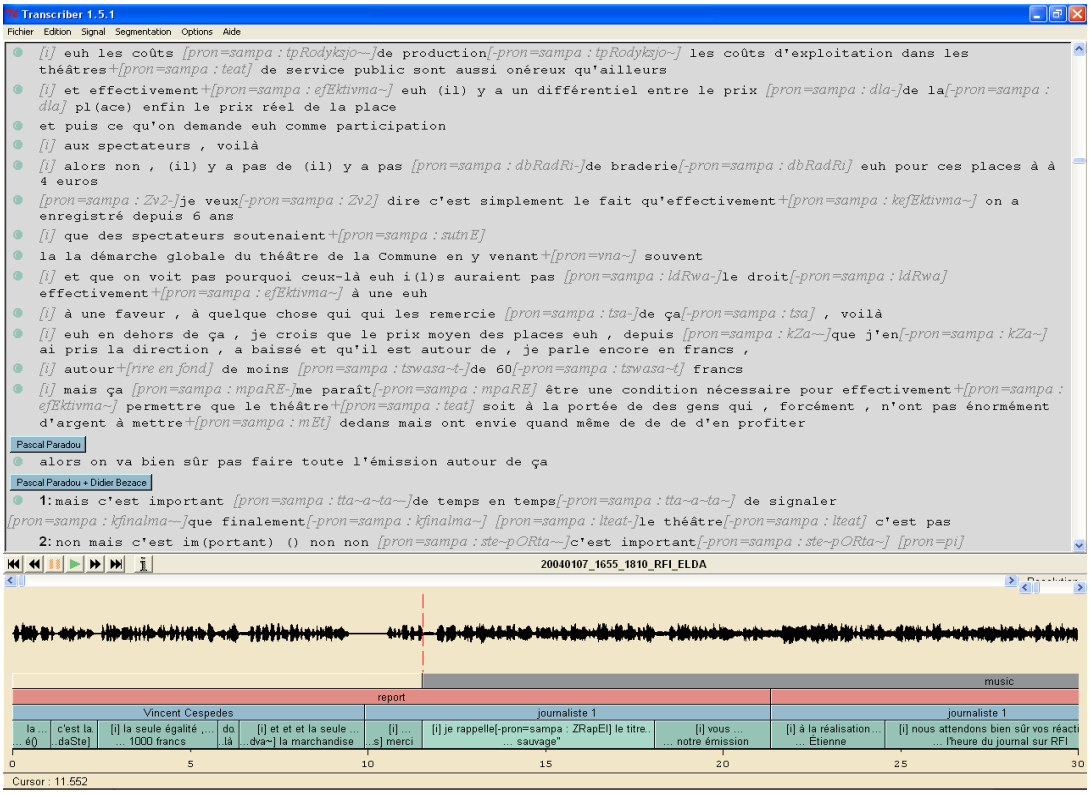


Figure 3 : Exemple d'annotation des élisions et des assimilations

Ensuite, et cette fois sur l'ensemble du corpus EPAC, les balises et l'alphabet SAMPA ont également été utilisés pour indiquer des prononciations erronées : erreurs de liaisons, bafouillages, articulation peu soignée... Un astérisque précédant le mot mal prononcé, conformément aux conventions ESTER, permet de distinguer ces cas précis de ceux que nous venons d'évoquer. Ce référencement autorise ainsi des visées intéressantes, notamment sur la qualité d'élocution des locuteurs, ou encore au sujet des mots ou expressions régulièrement déformés dans un cadre énonciatif. Selon le même principe, les erreurs grammaticales (faute d'accords, de genres...) ont elles aussi fait l'objet de balises idoines, ce qui pourrait s'avérer intéressant dans une optique didactique par exemple.

Toujours concernant la qualité d'élocution, le relevé systématique des bruits de bouche ou de gorge peut également s'avérer un élément prégnant. Étant entendu qu'ils traduisent fréquemment une hésitation ou la réorganisation d'un propos, ce sont des éléments pertinents si l'on souhaite mesurer l'aisance d'un locuteur face aux micros, ou sa faculté à produire un énoncé qui soit cohérent et compréhensible. Là encore, l'emploi permanent des balises normalisées [bb] et [bg] permet un accès rapide à ces données.

1.4.2.2 Études lexicales

De manière plus pragmatique, il est certain qu'un corpus de 100 heures de données, quelles qu'elles soient, offre de vastes perspectives lexicales. Dans le cadre plus restreint d'EPAC, nos données semblent particulièrement appropriées à des études telles que celle entreprise par Gougenheim [Gougenheim *et al.*, 1964] avec le Français Fondamental, à savoir relever quels étaient les mots de la langue française *vraiment* utilisés dans un discours naturel et non contraint. Quelques 45 ans après cet acte fondateur, le corpus EPAC nous a semblé une base tout à fait pertinente pour renouveler ce genre d'études.

Pour ce faire, nous avons utilisé le logiciel Word Frequency Counter[14], qui est un analyseur lexical robuste avec de nombreux éléments paramétrables (filtres de mots notamment). Nous lui avons demandé de traiter l'ensemble du corpus EPAC, soit une centaine d'heures de données représentant environ 1 333 000 mots. Il convient tout de suite de préciser que le corpus recueilli par Gougenheim et ses collègues est de bien moindre importance, puisqu'il ne compte que 312 135 mots. Par ailleurs, les conditions dans lesquelles ces corpus ont été réalisées sont également très différentes : les transcriptions du Français fondamental ont été réalisées à la volée, sans une quelconque possibilité d'enregistrement des entretiens réalisés. Ces transcriptions, saisies à la main, n'ont donc pas pu être retravaillées ou modifiées *a posteriori* à l'aide d'un support audio. Les conditions de réalisation du corpus EPAC ont été bien différentes : données audio numérisées permettant des écoutes infinies, utilisation d'un logiciel d'aide à la transcription... S'il est maladroit d'affirmer avec certitude que la qualité des transcriptions s'en ressent forcément, on pourra au moins avancer que les transcriptions EPAC devraient être plus abouties.

Enfin, le contenu même des deux corpus demeure relativement différent : le Français fondamental est composé d'entretiens avec 275 personnes, tous réalisés dans le cadre de vie quotidien des participants. EPAC est pour sa part constitué d'émissions radiophoniques, ce qui présuppose une certaine retenue dans les prises de parole des intervenants, qui savent (même s'ils peuvent parfois l'oublier) qu'ils sont enregistrés mais surtout écoutés par plusieurs milliers de personnes. Cela étant, en dépit de ces divergences, il nous a semblé intéressant de comparer les résultats de ces deux études, notamment pour tenter de dégager quelques grandes tendances du lexique oral. Le tableau ci-dessous présente les fréquences des dix mots les plus employés dans chacun des deux corpus :

14 http://www.hermetic.ch/wfc/

Français fondamental			**EPAC**	
Mot	*Fréquence*		*Mot*	*Fréquence*
de	10503	**1**	de	45192
je	7905	**2**	euh	37137
ce	7551	**3**	à	31199
il / ils	7515	**4**	la	26080
le	5851	**5**	et	22985
à	5629	**6**	le	22090
pas	5562	**7**	en	19568
la	5374	**8**	que	18746
et	5082	**9**	les	18316
on	4266	**10**	des	16377

Tableau 1 : Fréquence des dix mots les plus employés dans les corpus du Français fondamental et du projet EPAC

Avant toute analyse, nous nous devons d'apporter quelques précisions quant à ces résultats :

- les formes des verbes *être* et *avoir* ont été exclues de ce classement, car elles avaient été décomptées de façon trop différente d'un corpus à l'autre

- les chiffres concernant les mots comme *le*, *la* ou *pas* englobent leurs différentes acceptions et fonctions grammaticales.

Cela étant dit, on constatera tout d'abord que la préposition *de* occupe la première place des deux corpus, avec une marge assez importante sur les mots suivants. Ensuite, lorsque l'on voit l'importance qu'il revêt dans le corpus EPAC, il est étonnant de ne pas voir *euh* apparaître dans les premières places du corpus du Français fondamental, et encore plus de ne pas le trouver du tout à l'intérieur des relevés réalisés par Gougenheim et ses collègues. La seule explication plausible est que ce mot a volontairement été occulté des transcriptions et / ou des analyses, peut-être tout simplement parce qu'ils ne le considéraient pas comme tel. Pour notre part, nous expliquons dans le chapitre 5 les raisons pour lesquelles nous avons considéré ce *euh* d'hésitation comme un « morphème spécifique » de la langue orale.

Inversement, le pronom indéfini *on*, très fortement représenté dans les dialogues du

Français fondamental, n'apparaît qu'en 47ème position du côté du corpus EPAC. Si la connotation « oralisante » de *on* n'est pas aussi patente que celle de *euh (on* est régulièrement employé dans des journaux d'informations : « on apprend à l'instant que... », « on passe au sport... », « on est toujours sans nouvelles de... »), sa répartition dans les deux corpus a tout de même de quoi surprendre. De façon plus large, si l'on regarde attentivement le tableau, les pronoms personnels sont en fait beaucoup plus présents du côté du Français fondamental : outre *on*, on note également la présence de *je* et de *il* / *ils*, alors qu'aucun d'entre eux n'est présent dans la colonne du corpus EPAC.

Enfin, pour terminer ces quelques observations, nous nous arrêterons un instant sur le cas de *ben*. Gougenheim a été le premier à donner à ce « mot », dont nous parlerons largement en 5.2.3.2, un véritable statut au sein de la langue française parlée. Le fait qu'il arrive en 76ème position de son classement avait à l'époque constitué une petite révolution, et ouvert la voie à de nombreuses recherches sur le sujet par la suite[15]. Dans le corpus EPAC, ben ne se classe pour sa part qu'au 226ème rang, atteignant à peine la barre des 250 occurrences. Mais il est un autre chiffre tout aussi intéressant : la forme *bah*, jouant souvent un rôle similaire[16], apparaît elle à plus de mille reprises. Les caractéristiques acoustiques de ces deux mots étant très proches, on est en droit de se demander si, selon la perception auditive du transcripteur, les formes *ben* et *bah* ne sont pas parfois confondues. *bah* n'apparaît en effet à aucune reprise dans les statistiques du Français fondamental, à tel point que l'on peut penser qu'un consensus a été adopté pour transcrire *ben* ce que nous avons pour notre part orthographié *bah*, voire parfois *beh* ou *boh*.

Étant donné que le corpus EPAC contient également quelques heures de parole préparée, des analyses lexicales comparatives seraient également envisageables. Le fait qu'à l'intérieur d'EPAC se trouvent plusieurs enregistrements d'une même émission (*Le Téléphone Sonne*, *Sous les Étoiles Exactement*...) pourrait également être le point de départ de recherches lexicales contextuelles, s'intéressant notamment aux comportements de divers invités face à un même journaliste ou dans un même cadre énonciatif. Suivant un processus plus ou moins similaire, il serait aussi possible d'analyser les différentes prestations d'un même locuteur lors de diverses interventions (interview, débat...), afin de voir si son élocution varie en fonction

15 Voir notamment [Luzzati, 1982]

16 Voir 5.2.3.2

du contexte dans lequel il s'exprime.

Également, le fait d'avoir une importante somme de texte alignée avec un signal audio permet d'envisager des mesures précises sur le débit phonémique. Cela est d'autant plus intéressant que le corpus EPAC contient plusieurs « niveaux » de spontanéité, ainsi que quelques extraits de parole préparée : des comparaisons chiffrées pourraient donc être effectuées sur une masse de données conséquente et finalement assez hétérogène.

1.4.2.3 Études sémantiques

Une version étendue du corpus EPAC a été réalisée à l'automne 2009, comprenant des informations additionnelles sur les locuteurs et les émissions transcrites. Plus précisément, il s'est agi de spécifier le rôle, le métier et la fonction de chacun des intervenants du corpus, dès que cela était possible. Ainsi, un même locuteur pouvait avoir jusqu'à trois étiquettes (« invité / homme politique / ministre de l'intérieur », par exemple), suivant les informations disponibles. Dans le cas d'un débat contradictoire, une précision concernant l'opinion des intervenants était ajoutée, selon qu'ils étaient favorables ou non à la question posée. De même, chaque type d'émission était mentionné (débat, journal, entretien...), ainsi que son thème voire son sous-thème. Ce processus permet de disposer d'informations sémantiques riches et aisément exploitables car elles ont été directement intégrées *via* l'interface de TRANSCRIBER, et sont identifiables facilement dans les fichiers de sortie .trs.

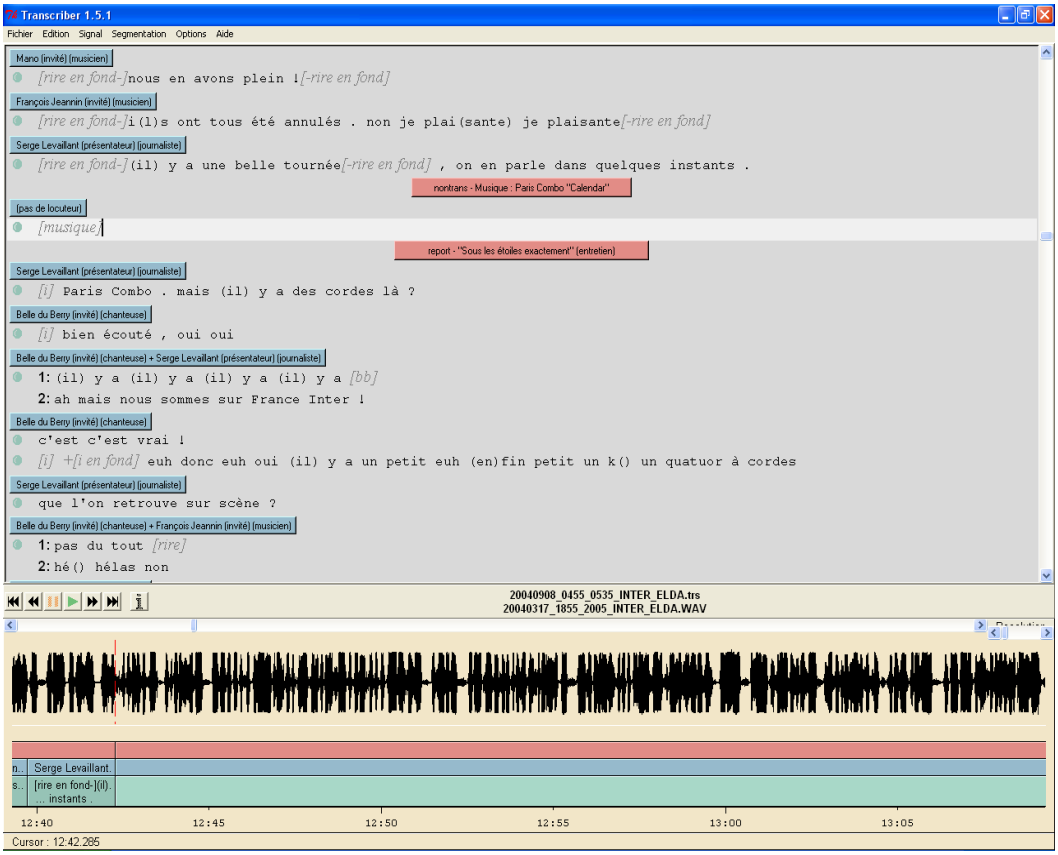


Figure 4 : Extrait du corpus EPAC annoté en genre d'émission et rôle du locuteur

Par ailleurs, couplées aux visées lexicales que nous venons d'évoquer, ces données supplémentaires permettraient d'opposer différents types de vocabulaires, i.e. celui des journalistes et des hommes politiques par exemple.

1.4.2.4 Études sur le traitement automatique de la langue

Enfin, le corpus EPAC ayant été intégralement ponctué, et bien que cette ponctuation demeure quelque peu artificielle, il est au moins un type de phrase pour lequel celle-ci peut être intéressante : les interrogatives. En effet, le point d'interrogation est sans doute le marqueur qui soit le moins sujet à caution ici, puisque certains éléments perceptibles à l'oral tels que l'intonation ou les structures interrogatives imposent quasiment d'eux-mêmes sa présence. Ainsi, notre corpus contenant beaucoup d'interrogatives (eu égard aux nombreux débats, interviews, entretiens...), il s'avère une base de données très intéressante pour quiconque s'intéresse d'une façon ou d'une autre à la modalité interrogative.

C'est pourquoi nous avons mené, en collaboration avec Benjamin Maza (doctorant au Laboratoire d'Informatique d'Avignon), une étude visant à développer un système capable de détecter automatiquement un énoncé interrogatif dans un flux audio « brut », c'est-à-dire sans segmentation préalable. Pour ce faire, nous avons commencé par isoler, à l'intérieur d'EPAC, des données susceptibles de contenir de nombreuses questions. C'est l'émission « Le Téléphone Sonne » qui a été retenue, car sa structure (des auditeurs qui appellent un standard pour s'exprimer sur un thème donné et qui, très souvent, posent une ou plusieurs questions aux invités présents) nous assurait une matière conséquente. Par ailleurs, quelques quarante numéros de cette émission ont été transcrits, ce qui représente un échantillon relativement large.

Ensuite, chaque question a été isolée par un annotateur : dans un premier temps, celui-ci a systématiquement recherché la présence de points d'interrogation pour établir son relevé, puis il a ensuite procédé à une minutieuse vérification manuelle pour s'assurer qu'aucune question n'avait été oubliée. Celles-ci ont alors été annotées selon la structure suivante :

- type de question : direct ou indirecte
- nature de l'interrogation : totale, partielle, alternative ou rhétorique
- marqueur interrogatif : intonation, inversion du sujet, terme complexe (« est-ce que »), adverbe interrogatif, pronom interrogatif, etc.

C'est le logiciel TRANSCRIBER qui a été retenu pour intégrer ces annotations au corpus. D'une part, cela évitait de convertir le corpus EPAC puisque celui-ci avait déjà été réalisé avec cet outil de transcription. D'autre part, grâce aux balises paramétrables, chaque catégorie de question a rapidement pu être intégrée au logiciel et assignée à un raccourci clavier spécifique. Voici comment apparaissent les annotations dans l'interface de TRANSCRIBER :

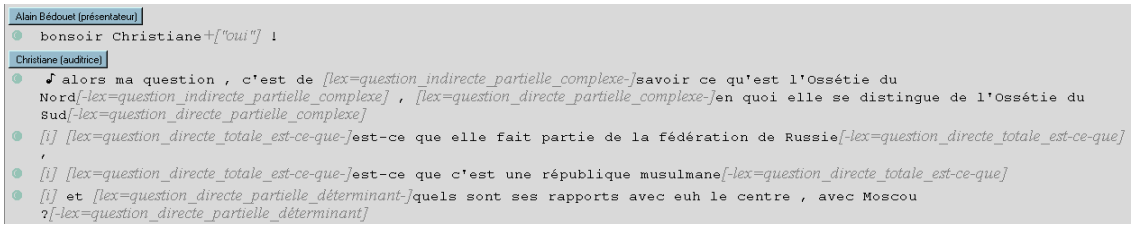


Figure 5 : Extrait du corpus EPAC annoté en type de questions

Quelques résultats liminaires ont déjà pu être obtenus sur le corpus du « Téléphone

Sonne » ; ils font état d'une précision de 75% et d'un rappel de 35%. Cela signifie que sur 100 occurrences relevées par le système, 75 sont effectivement des questions. Cependant, ce chiffre ne représente que 35% de toutes les questions bel et bien présentes dans le corpus. En d'autres termes, le système développé est (pour l'instant) plutôt efficace qualitativement, mais pas encore quantitativement.

Par ailleurs, à plus long terme, un deuxième objectif sera d'utiliser ces annotations pour proposer des méthodes de détection spécifiques selon le type de questions (basées sur l'intonation, la structure avec inversion du sujet, la présence d'adverbes interrogatifs, etc.).

1.5 Autres travaux de transcription ou d'annotation

Outre les travaux directement liés à EPAC, nous avons également abordé la transcription et l'annotation à travers différents projets et études. Nous présenterons donc, pour clore notre premier chapitre, un aperçu de ces quelques contributions.

1.5.1 Transcription d'émissions télévisées

Parallèlement aux transcriptions d'émissions radiophoniques réalisés pour le corpus EPAC, nous nous sommes également intéressés à la transcription d'émissions télévisées, afin de voir en quoi les deux tâches différaient. Nous avons ainsi constitué un corpus d'environ quatre heures, composé des émissions suivantes (toutes diffusées sur la chaîne France 5) : *C dans l'air*, *C'est notre affaire*, *Arrêt sur images*, *Ripostes* et *Madame, Monsieur, Bonsoir*.

Ce travail a nécessité une approche quelque peu différente de celle utilisée jusqu'alors avec les données radiophoniques. Si le support de l'image offre une possibilité supplémentaire quant à l'identification d'un locuteur par exemple, il est également à double-tranchant : le présentateur, sachant que les téléspectateurs entendent *et* voient le contenu de son émission, ne prend que peu souvent la peine de nommer ses intervenants lorsqu'ils prennent la parole spontanément (voir par ailleurs le chapitre 5.4.2.). Ces ellipses sont parfois problématiques : dans notre cas, et pour des raisons liées au format des fichiers, nous avons réalisé ces transcriptions avec TRANSCRIBER. Or, nous aurons l'occasion d'y revenir dans notre troisième chapitre, ce logiciel n'accepte en entrée que des fichiers audio. Il faut donc, pour bénéficier de l'aide de la vidéo, l'utiliser en parallèle avec un lecteur indépendant. Cela n'est pas poser quelques problèmes de synchronisation, qui au final sont synonymes de perte de temps.

Ce phénomène en entraîne un autre dans son sillage : le rythme de l'émission. À la radio, les locuteurs sont contraints d'apporter un minimum de soin à leurs propos, car c'est par le seul canal audio que les auditeurs y auront accès. À la télévision, les données sont tout autres : les mouvements de caméra offrent au téléspectateur une sorte de fil rouge où sont captés les gestes, les regards, les positions... La parole est certes un vecteur de communication essentiel, mais elle n'est plus seule. Sans aller jusqu'à dire qu'elle est facultative, il est parfois tout à fait envisageable de comprendre les tenants et les aboutissants d'une émission télévisée

sans bénéficier du son. En conséquence, les propos tenus à la télévision ont plus facilement tendance à se désarticuler, se déformer ou s'enchevêtrer. De telles fluctuations ne facilitent pas la tâche du transcripteur, qui doit alors consacrer à ces données un temps d'écoute bien plus important.

1.5.2 Annotation en frames sémantiques du corpus MEDIA[17]

Les compétences acquises lors de la phase de transcription et d'annotation du corpus EPAC nous ont également amené à travailler sur le corpus MEDIA [Bonneau-Maynard *et al.*, 2005]. MEDIA est un corpus de 1250 dialogues en français, qui a été enregistré selon le protocole du Magicien d'Oz (un locuteur croit s'adresser à un système automatisé, alors qu'il parle en fait à un être humain simulant le processus) [Kelley, 1984]. Une fois manuellement transcrit et annoté en concepts par la société ELDA, le LIA s'est occupé de l'annotation de MEDIA en frames sémantiques [Meurs *et al.*, 2008].

L'appellation « frames » désigne donc des éléments sémantiques pouvant être reliés entre eux *via* des sous-frames : par exemple, les frames *lodging* (hébergement) et *room* (chambre) peuvent être coordonnées à l'aide de la sous-frame *lodging_room*. De la frame *room* peuvent ensuite découler des sous-frames comme *nb_rooms*, *room_quality* ou *room_type*, qui viendront ensuite se lier directement aux unités lexicales concernées dans le dialogue.

Les dialogues du corpus MEDIA ont donc été annotés automatiquement selon une ontologie définie de frames et de sous-frames, et notre rôle a été de corriger et / ou de valider ces annotations. Nous avons pour cela utilisé le logiciel SALTO, qui permet d'avoir une vision dynamique et ajustable de la représentation en frames sémantiques. Voici un exemple de l'interface de SALTO, avec un dialogue issu du corpus MEDIA.

17 Ce travail a été réalisé dans le cadre du projet LUNA (contrat IST n°33549), inscrit dans le sixième programme-cadre de recherche (FP6) de l'Union Européenne. Site internet du projet LUNA : http://www.ist-luna.eu

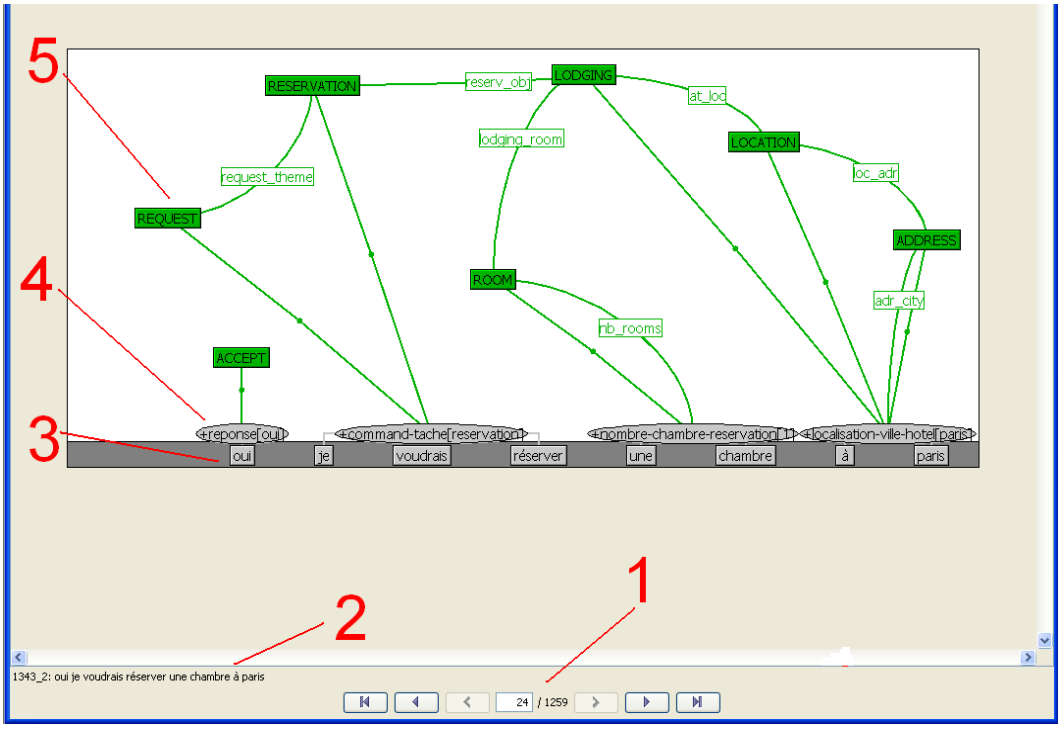


Figure 6 : Interface principale du logiciel SALTO

Il est intéressant de distinguer sur la figure ci-dessus :

1 : les commandes permettant de naviguer entre les différents dialogues

2 : le dialogue actuellement affiché, précédé d'un numéro d'identification

3 : le dialogue découpé en unités lexicales minimales

4 : l'annotation en concepts

5 : l'annotation en frames et sous-frames sémantiques

Chaque dialogue comme celui ci-dessus a donc été manuellement vérifié, pour s'assurer qu'aucune ramification logique ne manquait entre les frames. Des corpus de test et de développement d'environ 3000 et 1300 tours de parole (un dialogue contenant plusieurs tours successifs) ont ainsi été examinés puis validés, devenant de solides références sur lesquelles s'appuyer pour mener à bien les autres tâches du projet LUNA.

1.5.3 Segmentation en locuteurs du corpus vidéo LIA GERARD

Enfin, lors de travaux plus récents ayant fait l'objet d'une soumission à Interspeech 2010, nous avons segmenté en locuteurs le corpus LIA GERARD (multi**GE**n**R**e **A**udio/video

F**R**ench **D**atabase), composé de 130 vidéos issues d'internet. Il s'agissait de documentaires, de bandes-annonces, d'actualités, de publicités ou de dessins animés, dont la longueur variait de 1 à 10 minutes. La segmentation en locuteurs a été réalisée avec TRANSCRIBER. Celle-ci n'était pas nominative, mais distinguait les voix masculines et féminines sous les appellations M0, M1, M2... F0, F1, F2, etc. Les segments de non-parole était quant à eux attribués à un « locuteur » baptisé Z.

À terme, l'objectif de ces travaux est de parvenir à proposer un système capable d'établir automatiquement une segmentation en locuteurs sur des données vidéo, en proposant une distinction sexuée homme / femme.

1.6 Conclusion

Le projet EPAC, qui a été mené à son terme fin 2010, a été l'initiateur de cette thèse. Plus précisément, la transcription et l'annotation d'un corpus équivalent à plus d'un million de mots a orienté de façon prégnante nos recherches, et a bien souvent démenti nos intuitions ou certitudes initiales. Avant d'avoir été un moteur pour les autres sous-projets d'EPAC, ce corpus nous a permis d'appréhender de façon concrète la transcription de données. Il a aussi et surtout offert, une fois constitué, une base de travail très riche notamment grâce à la granularité fine des annotations.

Mais encore, il nous a fourni une alternative pertinente aux travaux de [Gougenheim *et al.*, 1964], et peut éventuellement prétendre à devenir un nouveau maître-étalon dans ce domaine. À ce titre, nous avons nous-même réalisé ou contribué à différentes expériences, en espérant que les facilités d'acquisition proposées pour le corpus EPAC en engendrent d'autres par la suite.

Enfin, les compétences acquises durant ce projet nous ont permis d'intégrer ensuite d'autres projets de transcription et/ou d'annotation, et ainsi de diversifier nos compétences en les confrontant à d'autres outils et données. Tous ces horizons présentaient cependant une nécessité commune : celle d'avoir une base de travail à disposition, en d'autres termes un corpus transcrit (ou à transcrire). Celui-ci, selon les besoins, peut être constitué d'une ou de plusieurs dizaines d'heures de données, transcrites avec plus ou moins de métadonnées, de façon manuelle ou semi-assistée...

C'est à tout cet ensemble de paramètres que nous allons désormais nous intéresser dans notre deuxième chapitre, consacré intégralement à cette tâche de transcription.

Chapitre 2 : La transcription

Ce chapitre aura pour objectif de définir un aspect fondamental de notre travail : la transcription de données orales.

La transcription pourrait se résumer ainsi : matérialiser sur un support (papier ou numérique principalement) la perception auditive que l'on a d'un flux audio. Le mot « perception » est ici lourd de sens puisque celle-ci peut varier d'un transcripteur à l'autre, et donc donner lieu à des représentations différentes d'une même production sonore. Dans le cadre du projet EPAC, la transcription concernait des émissions radiophoniques, mais il peut aussi s'agir dans un cadre plus large de réunions, d'entretiens, d'activités scolaires, etc.

Une transcription peut être effectuée de deux façons :

- à la volée, c'est-à-dire en recueillant puis en transcrivant instantanément des propos que l'on a entendus. Cette méthode, arbitraire et peu précise, était utilisée par défaut à l'époque où les appareils d'enregistrement n'existaient pas encore (ou étaient encore trop peu exploitables). Des pionniers tels que Damourette et Pichon [Damourette et Pichon, 1911-1940] ou bien Frei [Frei, 1929] ont ainsi, nous le verrons, été à l'origine des premiers « corpus » de langue française, il y a bientôt un siècle de cela.

- *a posteriori*, ce qui est la méthode communément utilisée aujourd'hui. Cela sous-entend qu'une phase d'enregistrement précède celle de transcription. Grâce à l'avènement du matériel numérique, celle-ci est devenue de moins en moins fastidieuse, et les résultats obtenus sont souvent excellents en terme de qualité sonore. Par ailleurs, le numérique permet également à ces données d'être facilement conservées, sous la forme de fichiers audio facilement exploitables et n'occupant qu'un espace virtuel sur un disque dur ou même une clé USB. Le temps des bandes magnétiques et même des cassette audio, supports fragiles et encombrants, semble désormais bien loin... Ce confort de stockage et d'écoute offre la possibilité d'effectuer des transcriptions précises, à l'aide de logiciels qui permettent de coordonner un flux audio avec le texte saisi par l'utilisateur. Le chapitre 3 sera d'ailleurs intégralement consacré aux plus répandus de ces logiciels d'aide à la transcription.

Auparavant, nous envisagerons donc dans ce deuxième chapitre la tâche de transcription en elle-même, sous un double regard synchronique et diachronique. Nous évoquerons les premiers travaux fondateurs sur le sujet, pour en arriver aux méthodes actuelles. Ce panorama sera l'occasion de dresser un inventaire des principaux corpus de parole disponibles à l'heure actuelle, en France comme à l'étranger.

Nous nous intéresserons ensuite aux conventions de transcription, c'est-à-dire aux règles adoptées par toute personne ou tout groupe de recherche désirant transcrire des données audio. Nous verrons que ces conventions ne sont que rarement unifiées, ce qui posent de nombreux problèmes d'accessibilité et de partage.

Enfin, nous clôturerons ce chapitre en nous interrogeant sur la pertinence d'une transcription « assistée ». Une expérience réalisée dans le cadre d'un article présenté à LREC en 2008 sera notamment présentée : celle-ci avait pour but d'évaluer le gain de temps qu'il était possible d'obtenir en fournissant au transcripteur l'aide d'un système de reconnaissance automatique de la parole. Les résultats obtenus seront l'occasion de se pencher sur les principales erreurs que commettent les SRAP, et nous proposerons quelques pistes en vue de les résoudre.

2.1 Corpus disponibles et conventions existantes

2.1.1 Historique

Se lancer dans l'établissement d'un corpus de parole spontanée sous-entend des possibilités d'enregistrement et de traitement post-enregistrement importantes. Comment en effet envisager d'analyser un objet dont on ne saurait avoir une quelconque trace ? A cet égard, l'historique des corpus qui nous intéressent est étroitement lié aux avancées technologiques du vingtième 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.

Pourtant, avant même que ne naisse cette invention, Damourette et Pichon [Damourette et Pichon, 1911-1940], en s'appuyant sur des conversations recueillies auprès d'un médecin, d'une institutrice… avaient dessiné les premiers contours morpho-syntaxiques 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. Toutefois, l'objectif de Frei était moins l'analyse de ce style oral que la hiérarchisation des « fautes » qui pouvaient s'y trouver.

Mais ce sont véritablement Gougenheim, Sauvageot, Michée et Rivenc [Gougenheim *et al.*, 1964] qui, s'appuyant sur 275 enregistrements sonores, ont révélé sans que cela constitue leur objectif premier la véritable nature de l'oral spontané. Ils souhaitaient avant tout proposer un équivalent du *Basic English* [Ogden, 1932], et ainsi favoriser l'apprentissage du français langue étrangère. Ils ont effectué notamment une étude quantitative du nombre d'occurrences des formes rencontrées. Il apparut alors que des mots comme « on », « hein » ou « ben » étaient parmi les plus utilisées de la langue française parlée, ce qu'aucune grammaire de l'époque ne prenait en compte. Aujourd'hui, certaines d'entre elles continuent de ne pas référencer ces termes. Il faut dire que les intégrer à partir des catégories en usage n'est pas

chose facile. S'agit-il d'interjections, d'onomatopées, d'adverbes, de conjonctions, etc. ? Leur mention passe donc par l'invention de nouveaux concepts qui prennent du temps à s'installer. Cela va des appuis du discours aux conjonctions, en passant par les connecteurs [Roulet *et al.*, 1985].

À l'époque où Damourette et Pichon entreprirent leurs recherches, les 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 ne permettait pas de faire des études précises : il était par exemple impossible d'enrichir ou de modifier une transcription après-coup, puisque l'objet sonore n'avait pas été capturé. Par la suite, les ordinateurs ont changé la donne : le traitement de texte a tout d'abord permis de numériser les transcriptions, et d'en assurer une certaine pérennité. Un grand pas a ensuite été franchi lorsqu'il est devenu possible de numériser également le signal sonore, évitant ainsi l'inévitable dégradation dont souffraient les bandes magnétiques, microfilms, cassettes, etc. Parallèlement à ces avancées, l'arrivée de logiciels permettant d'aligner le signal audio avec le texte de la transcription a été une petite révolution : il est désormais devenu possible, en quelques secondes, d'écouter n'importe quelle partie d'un enregistrement, et de voir apparaître à l'écran la transcription qui en a 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 présenterons en 2.2. quelques-uns des plus utilisés à l'heure actuelle.

2.1.2 Inventaire des corpus existants / disponibles

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 400 000 mots - sont la base de données CLAPI (Lyon, équipe ICAR), le corpus CRFP (Aix-en-Provence, équipe DELIC), le projet ASILA [ASILA, 2003], ainsi que deux initiatives belges, VALIBEL et ELICOP. À ces derniers viennent s'ajouter ceux issus de la communauté « parole » (ESTER par exemple), ainsi que divers autres, tels que le corpus PFC (Phonologie du Français Contemporain) [Durand *et al.*, 2002 ; Durand *et al.*, 2005] ou les projets RHAPSODIE et VARILING. 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 morpho-

syntaxe, VARILING s'inscrit dans une perspective sociologique, RHAPSODIE traite de la prosodie, tandis que PFC considère les aspects phonétiques et phonologiques de la langue.

Outre 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 (Centre de Ressources pour la Description de l'Oral) parisien 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 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 » [Cappeau et Seijido, 2005], 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.

La disponibilité et le contenu de ces données sont d'ailleurs des sujets qui font débat. Tous les corpus existants sont loin d'être de même nature et de proposer une même accessibilité [Baude, 2008]. Les plus anciens ne comportent que du texte sans numérisation du signal. Les plus récents comportent plusieurs sources vidéo. Certains restent fermés à toute personne n'ayant pas participé à leur élaboration, d'autres nécessitent une autorisation d'accès, d'autres encore sont commercialisés... 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. Il faut dire que les conditions juridiques d'exploitation [Baude, 2006] freinent l'accessibilité des données, dès lors que les locuteurs enregistrés ne sont pas « publics » (c'est le cas des médias) ou qu'ils n'ont pas dûment signé un document autorisant l'exploitation et la diffusion des enregistrements. L'enregistrement de parole spontanée est d'ailleurs en soi un problème, dans la mesure où la spontanéité en question présuppose des enregistrements à l'insu des locuteurs, dont il faudrait recueillir l'assentiment après coup...

Fort heureusement, des initiatives d'ampleur telles que celle diligentée par le Ministère de la Culture et de la Communication voient peu à peu le jour en France. Intitulée sobrement

« Corpus de la Parole », elle s'avère pourtant une mine de données impressionnante, qu'il est impossible de chiffrer avec précision. En quelques mots, le Corpus de la Parole se veut être le témoignage écrit et sonore de « Comment on parle en France aujourd'hui ». À ce titre, il regroupe donc des corpus de différents parlers et dialectes français, qui vont du picard au kabyle tout en passant par le créole ou le wallisien. La particularité majeure de ce projet est que tout son contenu est en libre accès, sans même une inscription ou une autorisation préalable : il est ainsi possible de consulter instantanément les transcriptions et/ou les enregistrements archivés, avec un alignement temporel lorsque les deux ressources sont disponibles conjointement. En contre-partie, aucun téléchargement n'est possible de façon directe, mais une telle initiative demeure quoi qu'il en soit unique en France (à notre connaissance), et ne trouve son équivalent que dans quelques pays étrangers comme l'Écosse, qui a à sa disposition le remarquable *Scottish corpus of Text and Speech*.

Dans le cadre de notre étude, c'est ce genre de travaux que nous avons souhaité mettre en valeur, parce qu'ils nous ont paru être les plus prometteurs. Référencer des corpus qu'il est impossible de consulter amène inévitablement à se poser la question de la « réalité » de leur existence. Nous avons ici préféré établir une liste, assurément non exhaustive mais aussi complète que possible, de données vraiment disponibles, parfois simplement au prix d'une connexion internet et d'un clic de souris. Certains corpus payants, notamment ceux distribués par ELRA, ont également intégrés ce classement. Si la somme pour les acquérir est parfois élevée, il n'en reste pas moins qu'ils peuvent « techniquement » être utilisés dans un cadre scientifique ou personnel par tout un chacun.

Cette étude doit beaucoup au travail de Paul Cappeau et Magali Seijido, que nous avons brièvement évoqué tout à l'heure. Il fut une base de départ très précieuse, même si notre critère de disponibilité a exclu une partie importante des corpus qu'ils avaient référencés. Lorsque cela était possible, nous avons pris soin de fournir un maximum d'informations possibles concernant la quantité de données, leur(s) format(s), la disponibilité des fichiers multimédia, le prix pour les acquérir, etc. Ces critères nous ont paru essentiels pour obtenir une description scientifique aussi précise que possible, même si dans certains cas tous n'ont pas pu être renseignés. L'objectif est ici de permettre au lecteur de savoir rapidement si un corpus est susceptible de l'intéresser ou non, et de lui proposer un accès direct à celui-ci.

Nombre d'informations, dont les liens fournis dans la colonne « Adresse », sont susceptibles de péremption, en quelques années voire en quelques mois. Nous espérons

cependant que les autres informations fournies permettront alors de retrouver facilement la trace du corpus recherché.

Nom du corpus	Responsable	Données	Format	Disponibilité	Adresse	Remarques
Air France	Sorbonne-Nouvelle – Paris III – Centre de Linguistique Française	- 73 dialogues	- 1 fichier .gz, mais impossible à décompresser. Le renommer en .txt permet plus ou moins d'accéder aux données. - Également disponible en XML	- Transcriptions libres	http://www.loria.fr/projets/asila/corpus_en_ligne.html	
Allaire	Suzanne Allaire	- Un article concernant l'étude de la subordination à travers un corpus de débats radiodiffusés		- Payant ; 15 € pour l'article (267 pages)	http://www.ilab.org/db/detail.php?lang=fr&booknr=355447913	
ALMURA	Jeanine-Elisa Médélice (Centre de Dialectologie de Grenoble)	- Enregistrements audio des parlers en Rhône-Alpes et régions limitrophes	- Mots ou expressions découpés en .mp3	- Données audio libres *via* un système de navigation par thèmes, expressions et régions	http://w3.u-grenoble3.fr/almura/atlas-theme.php	
Blanche-Benveniste	Blanche-Benveniste, Rouget, Sabio	- 36 extraits de français parlé	- Transcriptions intégrées au livre - CD-ROM pour audio	- Publié en livre	http://www.lavoisier.fr/notice/frIWO2AAOA2AWKL3.html	
Café	Daniel Luzzati	- 10 transcriptions d'enregistrements faits dans des cafés de 1979 à 1980	- Transcriptions en .doc ou .pdf - Pas d'audio	- Transcriptions libres	http://www.loria.fr/projets/asila/corpus_en_ligne.html	

CIO	*Information non disponible*	- 34 dialogues téléphoniques répartis en 3 phases	- Transcriptions en .txt - Pas d'audio	- Transcriptions libres	http://www.loria.fr/projets/asila/corpus_en_ligne.html#cio	
CLAPI	Équipe ICAR	- 45 corpus - 327 enregistrements (135h) - 514 transcriptions	- Transcriptions en .doc - Audio en .mp3	- Transcriptions et audio libres (14h téléchargeables, 35h accessibles)	http://clapi.univ-lyon2.fr/	- Navigation un peu confuse entre les données disponibles ou pas, accessibles, consultables...
C-ORAL-ROM	Equipe DELIC	- 1200000 mots - corpus multilingue de parole spontanée		Distribué par ELDA. Payant : 1500€ si membre d'ELDA, 3000€ sinon	http://www.elda.org/catalogue/fr/speech/S0172.html	
Corpus de la parole	Ministère de la culture et de la communication	- Portail regroupant une somme considérable de données, un peu à la manière de CLAPI	- Transcriptions affichés à l'écran - Audio aligné par segment	- Libre à la consultation - Pas de téléchargement	http://www.corpusdelaparole.culture.fr/spip.php?rubrique3	- Énorme initiative avec un nombre de données colossal - L'alignement des fichiers audio est très judicieux
CREDIF	Janine Leclerq	- 14 conversations d'enfants reproduites dans un ouvrage numérisé de 230 pages	- Transcriptions en .pdf	- Transcriptions libres	http://www.eric.ed.gov/ERICDocs/data/ericdocs2sql/content_storage_01/0000019b/80/37/16/63.pdf	

CRFP	DELIC	- 440 000 mots	*Information non disponible*	- Un exemple disponible	http://sites.univ-provence.fr/delic/corpus/index.html	- Possibilité d'utiliser un démonstrateur pour effectuer des recherches au sein du corpus
Ecole Massy [Antoine *et al.*, 2002]	VALORIA	- 31 dialogues - 5300 mots	- Transcriptions en .xml, ASCII, .doc, .odt, .pdf	- Transcriptions libres - Audio : 15 €	http://www.info.univ-tours.fr/~antoine/parole_publique/Massy/index.html	
EPAC	LIUM	- 100 heures de données radiophoniques	- Transcriptions au format .trs (Transcriber) - Audio en .wav ou .sph	- Transcriptions libres pour les participants à ESTER - Audio disponible moyennant la facturation d'un disque dur pour stocker les fichiers	http://epac.univ-lemans.fr/	
ESLO-1	CORAL	- 300 heures - 4,5 millions de mots	*Information non disponible*	Omelettes	http://www.univ-orleans.fr/eslo/spip.php?rubrique2	
ESLO-2	CORAL	- 400 heures - 6 millions de mots	*Information non disponible*		http://www.univ-orleans.fr/eslo/spip.php?rubrique3	- Corpus en cours de réalisation

GARS / Corpaix	GARS	Plus d'1 million de mots	*Information non disponible*	Compliquée ; quelques extraits publiés	http://www.lavoisier.fr/notice/frIWO2AAOA2AWKL3.html	
GOCAD	*Information non disponible*	- 32 dialogues (8 sujets x 4 tâches)	- Transcriptions au format .htm - Pas d'audio	Libre	http://www.loria.fr/projets/asila/corpus_en_ligne.html#gocad	- Format .htm peu pratique ; semble avoir des problèmes avec caractères accentués
Grenouille	DDL	- 6 heures - 96 enregistrements - 200 locuteurs	- Transcriptions au format CLAN (.cha) - Audio en .mp3	- Une partie est disponible sur la plate-forme CLAPI (cf plus haut)	http://clapi.univ-lyon2.fr/feuilleter.php?etude=N&locuteur=N&encours_corpus=20	
Archive Lacito	Lacito ; Michel Jacobson	- 200 documents en 45 langues, dont la plupart sont dites « rares »	- Transcriptions + traductions directement visibles à l'écran - Audio dans une application Java	- Pas de téléchargement	http://lacito.vjf.cnrs.fr/archivage/index_fr.htm	- Il est dommage qu'aucune donnée ne puisse être téléchargée - Absence d'un véritable alignement texte / son
Microfusées	Équipe COAST	- 4 enregistrements - 30 minutes	- Transcription en XML ou SGML - Pas d'audio	- Disponible la plate-forme ASILA	http://www.loria.fr/projets/asila/corpus_en_ligne.html#micro	

Les accents des Français	Pierre Léon, Fernand Carton, Mario Rossi	- 29 textes - Durée : environ une heure - thématique : accents des régions françaises	- À la base, livre + cassette audio - Récente adaptation numérique sous la forme d'un site internet	- Transcriptions visibles sur le site, avec conventions - Audio téléchargeable en .aif ou .rm	http://accentsdefrance.free.fr/	- Initiative très originale, parfaitement légale et pouvant servir d'exemple pour d'autres données au format « papier »
OTG [Antoine, 2002]	VALORIA	- 315 dialogues - 26000 mots - Durée : 2 heures	- Transcriptions en .pdf, .doc, .odt, .xml, ASCII	- Transcriptions gratuites - Audio : 15 €	http://www.info.univ-tours.fr/~antoine/parole_publique/OTG/index.html	
OZKAN	Nadine Ozkan, Jean Caelen	- 33 dialogues	- Transcriptions en .xml - Pas d'audio	- Transcriptions disponibles sur la plate-forme ASILA	http://www.loria.fr/projets/asila/corpus_en_ligne.html#ozkan	
PARIS-V	Denise François	*Information non disponible*		- Transcriptions disponibles dans un livre	http://www.peeters-leuven.be/boekoverz.asp?nr=5672	
PFC	Jacques Durand, Bernard Laks, Chantal Lyche (MoDyCo)	- 462 locuteurs francophones - 30 régions ou pays représentés - Parole lue ou spontanée	- Transcription à l'écran - Exportation possible en .trs ou .textgrid - Audio disponible	- Transcriptions libres, mais pas téléchargeables directement	http://www.projet-pfc.net/index.php?option=com_wrapper&view=wrapper&Itemid=152	- Corpus très riche et très diversifié
PIC	Laboratoire GREYC (Caen)	- Un dialogue	- Transcriptions en .xml ou .doc - Pas d'audio	- Transcriptions libres	http://www.loria.fr/projets/asila/corpus_en_ligne.html#pic	- Format .doc illisible (sous Openoffice)

Renault	Dominique Martini, Agnès Gryl et Xavier Briffault	- 27 dialogues	- Transcriptions en .xml ou .sun - Pas d'audio	- Transcriptions libres	http://www.loria.fr/projets/asila/corpus_en_ligne.html#renault	- Le format .sun est difficilement lisible avec un éditeur de texte - Les fichiers .doc complémentaires ne sont pas tous lisibles avec OpenOffice
RITEL	Limsi	- 5362 énoncés - 582 dialogues		- Distribué par ELRA	http://universal.elra.info/product_info.php?cPath=37_38&products_id=1422	
SNCF	*Information non disponible*	- 61 communications	- Transcriptions en .txt - Pas d'audio	- Transcriptions libres	http://www.loria.fr/projets/asila/corpus_en_ligne.html#sncf	
THESOC	Jean-Philippe Dalbera	*Information non disponible*	- Mots visibles à l'écran : API, graphie phonologique, lemme - Pas d'audio	- Transcriptions libres	http://thesaurus.unice.fr/	

Tableau 2 : Principaux corpus français "disponibles"

Nom du corpus	Responsable	Données	Format	Disponibilité	Adresse	Langue(s)
American National Corpus [Fillmore *et al.*, 1998]	Randi Rippen	- À terme, 100 millions de mots - 15 millions en libre distribution dont 3 millions de langue parlée	- Transcriptions à l'écran - Pas d'audio	- Consultable en ligne - Téléchargeable à hauteur de 15 millions de mots	http://www.americannationalcorpus.org/	- Anglais américain
Archiv für Gesprochenes Deutsch	Institut für Deutsche Sprache	*Information non disponible*	*Information non disponible*	- Apparemment, payant	http://agd.ids-mannheim.de/html/index.shtml	- Allemand
BigBrother corpus [Johannessen *et al.*, 2007]	Janne Bondi Johannessen	- 550 000 mots	*Information non disponible*	- Textes, audio et vidéos consultables	http://www.tekstlab.uio.no/nota/bigbrother/	- Norvégien
BNC	Université d'Oxford	- 100 millions de mots au total - 10 millions concernant la langue parlée	- Transcriptions en .xml	- Consultable en ligne gratuitement. - 75 £ pour la version DVD - Pas d'audio disponible	http://www.natcorp.ox.ac.uk/	- Anglais
Corpus del Español	Mark Davies	- 100 millions de mots	- Transcriptions affichées à l'écran avec hyperliens	- Consultable en ligne gratuitement	http://www.corpusdelespanol.org/	- Espagnol
Corpus de Referencia del Español Actual	Real Academia Española	*Information non disponible*	- Transcriptions affichées à l'écran avec hyperliens	- Consultable en ligne gratuitement	http://corpus.rae.es/creanet.html	- Espagnol

Corpus Diacronico del Español	Real Academia Española	*Information non disponible*	- Transcriptions affichées à l'écran avec hyperliens	- Consultable en ligne gratuitement	http://corpus.rae.es/cordenet.html	- Espagnol
Corpus of Contemporary American English	Mark Davies	400 millions de mots, a priori 60 millions sur langue parlée	- Transcriptions affichées à l'écran avec hyperliens	- Consultable en ligne gratuitement	http://www.americancorpus.org/	- Anglais américain
ELICOP [Mertens, 2002]	Université de Louvain	Environ 100 heures de transcription graphique disponible	Textes en ligne	Transcriptions consultables via moteur de recherche	http://bach.arts.kuleuven.be/pmertens/corpus/search/s.html	- Belge
EMILLE	Lancaster University	- 92 millions de mots de langue parlée - Données concernant 14 langues d'Asie du Sud		- Distribué par ELRA - Une partie du corpus semble accessible gratuitement	http://www.lancs.ac.uk/fass/projects/corpus/emille/	- Assamese, Bengali, Gujarati, Hindi, Kannada, Kashmiri, Malayalam, Marathi, Oriya, Punjabi, Sinhala, Tamil, Telegu et Urdu
FLLOC	Florence Myles, Rosamond Mitchell	- 4000 fichiers	- Transcriptions en .cha (format CLAN) ou .xml - Audio en .mp3 ou .wav	- Transcriptions libres - Fichiers audio libres	http://www.flloc.soton.ac.uk/	- Anglais - Français
Nationwide Speech Project Corpus [Clopper et Pisoni, 2006]	Cynthia Clopper	*Information non disponible*	*Information non disponible*	Contacter l'auteur pour avoir les données	http://www.ling.ohio-state.edu/~cclopper/nsp/index.html	- Anglais américain

Nota	Janne Bondi Johannessen	- 900 000 mots	- Données vidéo	- Les données sont *a priori* disponibles mais nécessitent un mot de passe	http://www.tekstlab.uio.no/nota/oslo/index.html	- Norvégien
Scandinavian Dialect Syntax	Janne Bondi Johannessen	- 350 000 mots - Objectif à terme : 1 million de mots	*Information non disponible*	- Les données sont *a priori* disponibles mais nécessitent un mot de passe	http://www.tekstlab.uio.no/nota/scandiasyn/index.html	- Scandinave
Scottish corpus of Text and Speech	Université de Glasgow	- 4 millions de mots au total, dont 800 000 en langue parlée	- Transcription en .txt - Audio et/ou vidéo en .mp4	- Tout est consultable en ligne (avec alignement texte/signal) ou téléchargeable	http://www.scottishcorpus.ac.uk/	- Écossais
Spoken Dutch Corpus [Oostdijk *et al.*, 2002 ; Boves et Oostdijk, 2003]	Dutch Language Union	- 10 millions de mots	- Transcriptions en .textgrid - Audio en .wav	- Distribué par le Dutch HLT Centre	http://lands.let.kun.nl/cgn/ehome.htm	- Néerlandais
Swedish Gøteborg Spoken Language Corpus	Université de Göteborg	- 182 heures	*Information non disponible*	- Attestation à remplir, puis accès aux transcriptions + audio	http://www.ling.gu.se/projekt/tal/index.cgi?PAGE=3	- Suédois
TAUS	Eskil Hanssen	- 212 000 mots	*Information non disponible*	*Information non disponible*	http://www.tekstlab.uio.no/nota/taus/index.html	- Norvégien
VALIBEL [Dister et Simon, 2008]	Université Catholique de Louvain	- 373 heures	*Information non disponible*	- Transcriptions consultables *via* un moteur de recherche	http://moca.fltr.ucl.ac.be/	- Belge

Tableau 3 : Principaux corpus étrangers "disponibles"

2.1.3 Les conventions de transcription

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 ?

Le premier aspect à aborder est celui de la représentation textuelle. Comment représenter par écrit des conversations orales ? La majorité des corpus francophones créés jusqu'à aujourd'hui ont adopté une orthographe normalisée, semblable à celle que l'on trouve dans les dictionnaires. Bien que ce choix suive une certaine logique de normalisation, il présente cependant quelques inconvénients. Ainsi, dans la langue parlée, il n'est pas rare que la prononciation de certains mots soit déformée, voire transformée. Les exemples les plus courants concernent l'élision de certaines lettres comme les « e » muets, dont nous aurons l'occasion de reparler longuement dans notre cinquième et dernier chapitre. Pour le dire simplement ici, un mot tel que *petit* sera souvent prononcé [pti] à l'oral : se pose alors la question de la représentation à l'écran.

Retranscrire le mot sous sa forme « petit » est la démarche la plus naturelle, et évite les processus artificiels comme l'emploi de l'apostrophe (« p'tit »). Cela étant, l'information concernant l'élision du « e » est perdue, et rien ne permet alors de distinguer à l'écrit les prononciations [p@ti] et [pti]. Une des solutions permettant de pallier ce problème est l'emploi de l'alphabet phonétique international (ou de la norme équivalente SAMPA). Néanmoins, transcrire une masse de données importante en API est une tâche démesurément longue, et qui pose entre autres de nombreux problèmes acoustiques [Durand et Tarrier, 2006]. Un compromis intéressant peut être le suivant : adopter une transcription orthographique « classique », complétée par l'utilisation de l'API pour les phénomènes que l'on souhaite mettre en valeur. C'est, nous l'avons vu dans notre premier chapitre, cette méthode qui a été adoptée lors de la réalisation du corpus EPAC. Bien qu'assez coûteuse en temps si la granularité de l'annotation est fine (réalisation ou non des schwas, liaisons, assimilations, etc.), elle offre des possibilités d'analyse *a posteriori* très riches. D'autres études proposent même de réaliser deux version d'un même corpus : l'une richement annotée, et l'autre beaucoup plus épurée [Benzitoun et Véronis, 2005].

Nous avons également déjà évoqué le problème de la ponctuation, qui nous amène ici à poser la question de sa légitimité. Si, dans le corpus EPAC, il a été décidé d'en ajouter une,

c'est plus par souci de lisibilité que par pertinence linguistique. En effet, matérialiser des données orales sous la forme d'une transcription est déjà en soi une démarche artificielle, puisque cela revient à transformer un signal sonore en représentation scripturale. En cela, l'ajout de la ponctuation est assez ambigu : est-il la matérialisation d'informations prosodiques (intonation montante ou descendante, élévation de la voix...), ou bien doit-il s'appuyer sur des indices morpho-syntaxiques ou lexicaux (emploi d'adverbe interrogatifs, exclamatifs ou d'interjections ; inversion du sujet...) ?

Les conventions d'annotation de la campagne ESTER sont par exemple assez ambiguës à ce sujet. Concernant l'usage du point, il y est ainsi recommandé de l'utiliser pour « marquer une pause relativement longue », de même que pour « délimiter une phrase, un syntagme ou un groupe de sens important ». Il y a donc ici confusion entre indices prosodiques (« pause relativement longue ») et éléments morpho-syntaxiques (« phrase », « syntagme »).

Quant au point d'interrogation il doit simplement, toujours selon ces mêmes conventions, « marquer la fin d'une phrase interrogative ». La question est alors de savoir si par « phrase interrogative », le transcripteur doit comprendre « phrase avec une structure interrogative explicite » ou bien « énoncé dont l'intonation montante indique une velléité interrogative ». Les deux critères sont certes parfois réunis, mais ce n'est pas toujours le cas (*tu viens demain ?*). La documentation d'ESTER conclut d'ailleurs le chapitre consacré à la ponctuation sur une phrase qui ne clarifie pas vraiment la situation : « Ces informations sont fortement liées à la prosodie et à la compréhension ».

Dans le cadre du corpus EPAC, nous nous sommes donc efforcés d'établir nos choix en fonction de la morpho-syntaxe et / ou de la prosodie, non sans être confrontés à des situations très délicates à traiter – les interrogations indirectes notamment.

D'autres projets adoptent une stratégie différente vis-à-vis de la ponctuation : VARILING, chargé de la transcription des Enquêtes Socio-Linguistiques d'Orléans (ESLO) [Abouda et Baude, 2005 ; Abouda et Baude, 2006], demande par exemple à ses transcripteurs de ne ponctuer que les phrases interrogatives. Cela sans doute parce le point d'interrogation est le signe de ponctuation le moins équivoque lorsqu'il s'agit de ponctuer de l'oral, en dépit des quelques cas épineux que nous avons mentionnés.

Pour sa part, le projet PFC (Phonologie du Français Contemporain) dit adopter une « ponctuation plus réduite que la norme française » [Durand et Tarrier, 2006], mais sans réellement expliciter en quoi cette réduction consiste. Si l'on regarde du plus près les

conventions de transcription de ce projet, il y est préconisé l'emploi de la virgule pour une « pause brève », du point pour « une pause plus longue (ou une fin d'énoncé marquée mélodiquement) » et du point d'interrogation pour une « question ». Il semblerait donc que l'usage du point d'exclamation ou du point-virgule ne soient pas pris en compte, et que la morpho-syntaxe n'entre que peu en compte dans les choix à effectuer.

Enfin, le groupe de recherche aixois du DELIC (récemment devenu TALEP) n'ajoute dans ses transcriptions aucun signe de ponctuation, quel qu'il soit. De façon plus large, il se refuse à l'adjonction de « toute indication prosodique ».

Ainsi, nous le voyons, la ponctuation est sujette à de nombreux choix et interprétations suivant les projets et groupes de recherche. Pour autant, cette diversité est très représentative de l'utilisation des conventions de transcription en général. En effet, outre la ponctuation, de nombreux autres phénomènes linguistiques ou extra-linguistiques sont régulièrement mis en valeur dans une transcription. Et, d'un corpus à l'autre, leurs représentations se recoupent parfois, mais divergent également dans de nombreux cas.

Conventions	**Annotation proposée**
ESTER	des idées ré() révolutionnaires
LDC	des idées [ré-] * révolutionnaires
DELIC, ICOR, VARILING	des idées ré- révolutionnaires
PFC / VALIBEL	des idées ré/ révolutionnaires

Tableau 4 : Exemple d'annotations proposées pour la troncation

Comme l'indique le tableau 4, et même si tous les exemples d'annotations divergentes ne sont pas aussi représentatifs, il y a bien ici un problème d'harmonisation : pour un seul élément, il existe au moins quatre représentations différentes. Il serait pourtant judicieux de s'orienter vers une annotation unifiée, ne serait-ce que pour permettre un échange et une lecture des données plus simples. D'autant que, si la représentation d'un élément diffère souvent d'un corpus à un autre, le choix même des éléments à annoter est lui aussi très variable. Selon les objectifs et les besoins de chaque projet, il ne sera pas forcément pertinent de mettre en valeur les mêmes phénomènes.

Mais avant même de se poser ces questions, il faut toutefois se soucier de plusieurs problèmes généraux concernant la représentation à l'écran d'un enregistrement sonore :

- les nombres : doivent-ils être retranscrits en chiffres ou en lettres ?

- les adresses internet : que transcrire exactement ? Les chiffres et lettres tels qu'ils sont prononcés (3 W point ...), ou bien l'adresse normalisée (www. ...) ?

- les sigles : comment distinguer ceux qui sont lus (FIFA) de ceux qui sont épelés (CIO), et comment traiter les ambiguïtés (ONU par exemple, qui peut être aussi bien lu qu'épelé) ?

- interjections : faut-il normaliser une orthographe pour les expressions comme « hmm », « ben », « euh »... ?

Les exemples pourraient ainsi se multiplier, mais les quatre que nous avons retenus montrent bien qu'avant de se lancer dans une annotation précise, il faut impérativement définir quelques règles de base pour que le corpus obtenu demeure aussi homogène que possible.

Une fois ces principes définis, il convient de s'intéresser à l'annotation spécifique en elle-même. Concernant les corpus de parole, certains phénomènes paraissent incontournables : les troncations, répétitions ou prononciations erronées se doivent par exemple d'être clairement mentionnées, soit par un signe diacritique quelconque (parenthèse, slash, crochets...) soit par une balise comme celles de TRANSCRIBER. Les conventions de transcription des principaux corpus de parole proposent ainsi des solutions pour chacun de ces événements.

La parole simultanée est elle aussi un élément prépondérant, notamment lorsque l'on a affaire à de l'oral spontané. Là encore, les solutions divergent : soit le transcripteur exploite la solution proposée par le logiciel de transcription qu'il utilise (tous envisagent un moyen spécifique de traiter la parole superposée), soit il se réfère à des signes diacritiques tels que les chevrons (< >), notamment préconisés par le projet PFC et le groupe DELIC.

Mais des annotations extra-linguistiques sont également envisageables : de nombreuses conventions de transcription proposent ainsi des solutions pour annoter les bruits de fond, bruits de micro, souffles, rires, etc. Des logiciels comme XTRANS ou TRANSCRIBER offrent même, nous le verrons, des propositions internes pour mettre en valeur ces éléments.

Enfin, il existe également des conventions de transcription qui considèrent quelques éléments prosodiques, comme l'intonation ou l'allongement vocalique. DELIC propose par exemple l'adjonction des deux points « : » pour le second phénomène cité (« mais vous savez, je pense **que:** à mon avis, c'est une erreur »). Quant à l'intonation, le groupe de recherche ICOR (Lyon) propose des annotations pour ce qu'il appelle les « montées et chutes

intonatives ». La montée se représente avec une barre oblique (/), et la chute une barre oblique inversée (\). En cas de phénomènes particulièrement accentués, ICOR propose de redoubler l'annotation (// ou \\).

2.2 Transcription manuelle ou assistée ?

Transcrire de la parole est, nous l'avons vu, une tâche assez complexe. Procéder de façon entièrement manuelle est très coûteux en temps, tandis que faire pleinement confiance aux SRAP n'est pas encore envisageable si l'on souhaite des transcriptions très précises, bien que les progrès effectués dans ce domaines soient constants [Geoffrois, 2004]. La transcription assistée apparaît donc comme un bon compromis pour obtenir de tels résultats dans un temps raisonnable.

C'est pourquoi nous présentons ci-après les résultats d'une étude [Bazillon, 2008b] qui a permis de quantifier le gain de temps qui peut être obtenu, pour la parole préparée et spontanée, avec l'aide d'un système de reconnaissance automatique de la parole. Celui qui a été utilisé est LIUM RT, un SRAP basé sur le décodeur CMU Sphinx et développé au Laboratoire d'Informatique de l'Université du Maine. L'apprentissage de LIUM RT a été en grande partie réalisé grâce à des articles issus du journal *Le Monde*, ce qui signifie que ce système n'est par essence pas optimisé pour reconnaître de la parole spontanée, telle qu'elle apparaît dans les débats ou les interviews.

2.2.1 Transcription manuelle *vs* assistée : une étude chiffrée

2.2.1.1 Protocole

Les données utilisées sont les suivantes : 24 segments d'environ 10 minutes, sélectionnés parmi les données non transcrites du corpus ESTER (France Inter, France Culture, RFI, France Info). 12 sont considérés comme étant de la parole spontanée (débats ou interviews), et 12 sont considérés comme étant de la parole préparée (informations).

France Info n'a pas été pris en compte pour la parole spontanée, car le caractère même de cette radio fait qu'elle ne contient que très rarement ce type de données. Par ailleurs, quand notre corpus comprenait des éléments non pertinents (musique...), ils n'ont pas été pris en considération. À l'aide de TRANSCRIBER, une transcription manuelle et une transcription assistée ont été effectuées sur chacun de ces fichiers par le même transcripteur. La seconde était réalisée suffisamment longtemps après la première pour qu'elle soit aussi peu influencée que possible, bien que le fait d'enchaîner rapidement plusieurs transcriptions différentes facilitait cette neutralité. Pour la transcription assistée, l'annotateur bénéficiait du texte brut

généré par LIUM RT, ainsi que d'une segmentation automatique (donc susceptible d'être erronée) en tours de parole et locuteurs, sans identification nommée de ces derniers.

Enfin, sur ces 24 segments, nous avons mesuré le temps nécessaire pour effectuer chacune des tâches suivantes :

- la transcription et la segmentation en tours de parole ;
- l'assignation des locuteurs ;
- la vérification orthographique.

Dans la mesure du possible, lorsqu'une tâche commençait à être chronométrée, le transcripteur essayait de ne pas s'interrompre avant de l'avoir terminée entièrement. Lorsque cela arrivait, le temps de la pause était décompté afin de ne pas biaiser les résultats. Par rapport aux transcriptions réalisées dans le cadre du projet EPAC, celles réalisées pour cette expérience différaient quelque peu : elles n'ont pas été ponctuées, et aucune balise n'a été ajoutée. Il n'y avait donc que le texte à l'état brut, afin d'apporter une certaine neutralité à l'ensemble. Ponctuer de l'oral est une démarche assez subjective, et les balises sont liées à des besoins spécifiques inhérents au projet EPAC. Or, nous voulions ici, pour le crédit scientifique de l'expérience, rester le plus objectif possible.

2.2.1.2 Principaux résultats

	Parole préparée	**Parole spontanée**
Transcription manuelle	17h36	19h33
Transcription assistée	8h31	15h44

Tableau 5 : Durée totale de la transcription (durées respectives des corpus : 2h08 et 2h10)

Le tableau 5 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. À titre de comparaison, 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 de façon 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. 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	13h36	16h15
Transcription assistée	5h06	12h41

Tableau 6 : Transcription du texte et segmentation

C'est lors de la tâche de transcription du texte (tableau 6) 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 (5h06 vs 13h36). Pour être plus précis, le fichier le plus significatif a obtenu un score de 3,75 : pour une durée effective de 00’’08’’55 (hh/mm/ss), la transcription assistée a ainsi demandé 00’’14’’49, et la transcription manuelle, 00’’55’’34. Ces chiffres sont très significatifs, notamment s'ils sont comparés à ceux obtenus avec la parole spontanée. Pour une durée sensiblement équivalente, le rapport global chute à 1,28. Quant au fichier qui présente le gain le plus important, celui-ci n’est que de 1,95. Pour 00’’11’’18 de parole, la transcription assistée nécessite 00’’47’’03, et la transcription manuelle, 01’’31’’52. Ces écarts mettent en évidence le fait que les système 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 (84 locuteurs et 180 tours de parole)	**Parole spontanée** (46 locuteurs et 648 tours de parole)
Transcription manuelle	1h17	2h13
Transcription assistée	1h17	2h13

Tableau 7 : Assignation des locuteurs

En ce qui concerne l'assignation des locuteurs (tableau 7), les chiffres rigoureusement identiques que nous avons obtenus viennent confirmer une évidence : un système de reconnaissance automatique de la parole n'est d'aucune aide pour cette tâche, dans la mesure où il n'identifie pas nommément les locuteurs. Il est surtout important de retenir que cette assignation demande presque deux fois plus de temps quand la parole est spontanée. Cela s'explique relativement facilement : la parole spontanée, avec ses nombreux tours de parole, contraint le transcripteur à leur assigner un locuteur, quand bien même il peut n'y en avoir que deux différents 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 de parole spontanée se trouve parfois de la parole superposée, et lorsque trois locuteurs ou plus sont susceptibles de prendre la parole, et il est parfois long et difficile de déterminer qui parle réellement.

	Parole préparée	**Parole spontanée**
Transcription manuelle	2h43	1h05
Transcription assistée	2h08	0h51

Tableau 8 : Correction orthographique

Le minutage de la correction orthographique (tableau 8) a permis d'observer que, si la différence spécifique entre transcription manuelle et assistée n'est certes pas très significative, celle entre parole préparée et spontanée l'est bien davantage. La raison en est simplement que les segments de parole préparée contiennent essentiellement de l'information radiophonique, c'est-à-dire un type de données très riche en noms propres, assimilables à des entités nommées (reporters, interviewés, personnalités, villes...), dont les orthographes exactes ne peuvent être systématiquement connues de l'annotateur. Les rechercher peut être une tâche assez longue, notamment dans le cas de noms étrangers. Inversement, les fichiers de parole spontanée, qui sont majoritairement des interviews ou des débats, comportent peu de noms propres, les thèmes abordés ne nécessitant en général qu'un faible recours à ceux-ci.

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

Tableau 9 : Taux d'erreur mot (%)

Les dernières observations effectuées (tableau 9) concernent le taux d'erreur mot, ou *Word Error Rate* (WER) [McCowan *et al.*, 2004]. Cet indicateur se calcule selon la formule suivante :

$$WER = \frac{S+D+I}{N}$$

où :

- S est le nombre de substitutions (le SRAP propose un autre mot que celui présent dans la transcription de référence)
- D est le nombre de suppressions (le SRAP supprime un mot par rapport à la transcription de référence)
- I est le le nombre d'insertions (le SRAP ajoute un mot par rapport à la transcription de référence)
- N est le nombre total de mots de la transcription de référence

On obtient alors un chiffre entre 0 et 1 (et même supérieur à 1 en cas de très mauvais résultats, puisqu'un seul mot dans la référence peut susciter plusieurs erreurs). Dans notre tableau, nous l'avons ramené à un pourcentage par souci de lisibilité. Ainsi 0,1695 est devenu 16,95% ; 0,3521 35,21% ; etc.

Le taux d'erreur mot a été mesuré à partir des sorties automatiques générées par le système LIUM RT, que nous avons comparées aux transcriptions manuelles puis assistées réalisées par l'annotateur (la « référence » désigné par la lettre *N* dans la formule). Les moyennes indiquées ci-dessus confirment ce que nous avons constaté précédemment, à savoir que le système de reconnaissance automatique du LIUM n'est pas aussi performant sur la parole spontanée que sur la parole préparée, ce qui est le cas des autres systèmes francophones équivalents (ceux développés au LIMSI ou au LIA par exemple). Les différences observées entre les tâches manuelles et assistées peuvent s'expliquer par le fait que

le transcripteur n'a pas forcément transcrit le même texte à chaque fois, dans la mesure où il est parfois difficile de percevoir clairement des phénomènes tels que les répétitions, les faux départs ou encore la parole superposée. Leur transcription ne sera donc pas toujours identique, même lorsqu'elle est réalisée deux fois par la même personne.

Ainsi, des résultats plus détaillés montrent que le segment qui a obtenu le plus faible taux d'erreur mot est celui qui a nécessité le moins de temps pour être transcrit. De même, il représente le gain le plus important entre la tâche de transcription manuelle et celle de transcription assistée (0h56 vs 0h15). Toutefois, le segment qui possède le taux d'erreur mot le plus fort (54,5%) n'a quant à lui pas demandé plus de temps qu'un autre pour être transcrit. Son principal locuteur est toutefois un photographe âgé qui ne s'exprime que de façon très sporadique, d'une voix presque chuchotée, sans articuler, ce qui explique en grande partie le pourcentage élevé. Ainsi, dans notre corpus, ce fichier est l'un des seuls pour lequel la transcription assistée a demandé à peu près le même temps de travail que la transcription manuelle (0h58 vs 0h57). Effacer les formes erronées puis ré-écrire le texte exact est un très long travail quand le taux d'erreur mot est élevé.

Enfin, si l'on considère la tâche de transcription assistée, on s'aperçoit finalement que le taux d'erreur mot est en adéquation avec le rapport entre durée totale de la transcription et durée totale des fichiers (Table 2). Un transcripteur professionnel a approximativement besoin de deux fois plus de temps pour corriger un segment spontané qu'un segment préparé, et le taux d'erreur mot est approximativement deux fois plus important avec la parole spontanée qu'avec la parole préparée.

2.2.1.3 Principaux enseignements

Ce travail montre que la transcription, qu'elle soit réalisée de façon manuelle ou assistée, est un travail qui nécessite beaucoup de temps. Les meilleurs résultats que nous ayons obtenus sont ceux qui combinaient parole préparée et transcription assistée, mais même dans ce cas de figure, transcrire un corpus de 10 heures signifie approximativement 40 heures de travail. Quant à la transcription manuelle de 10 heures de parole spontanée, elle en demanderait le double. Notre étude illustre également le fait que, hormis quelques cas isolés, recourir à un système de reconnaissance automatique de la parole permet de gagner du temps, quand bien même ce gain peut varier considérablement d'un fichier à l'autre.

L'assignation des locuteurs peut sembler relativement coûteuse en temps sur la parole

spontanée, mais il faut nuancer ce constat. En règle générale, transcription, assignation des locuteurs et correction orthographique sont effectuées conjointement par l'annotateur, et non de façon consécutive. En ce qui concerne les locuteurs, cela possède son importance car le fait de les assigner après avoir transcrit le texte contraint le transcripteur à « relire » son travail depuis le début. Pour la parole préparée, cette relecture prend assez peu de temps puisque les tours de parole sont généralement longs et clairement distincts. A l'inverse, la parole spontanée contient fréquemment des tours de parole brefs, avec de nombreux changements de locuteurs. Ainsi, le transcripteur doit réécouter presque chaque segment séparément, ce qui demande beaucoup plus de temps.

La fréquence des noms propres est une donnée intéressante dans la perspectives de futurs travaux. En tant que moyen de nature ontologique permettant de détecter la parole spontanée, elle pourrait être très utile et permettre un important gain de temps, notamment en ce qui concerne de grands corpus tels que ceux du projet EPAC.

2.2.2 Relevé, classement et analyse des principaux types d'erreur de LIUM RT

Cette expérience, parce qu'elle nous a confrontés directement aux sorties d'un SRAP (LIUM RT), a permis d'établir une typologie des erreurs commises par celui-ci. Si certaines sont inhérentes à la parole spontanée (problèmes avec les mots répétés, tronqués ou les assimilations), d'autres apparaissent indifféremment dans des données préparées ou spontanées. C'est notamment le cas des problèmes d'homonymie / paronymie ou d'ouverture de la voyelle « e ».

2.2.2.1 Homonymes / paronymes

La principale difficulté éprouvée par le système de reconnaissance automatique LIUM RT est de traiter les phénomènes d'homonymie et de paronymie. Ceux-ci sont particulièrement importants en français car les monosyllabes homophones sont beaucoup plus nombreux et employés que dans d'autres langues (à l'exception notable de l'anglais). Il n'est qu'à regarder le relevé établi par Daniel Luzzati dans *Le Français et son orthographe* [Luzzati, 2010] pour mesurer l'importance de ce phénomène.

À quoi cela tient-il ? En fait, à une réduction syllabique importante des mots français

tirés du latin (et beaucoup de mots français ont une étymologie latine). Ainsi, *hominem* a par exemple donné *homme*, passant de trois à une seule syllabe. Par opposition, le terme espagnol *hombre*, descendant du même étymon, compte lui deux syllabes.

Voici par exemple un tableau mettant en regard les différentes graphies de la séquence monosyllabique [ta~] et leurs traductions en italien, en l'occurrence toutes fondées sur les mêmes étymons latins.

Français	**Italien**
tant	tanto
temps	tempo
taon	tafano
tends (tendre pers 1)	tendo
tends (tendre pers 2)	tendi
tend (tendre pers 3)	tende
t'en (je t'en parle)	te ne (te ne parlo)

Tableau 10 : Graphies [ta~] et traductions en Italien

Comme on le voit dans ce tableau 10, toutes les traductions italiennes sont, sans exception, composées d'au moins deux syllabes. Dans la pratique, pour une seule voyelle française isolée, les possibilités monosyllabiques sont donc très diverses. Il n'est d'ailleurs même pas besoin de s'appuyer sur une combinaison consonne / voyelle pour le démontrer. Le son [o] peut, à lui seul, être orthographié de bien des manières : *oh*, *ho*, *au*, *aux*, *os*, *auls*, *haut*, *hauts*, *eau*, *eaux*...

Nous l'avons dit, ces combinaisons phonologiques consonne / voyelle sont également très nombreuses dans la langue anglaise. Cela étant, près de 35% d'entre elles ne renvoient à aucun mot usité. Tandis que dans la langue française, à peu près n'importe laquelle de ces combinaisons est susceptible de renvoyer à plusieurs entrées du dictionnaire ou du Bescherelle. En voici la démonstration avec la voyelle « u » :

- bu, bues, bue...
- chu, chus, chut...
- du, dû, dus...
- fut, fût, fus...
- jus, j'eus...
- lu, lue, lues...

- mû, mue, m'eus...
- nu, nus, nue...
- pu, pus, pue...
- cul, culs, qu'eut
- rue, ru, ruent...
- su, sut, c'eût...
- tu, tue, t'eut...
- vu, vus, vue...

Il est évident que cette multiplicité est un souci majeur pour les SRAP, qui se retrouvent ainsi confrontés à de multiples possibilités face auxquelles même un annotateur humain demeure parfois perplexe.

A cela vient s'ajouter un paramètre supplémentaire : il n'y a pas que les monosyllabes à être régulièrement homophoniques. Par le jeu des assimilations et des élisions propres à l'oral spontané, des formules telles que « pas de problèmes » ou « parce que » deviennent elles aussi sources de confusion. Ainsi, il est arrivé au cours de nos expériences que LIUM RT les orthographie « patte problème » ou « pas ce que ».

Voici une petite liste d'exemples similaires que nous avons relevés (à gauche se trouve la transcription de référence, et à droite la transcription automatique générée par LIUM RT) :

1) « là j(e) viens d'ouvrir » : *l'âge vient d'ouvrir*
2) « affirment elles avoir interpellé » : *affirmaient l'avoir interpellé*
3) « proches hein » : *prochain*
4) « chevauchement de compétence » : *chevauchent Mende compétences*
5) « sont là statiques i(l)s bougent pas » : *sont lasse Tati qui bougent pas*

Comme annoncé en préambule de ce paragraphe, la distinction parole préparée / parole spontanée n'est pas forcément la cause des erreurs du système. Dans l'exemple 5, l'élision du « l » appartient certes 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 2, qui est issu d'un flash d'informations, aucune altération phonologique n'apparaît. Pourtant, cela n'empêche pas le système de proposer une séquence,

acoustiquement exacte mais sémantiquement erronée. Ce type d'erreurs, difficilement évitable au vu des technologies dont on dispose aujourd'hui, est donc susceptible d'apparaître quel que soit le contexte langagier dans lequel on se trouve. En effet, comme nous le disions précédemment, même un annotateur humain peut ici hésiter entre les deux propositions, s'il ne bénéficie pas d'informations suffisantes sur ce qui a été dit avant. En conséquence, il est impossible pour une machine de faire la même démarche, car cela nécessiterait une puissance de calcul beaucoup trop importante, du moins à l'heure actuelle.

2.2.2.2 « 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 sources de confusions (*je 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 :

6) « j'ai été » : *j'étais*
7) « le papa c'est une » : *le pas passé une*
8) « c'est cool » : *s'écoule*
9) « traîner » : *Trénet*
10) « 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 :

11) « et ça » : *SA*
12) « et euh » : *et E*

13) « j'ai j'ai » : *g g*

14) « c'était » : *CT*

Il est amusant de constater ici l'analogie avec le langage SMS, friand de cette forme d'orthographe simplifiée à l'extrême. On note également la confusion récurrente entre « e » ouvert et « e » fermé, notamment dans l'exemple 11 où l'écoute du segment concerné fait clairement entendre la prononciation [esa] et non [Esa]. Paradoxalement, il se peut aussi que le système ne reconnaisse pas un sigle et le transcrive sous forme de mots :

15) « MSA » : *mais ça*

Pour poursuivre l'analogie, on pourrait ici assimiler le raisonnement de LIUM RT à celui d'un traducteur de langage SMS. 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.

2.2.2.3 Assimilations

Cela dit, il est effectivement des spécificités de la parole spontanée qui sont source d'erreur d'interprétation du logiciel de reconnaissance automatique, et notamment l'assimilation. Nous reviendrons longuement sur la définition phonologique de l'assimilation en 5.2.4.2., mais ils convient déjà de l'expliciter pour permettre au lecteur de saisir les problèmes qu'elle pose aux SRAP. En quelques mots, l'assimilation est donc une variation phonétique entraînant la modification de la prononciation d'une consonne sourde au contact d'une consonne voisine sonore (ou l'inverse). Par exemple, la séquence « je crois », dans un cadre spontané, sera souvent prononcée [SkRwa], en raison du contact direct entre le « j » et le « c ». Ce contact est provoqué par la chute du « e », spécifique de l'oral.

Or, les systèmes de reconnaissance automatique de la parole, souvent peu entraînés à ce genre de phénomène, ne savent 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 d'en relever plusieurs occurrences lors de nos expériences :

16) « envie d(e) passer » : *vite passé*
17) « pas d(e) sanitaires » : *patte sanitaire*
18) « 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.

Des séquences avec les monosyllabiques « que », « se » et « te » produisent les mêmes effets phonologiques, à savoir des glissements respectifs vers les sons « gue », « ze » et « de ». A nouveau, il en résulte de possibles confusions homophoniques, théoriquement beaucoup moins répandues dans un cadre de parole préparée (puisque l'élision du schwa y est nettement moins fréquente).

19) « cet élan qu(e) vous chassez » : *c'était langue vous chassez*
20) « on s(e) demande quand même » : *onze demandes quand même*
21) « et tu t(e) dis que bon, voilà quoi » : *étude dit que bon voilà quoi*

2.2.2.4 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 :

22) « faut faut faut faut » : *faut fois font fois*

Les troncations ou les faux départs 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 :

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

2.2.2.5 Autres observations

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 un « que ».

2.2.3 Perspectives pour optimiser les systèmes de reconnaissance automatique de la parole

Pour conclure, nous nous proposons de réfléchir à des hypothèses qui permettraient d'optimiser les systèmes de reconnaissance automatique de la parole face à la parole spontanée. 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 souhaitable qu'on puisse les utiliser librement, à commencer par ceux qui ont été financés sur

des fonds publics. Ces données sont trop rares pour qu'elles ne puissent pas contribuer dans leur globalité à l'amélioration des systèmes de reconnaissance de la parole, qui passe par la masse de corpus transcrits et alignés mis à la disposition des laboratoires. Les incompatibilités de transcriptions, de codages, de formats... sont compréhensibles. La confidentialité des données ne l'est pas, surtout lorsqu'il s'agit d'organismes de recherche publics.

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 avons contribué à une étude explicitée dans [Dufour, 2010]. Celle-ci avait pour but d'ajouter au dictionnaire de prononciation de LIUM RT des variantes de prononciation spécifiques de l'oral (parce que = « pasque », celui-là = « çui-là »). Ces variantes, en plus d'être nombreuses, concernent des mots très fréquemment utilisés dans la conversation courante (notamment les pronoms personnels, voir 5.2.1.1). S'il ne la possède pas dans son dictionnaire, un SRAP va, s'il se trouve confronté à l'une de ces variantes, la supprimer ou bien lui substituer un autre mot. Afin d'éviter ces deux écueils qui, à grande échelle, fait augmenter de manière considérable le taux d'erreur mot, nous avons tout d'abord établi une liste de mots ou expressions susceptibles d'être concernés par ce problèmes. Principalement, il s'est agi de mots dont une ou plusieurs lettres étaient élidées. Il est arrivé également que des phénomènes d'assimilations rentrent en compte, notamment en ce qui concerne les verbes conjugués dont le sujet est « je » (voir 5.2.4).

Puis, les variantes de prononciation ont été ajoutées au dictionnaire de prononciations de LIUM RT. Celui-ci contenait initialement 165 000 variantes de prononciations pour environ 62 000 mots. Suite à cette démarche, 1761 prononciations ont été ajoutées concernant 631 mots, soit environ 1% du dictionnaire.

[Dufour, 2010] a ensuite mené plusieurs expériences afin de mesurer l'impact de ces ajouts dans un contexte spontané. En se basant sur une quinzaine d'heures issues du corpus EPAC, il a ainsi démontré que les suppressions et substitutions diminuaient sensiblement (3111 contre 3369), faisant descendre le taux d'erreur mot de 1,72 % (20,82 % contre 22,54 %). Cela étant, ces résultats n'étaient basés que sur les mots concernés par les variantes de

prononciations ajoutées au dictionnaire. Une fois étendu à l'ensemble des mots du corpus de quinze heures, le gain a diminué de moitié (0,81 %). Enfin, le fait de prendre en compte les insertions, en plus des suppressions et des substitutions, a finalement abouti à un gain négatif de -0,9 %.

Il semble donc qu'à défaut d'être moins nombreuses, les erreurs soient simplement différentes. Le fait d'intégrer de nouvelles prononciations a également multiplié les choix possibles, notamment concernant des « mots usuels courts » tels que *de*, *je* ou *il* [Dufour, 2010]. Également, cela a créé de nouvelles ambiguïtés homophoniques, qui plus est dans une langue qui en comprenait déjà beaucoup. Ainsi, lors d'une séquence telle que « c'est pas c(e) que j'ai dit », le système est désormais amené à considérer la solution « c'est parce que j'ai dit », hypothèse qu'il n'envisageait pas auparavant.

Cela étant, et moyennant quelques modifications et paramétrages supplémentaires, il nous semble raisonnable de penser que l'adaptation du dictionnaire de prononciations à un contexte précis (ici, la parole spontanée) est une piste à approfondir si l'on souhaite améliorer les performances d'un SRAP comme LIUM RT.

Chapitre 3 : Les logiciels d'aide à la transcription

Il existe de nombreux logiciels utilisés pour réaliser la transcription orthographique et l'annotation d'un fichier audio, voire video : TRANSCRIBER et PRAAT sont assurément les plus répandus [Antoine, 2008], mais avec des objectifs différents (transcription de gros corpus pour l'un, analyse prosodique pour l'autre). WINPITCHPRO, CLAN, ELAN, EXMARALDA ou encore TRANSANA sont moins sollicités à l'heure actuelle, pour des raisons que nous nous efforcerons de préciser par la suite. Chacun de ces outils va désormais être présenté en détails afin de cerner leurs possibilités. Ils sont orientés vers des besoins différents, et aboutissent non pas à des transcriptions divergentes mais à des représentations de ces transcriptions nettement distinctes.

Ces logiciels ne sont pas exclusivement réservés au monde scientifique ou professionnel. La simplicité d'accès de certains d'entre eux (et notamment TRANSCRIBER) les destinent également à des usages privés, lorsqu'un utilisateur souhaite simplement transcrire le contenu d'une réunion par exemple. Cela lui offre notamment la possibilité, par rapport à un traitement de texte classique, d'aligner sa transcription avec le signal audio (il faut bien sûr pour cela disposer de l'enregistrement numérisé de la réunion).

Toutefois, la plupart des autres logiciels que nous évoquerons sont réservés à des usages spécifiques, et ne concernent finalement que des communautés d'utilisateurs restreintes. Sans compter TRANSCRIBER, celle de PRAAT est sans doute la plus répandue, grâce aux possibilités que cet outil offre pour traiter la prosodie. Bon nombre de phonéticiens et de phonologues y ont ainsi recours, d'autant que le logiciel existe depuis maintenant dix-huit ans. Enfin, il est librement distribué et régulièrement mis à jour, ce qui contribue à sa popularité.

A l'inverse, de par leur mode de distribution restreint ou payant, des logiciels comme WINPITCHPRO, TRANSANA ou ANVIL ne sont pas aussi répandus. A l'heure où sur Internet, le logiciel libre connaît de plus en plus de succès (et notamment le site

http://www.sourceforge.net, qui héberge d'ailleurs TRANSCRIBER), les utilisateurs tendent désormais à refuser ce qui exigera d'eux une contribution financière, aussi modeste soit-elle.

Il sera enfin intéressant de voir, dans les cinq ou dix années à venir, l'importance que prendra EXMARaLDA. Plus qu'un logiciel, c'est en fait un ensemble d'outils destinés à la transcription de corpus. Encore relativement jeune (2001), sa gratuité et cet aspect multi-tâches le font de plus en plus connaître, d'autant que leurs auteurs allemands ont pris soin de traduire leur site en anglais et en français pour le rendre accessible au plus grand nombre.[18]

18 Quelques-unes des figures présentées dans les pages qui suivent sont directement issues des sites internet des logiciels concernés

3.1 TRANSCRIBER

Page Web : http://trans.sourceforge.net/en/presentation.php
Auteur(s) : Karim Boudahmane, Mathieu Manta, Fabien Antoine, Sylvain Galliano, Claude Barras
Date de création du logiciel : 25 mai 1998
Dernière mise à jour du logiciel : janvier 2005
Plates-formes : Windows, Unix, Macintosh
Statut : logiciel libre
Langue de l'interface : anglais, français, tchèque
Format natif : .trs
Autres formats supportés : .cha, .sdt, .sgml, .stm, .txt, .typ, .utf, .xml
Formats d'entrée audio : .aif, .aiff, .au, .mp3, .ogg, .sd, .smp, .snd, .sph, .wav
Formats d'entrée vidéo : .mov (seulement sous Macintosh)
Formats d'exportation : .html, .lbl, .stm, .txt, .typ
Publications : [Barras *et al.*, 1998], [Barras *et al.*, 2000], [Geoffrois *et al.*, 2000]

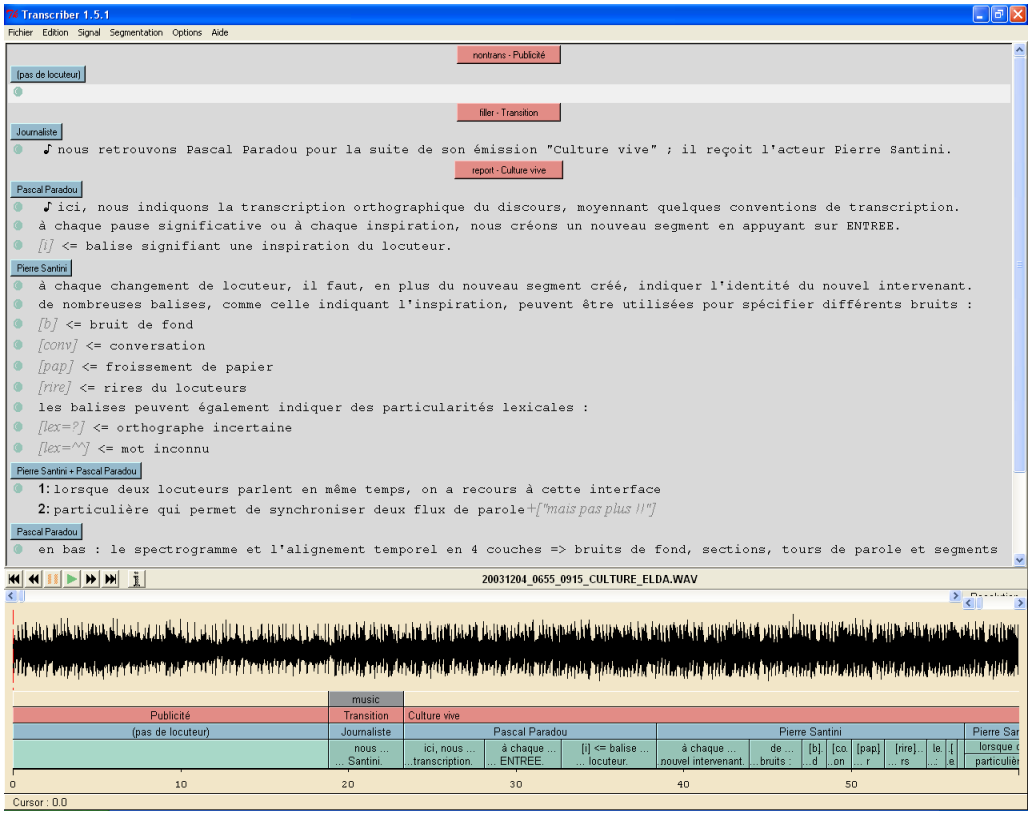


Figure 7 : Interface générale du logiciel TRANSCRIBER

TRANSCRIBER est le logiciel que nous avons majoritairement utilisé, puisque c'est avec celui-ci que les transcriptions du corpus EPAC ont été réalisées. Il présentait en la circonstances un nombre d'avantages non négligeables, bien que certaines de ses limites aient été mises en avant par la nature même du corpus EPAC, contenant principalement de la parole spontanée. En effet, à l'origine, TRANSCRIBER a été développé par la DGA (Direction Générale de l'Armement) pour transcrire essentiellement des journaux d'informations, donc contenant essentiellement de la parole préparée.

Au niveau de son interface globale, TRANSCRIBER se veut très accessible. A l'ouverture du logiciel, il est simplement demandé à l'utilisateur d'ouvrir un fichier son au format .wav. Une fois cette démarche accomplie, la transcription en elle-même peut commencer. Cette simplicité semblerait volontiers banale, mais aucun des autres logiciels que nous avons testés ne s'y soumet réellement. Il est même parfois difficile de ne faire « que » transcrire, sans manipulations ou démarches préalables, et nous aurons l'occasion d'y revenir par la suite. Précisons toutefois que sous la simplicité se cachent parfois quelques lacunes. En l'occurrence, au niveau des formats d'entrée, TRANSCRIBER ne propose aucune possibilité au sujet de la vidéo, si ce n'est le format .mov sous Macintosh seulement. Signalons enfin que TRANSCRIBER supporte en entrée de nombreuses extensions, notamment les fichiers .stm, .cha (CLAN), .txt, .xml, etc. Les fichiers qu'il génère sont en .trs, un format très proche du XML, qui assure une relative pérennité des données tout en facilitant leur interopérabilité. TRANSCRIBER permet également l'exportation des transcriptions en .html, .txt ou bien .stm, ce qui assure une certaine portabilité aux données, et ce pour tout type d'usages, du plus simple au plus complexe.

Les fichiers audio sont gérés de façon très robuste, ce qui est suffisamment rare pour être souligné. Nous avons plusieurs fois travaillé avec des fichiers .wav de très grande taille (12 heures d'audio, soit près de 1,5 Go), et cela n'a posé aucun souci de traitement au logiciel. La navigation temporelle est d'ailleurs particulièrement efficace, puisqu'il est possible d'agrandir ou de diminuer l'échelle du spectrogramme, et de se déplacer de façon plus ou moins précise selon que l'on utilise ledit spectrogramme ou bien la barre de navigation située au-dessus.

Outre le texte lui-même, TRANSCRIBER offre un espace spécifique dédié au(x) locuteur(s). Là encore, c'est une particularité d'autant plus utile qu'on la retrouve assez peu dans les autres outils d'aide à la transcription. En plus de l'identité d'un intervenant, il est ainsi

possible de préciser son sexe, son origine géographique, le type de parole employé, la qualité et le type du canal utilisé (studio ou téléphone). Bien sûr, aucune de ces informations n'est obligatoire et chacun peut juger bon de les spécifier ou non, suivant le temps dont il dispose et la précision qu'il souhaite donner à sa transcription. Toutefois, en vue de recherches futures, elles peuvent être précieuses (distinguer les locuteurs masculins et féminins, dissocier les tours de parole spontanés et préparés, ne traiter que les données de nature téléphonique, ne s'intéresser qu'aux locuteurs n'ayant pas le français pour langue maternelle...).

Une autre particularité remarquable de TRANSCRIBER est son système de balises. Le logiciel en intègre un nombre impressionnant, qui permettent en une combinaison de touches (pour peu que l'on ait paramétré un raccourci) d'annoter toutes sortes d'événements, qu'ils soient linguistiques, lexicaux ou encore phonologiques. Voici une liste presque exhaustive qui donnera une idée de la granularité d'annotation qu'il est possible de réaliser avec TRANSCRIBER, sachant que l'utilisateur peut ajouter à cette liste autant de créations personnelles qu'il le souhaite :

Bruits :

- respiration, inspiration, expiration, reniflement, souffle
- bruit de bouche, bruit de gorge, toux, rire, sifflement
- bruit indéterminé, conversation de fond, froissement de papier, souffle électrique, bruit micro, toux en fond, rire en fond, indicatif, jingle, top, musique...

Prononciation :

- mal prononcé, lapsus, inintelligible, inintelligible/faible, voix chuchotée
- sigle lu, sigle épelé...

Lexique :

- orthographe incertaine
- mot inconnu
- néologisme
- rupture de syntaxe

On le voit, les possibilités sont très larges et autorisent vraiment une annotation très

détaillée. Qui plus est, nous n'avons pas mentionné dans cette liste les entités nommées, qui représentent presque une interface à elles seules. En effet, il est possible de choisir, dans les options de TRANSCRIBER, d'afficher ou non la fenêtre les concernant. Lorsque cette option est activée, la fenêtre principale du logiciel se réduit et l'affichage prend alors cette forme :

Figure 8 : Fenêtre d'interface des entités nommées de TRANSCRIBER

Le fait de consacrer une place si importante à ce concept d'entités nommées peut paraître curieux, d'autant qu'aucun autre logiciel présenté ici ne s'y attache. En fait, il s'agit d'une mise à jour « sur mesure » pour la campagne d'évaluation ESTER, à laquelle le logiciel est étroitement lié : en effet, les transcriptions fournies aux participants avaient été réalisées avec TRANSCRIBER. Or, parmi les tâches qui devaient être réalisées durant cette campagne, il y avait l'annotation de corpus en entités nommées. Comme aucune fonction ne permettait jusqu'alors de réaliser ce travail efficacement, la DGA a implanté spécifiquement, dans la version 1.5 du logiciel, une interface dédiée.

A nouveau, ce sont donc ici de nouvelles possibilités d'annotation proposées au

transcripteur, bien que le processus soit un peu différent des balises que nous évoquions précédemment. Les entités nommées ont plutôt pour but l'indexation de données, agissant comme une sorte de répertoire des noms communs et des noms propres. Elles ne sont en l'occurrence pas directement dépendantes de la tâche de transcription.

D'autres possibilités d'annotation sont encore disponibles avec TRANSCRIBER. Une sorte de balise « passe-partout » tout d'abord, qui porte l'intitulé « commentaire », et qui permet de spécifier tout type d'informations, et ensuite deux couches d'annotation que nous appelleront « section » et « bruit de fond ».

Symbolisées sur la figure 8 par les rectangles de couleur rosée, les sections permettent à l'annotateur de renseigner deux champs distincts. L'un, contraint, laisse le choix entre *report* (les données jugées pertinentes par l'annotateur, et donc transcrites), *filler* (les données de transition comme les pages de publicités, pas forcément transcrites) et *nontrans* (tout ce qui ne sera pas transcrit). L'autre, facultatif, permet d'apporter des précisions supplémentaires, comme le nom des programmes transcrits ou le genre de publicité, de plages musicales, etc.

Les bruits de fond, quant à eux, se caractérisent simplement par la présence de notes musicales dans l'interface de TRANSCRIBER, et sont représentés par la strate grisée sur l'échelle temporelle. Quatre choix s'offrent à l'annotateur pour les caractériser : *music*, *shh* (bruits électriques de type micro par exemple), *speech* ou *autre*. Ainsi, plutôt que d'avoir à utiliser des balises qui devraient parfois couvrir des périodes extrêmement longues, à tel point que l'on ne saurait même plus ce que certaines ouvrent ou closent, il est possible de recourir à cette option. Elle est d'une part plus facilement utilisable dans ce cas de figure précis, et d'autre part bien mieux représentée à l'écran grâce à sa tire[19] attitrée.

Au sein même du logiciel TRANSCRIBER, beaucoup d'éléments sont paramétrables. En plus de la possibilité (très utile) de créer autant de raccourcis clavier que l'on souhaite pour les balises, il est possible de modifier la langue, les fontes, les couleurs, l'affichage de chaque élément...

TRANSCRIBER est un logiciel libre et multi plate-formes, deux arguments de poids dans le monde du logiciel aujourd'hui puisqu'ils permettent un large usage du produit, de même qu'ils offrent aux utilisateurs avancés la possibilité de le faire évoluer. Par exemple, Mathieu Quignard est parvenu a reprogrammer une partie du logiciel de façon à résoudre le

19 Ce que nous appellerons « tire », tout au long de notre étude, désignera un niveau ou une couche d'annotation.

problème des segments superposés[20]. Ainsi, il est désormais possible de rendre illimité le nombre de flux de parole pouvant être synchronisés temporellement.

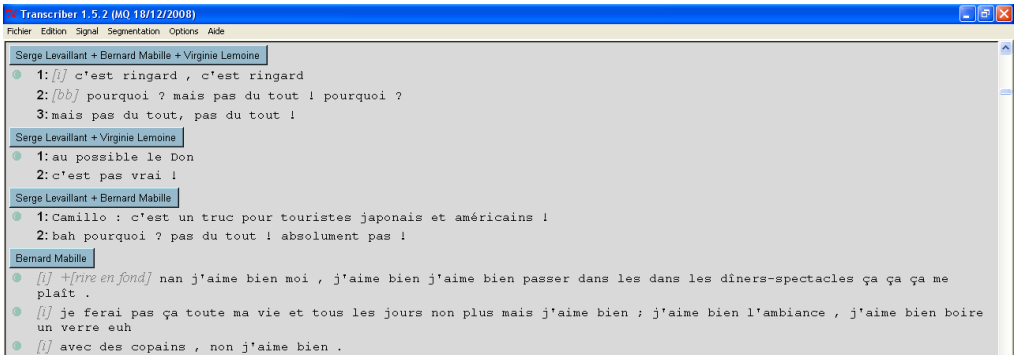


Figure 9 : Représentation d'un dialogue avec trois flux de parole simultanés

Au final, TRANSCRIBER se révèle naturellement très efficace pour transcrire de la parole préparée, et ce même avec des fichiers audio de plusieurs gigaoctets. Mais il souffre d'un gros défaut pour qui voudrait le confronter à des données multi-locuteurs : l'impossibilité de gérer pertinemment plusieurs flux de parole superposée.

Cela dit il est le seul à proposer, notamment grâce à son utilisation dans la campagne ESTER, un jeu d'annotations aussi complet et aussi modulable. Ce qui est presque un paradoxe étant donné que son interface est très simple à maîtriser, et que même une personne non-initiée aux logiciels informatiques s'y retrouvera au bout de quelques heures. La gestion très (trop ?) minimaliste de la vidéo est assez regrettable, d'autant qu'il est possible de créer ses propres annotations : il aurait alors été aisé de concevoir une nomenclature pour annoter les regards, les gestes, etc.

Une nouvelle version de TRANSCRIBER est annoncée depuis mai 2008, et celle-ci doit corriger la plupart des problèmes que nous avons mentionnés. Mais à l'heure où sont écrites ces lignes (septembre 2010), elle n'est toujours pas disponible.

20 http://www.loria.fr/~quignard/Transcriber

3.2 PRAAT

Page Web : http://www.fon.hum.uva.nl/praat/
Auteur(s) : Paul Boersma, David Weenink
Date de création du logiciel : 1992
Dernière mise à jour du logiciel : 11 mai 2011
Plates-formes : Windows, Unix, Macintosh
Statut : logiciel libre
Langue de l'interface : anglais
Format natif : .textgrid
Autres formats supportés : aucun
Formats d'entrée audio : .aifc (non compressé), .aiff, .au, .nist, .wav
Formats d'entrée vidéo : aucun
Formats d'exportation : aucun
Publications : [Boersma, 2001], [Delais-Roussarie *et al.*, 2006], [Goldman, 2006]

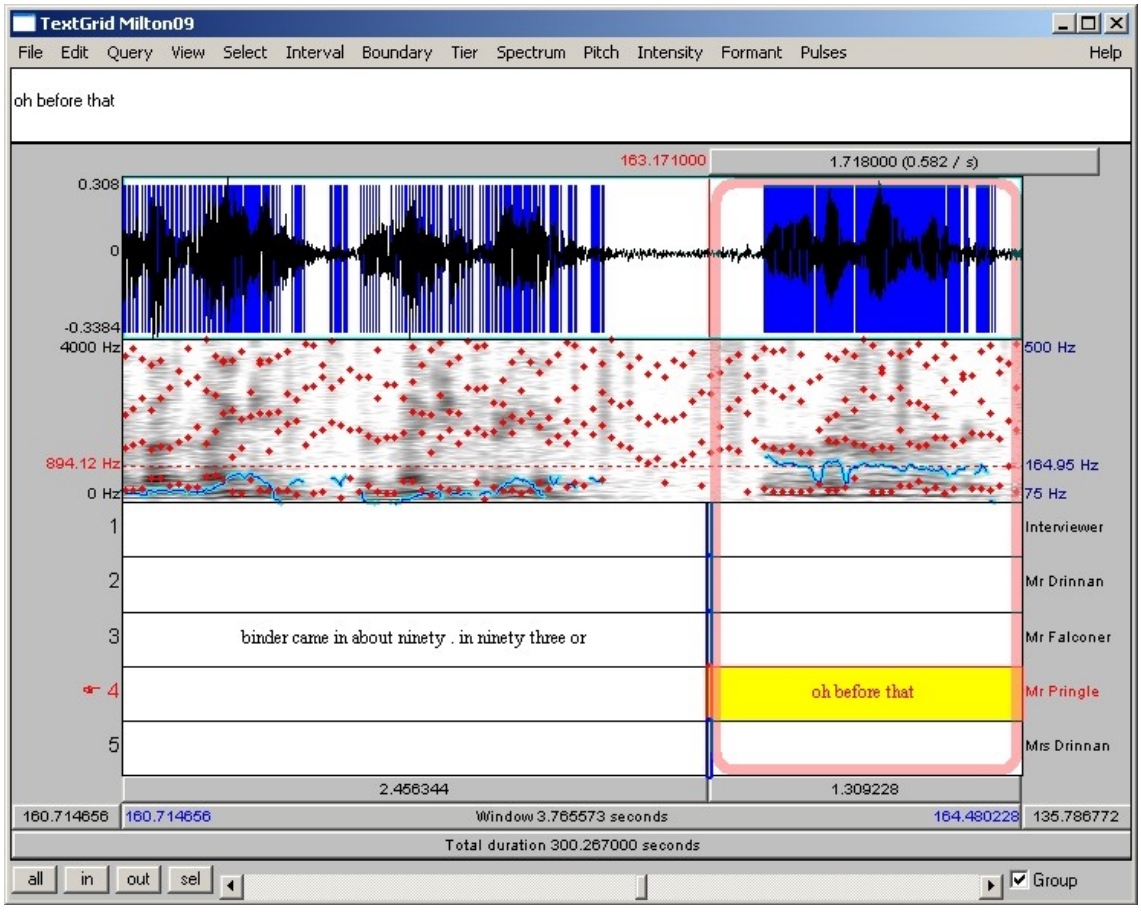


Figure 10 : Exemple d'une transcription réalisée sous PRAAT

A la différence de TRANSCRIBER, PRAAT n'est pas optimisé pour la transcription de gros corpus audio. Outre le fait qu'il ne supporte pas les fichiers volumineux (nous avons obtenu un message d'erreur avec un fichier .wav d'environ 450 Mo), ses fonctionnalités majeures le destinent plutôt à des usages spécifiques : phonétique, phonologie ou prosodie notamment (voir la figure 10 ci-dessus). A ce titre, de nombreux plug-ins sont disponibles pour permettre des niveaux d'analyse encore plus poussés, et font de PRAAT un logiciel très répandu, d'autant qu'il est multi plate-formes, libre et régulièrement mis à jour.

Même si ce n'est donc pas son usage premier, PRAAT propose bel et bien la possibilité d'aligner des segments de parole sur un fichier audio puis de les transcrire ; c'est pourquoi nous allons ici détailler son utilisation dans ce cadre.

Lors de sa première utilisation, PRAAT est quelque peu déroutant puisque son interface principale est divisée en deux fenêtres : l'une intitulée « Praat Objects », et l'autre « Praat Picture ». La première permet d'exécuter les fonctionnalités basiques du logiciel ; la seconde est destinée à des traitements plus spécifiques (traitement et impression de graphiques notamment) que nous ne détaillerons pas ici.

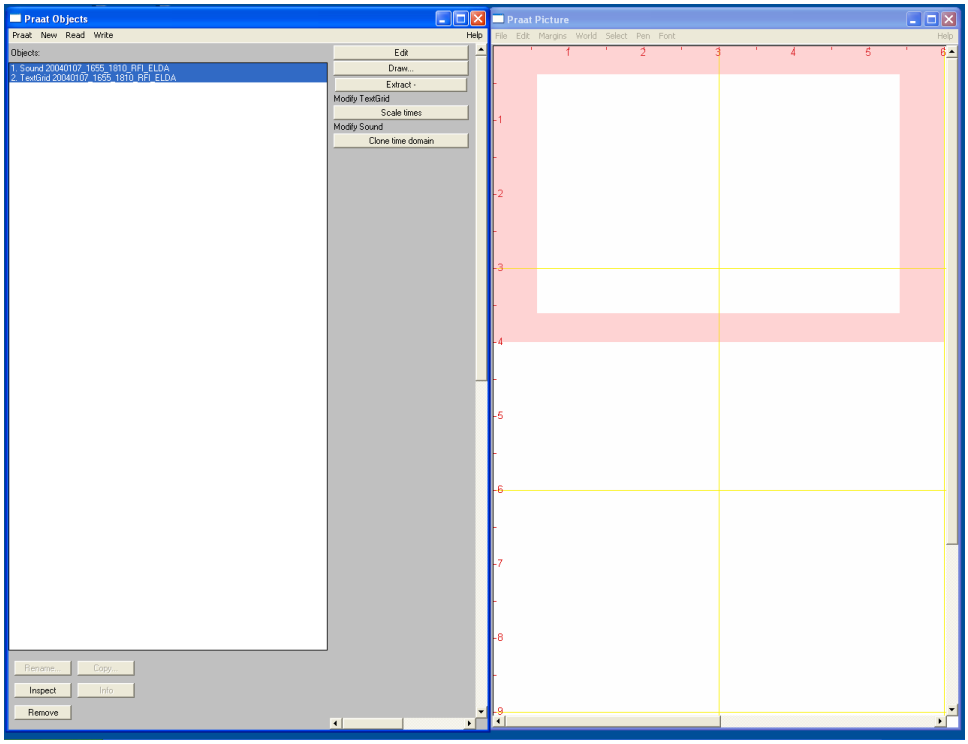


Figure 11 : Fenêtres principales de PRAAT

Pour accéder à la fenêtre permettant entre autres la transcription d'un signal audio (PRAAT ne gère aucune format vidéo), il faut tout d'abord la générer. Pour ce faire, il convient d'indiquer au logiciel l'emplacement du fichier son que l'on souhaite traiter, puis de préciser ensuite la tâche voulue, à savoir la transcription. Un fichier intitulé « textgrid » sera alors créé, qu'il conviendra de sélectionner conjointement au fichier son pour ensuite pouvoir passer à l'étape de transcription proprement dite.

Celle-ci se déroule donc dans la fenêtre présentée à travers la figure 10. Comme on le voit, la présentation globale est exclusivement horizontale, à la différence de TRANSCRIBER qui propose un déroulement vertical en ce qui concerne les segments et les tours de parole. Ce mode de fonctionnement a pour principal inconvénient d'imposer la création d'une tire par locuteur, tire qui restera présente à l'écran tout au long de la transcription même si ce locuteur n'intervient qu'une seule fois. Cet exemple illustre les limites de PRAAT concernant le traitement de corpus à grande échelle : plusieurs fichiers du corpus EPAC contiennent ainsi au moins dix locuteurs, et on imagine bien les problèmes de représentation qu'une telle densité poserait ici.

La fenêtre consacrée à la saisie du texte se situe dans la partie supérieure. Il s'agit d'un cadre tout à fait classique, et quelques options utiles sont proposées, comme la gestion de la taille des caractères. Aussitôt le texte saisi dans cet espace, il apparaît conjointement dans la strate allouée au locuteur concerné. A nouveau, PRAAT se montre ici différent de TRANSCRIBER. Si les segments à traiter sont relativement longs, leur affichage sera tronqué, avec des mots brutalement coupés en fin de zone de parole. Il est plus surprenant qu'un problème similaire arrive avec les patronymes des locuteurs : au-delà d'une quinzaine de caractères, ils seront tout simplement tronqués, sans aucune possibilité de les faire apparaître dans leur intégralité (figure 12).

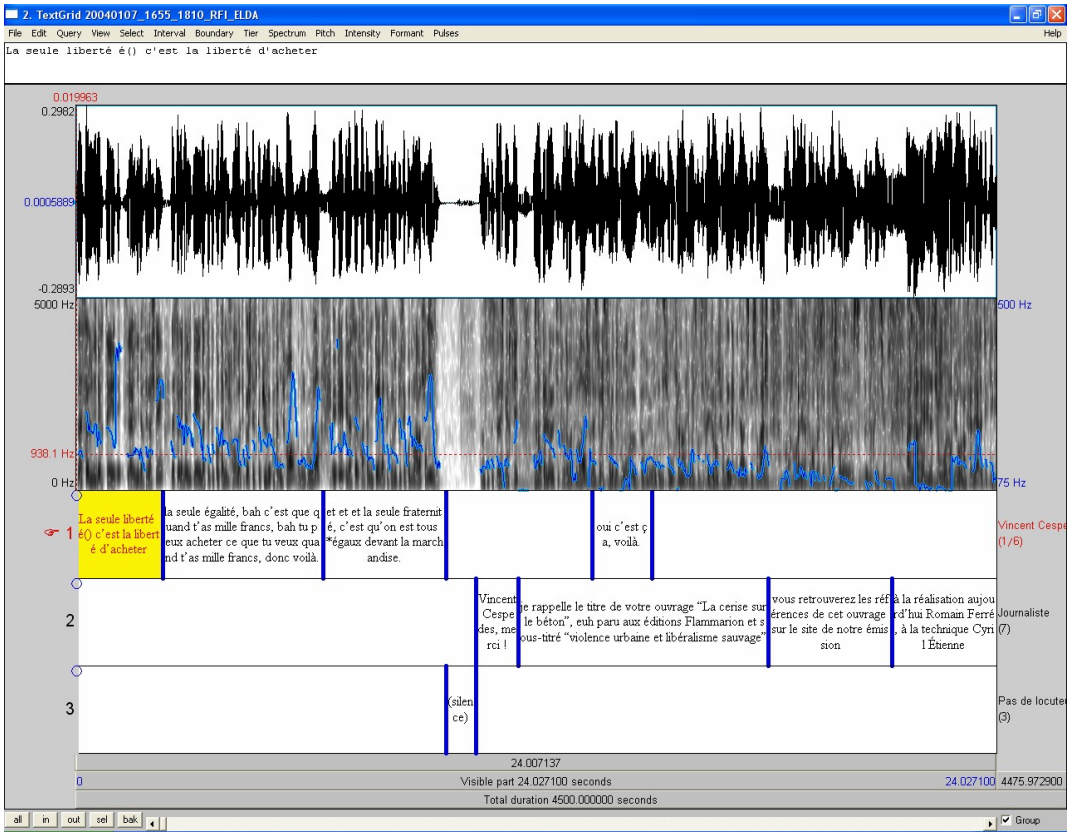


Figure 12 : Affichage des locuteurs sous PRAAT

La figure 12 nous permet également d'évoquer quelques autres possibilités de PRAAT, comme l'alignement temporel spécifique à chaque locuteur. Cette fonctionnalité est très précieuse pour annoter la parole superposée. En effet, et contrairement à TRANSCRIBER, il n'est nul besoin avec PRAAT de repérer précisément le passage où deux flux de parole se chevauchent. Il est tout à fait possible (et beaucoup plus pratique) de les traiter un par un, en ancrant séparément leurs segments respectifs sur l'échelle temporelle.

Pour autant, la gestion des locuteurs n'est pas sans poser quelques soucis. Comme on le voit ici, la troisième tire est intitulé « pas de locuteur », et il y figure un segment où se trouve la mention « (silence) ». Cela signifie tout simplement que les zones qui ne contiennent pas de données pertinentes à transcrire (dans notre cas, les silences, mais aussi les passages musicaux, les publicités...). Elles doivent faire l'objet d'une ou de plusieurs sections spécifiques qui peuvent alourdir considérablement la représentation générale de la transcription.

Mais, nous l'avons dit, PRAAT se destine avant tout à des études particulièrement précises, et cette visée explique les quelques absences que nous avons mentionnées. De même, on ne s'étonnera guère, pour la même raison, de ne pas trouver de balises ou toute autre possibilité d'annotation prédéfinies. Bien sûr, il est possible de créer ses propres conventions et d'annoter manuellement ce que l'on souhaite, mais au prix d'une perte de temps considérable puisqu'aucun raccourci clavier n'est paramétrable.

Les transcriptions ainsi générées sont en .textgrid, un format spécifique à PRAAT. Bien que ne proposant pas d'autres possibilités d'exportation, les fichiers .textgrid peuvent être convertis, grâce à des outils spécifiques, vers différents formats.

A l'inverse de TRANSCRIBER, PRAAT est donc un logiciel tout à fait exploitable en tant que « microscope », pour des analyses prosodiques précises ou des transcriptions alignées sur les phonèmes par exemple. En revanche, il se montre particulièrement inapproprié pour traiter des corpus de grande ampleur.

3.3 WINPITCHPRO

Page Web : http://www.winpitch.com/
Auteur(s) : Philippe Martin
Date de création du logiciel : 1996
Dernière mise à jour du logiciel : 13 décembre 2010
Plates-formes : Windows
Statut : logiciel payant (100 $) ; version d'essai gratuite (15 jours)
Langue de l'interface : anglais
Format natif : .wp2
Autres formats supportés : .trs, .textgrid, .txt., .rtf
Formats d'entrée audio : .wav, .mp3, .wma, .cda
Formats d'entrée vidéo : .avi, .mp4, .mpg
Formats d'exportation : .xml, excel
Publications : [Martin, 2003], [Martin, 2004], [Delais-Roussarie *et al.*, 2006]

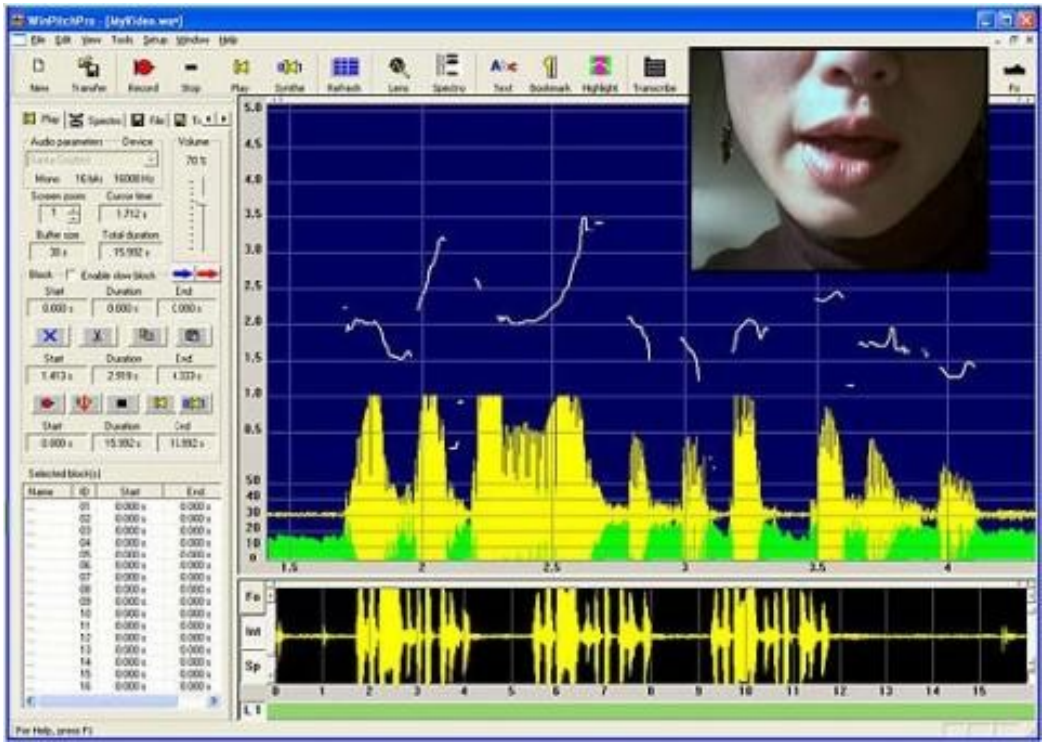


Figure 13 : Interface générale du logiciel WINPITCHPRO

WINPITCHPRO, une fois n'est pas coutume, est le fait d'une seule et même personne : Philippe Martin, qui en a assuré l'intégralité du développement ainsi que ses mises à jour. A l'instar de PRAAT, WINPITCHPRO est un logiciel qui permet certes de transcrire et d'aligner du texte avec un flux audio, mais qui propose aussi d'autres possibilités, dans le domaine de la prosodie ou de la phonétique notamment. Pour poursuivre le parallèle entre les deux outils, on pourra également préciser que l'interface réservée à la transcription est plus ou moins identique. Les tires viennent se placer en dessous du spectrogramme (du moins, par défaut, cette disposition est paramétrable), et il est possible d'en créer jusqu'à 96. Un tel chiffre laisse imaginer les diverses possibilités d'analyse et de traitement envisageables d'autant que, nous l'avons dit, WINPITCHPRO offre à l'utilisateur beaucoup d'angles d'approche. Il est notamment le seul des logiciels présentés ici qui permet de *créer* ses données, c'est-à-dire d'enregistrer un ou plusieurs flux de parole (moyennant naturellement l'utilisation de matériel adapté pour la prise de son). Tout est paramétrable, depuis la source audio jusqu'à la qualité des sorties, en passant par l'arrêt automatique programmable de l'enregistrement ou le volume de la prise de son. Lors de l'enregistrement proprement dit, WINPITCHPRO affiche instantanément les courbes correspondant à la fréquence fondamentale et l'intensité du signal, ce qui autorise des ajustements ou même des analyses en temps réel.

Puisque nous venons d'évoquer la fréquence fondamentale, il convient de préciser que WINPITCHPRO autorise un traitement très précis la concernant. Cinq moteurs d'analyse indépendants sont en effet disponibles, et peuvent être activés sur l'ensemble du signal ou sur quelques segments isolés. Ils offrent ainsi la garantie d'analyses robustes, même lorsque les données sont bruitées. Plus largement, il est possible d'indiquer sous forme graphique divers éléments prosodiques (intensité, pauses, durée des segments...), et ceux-ci peuvent être ajustés directement par l'utilisateur puisque leurs valeurs apparaissent dans une fenêtre indépendante (figure 14).

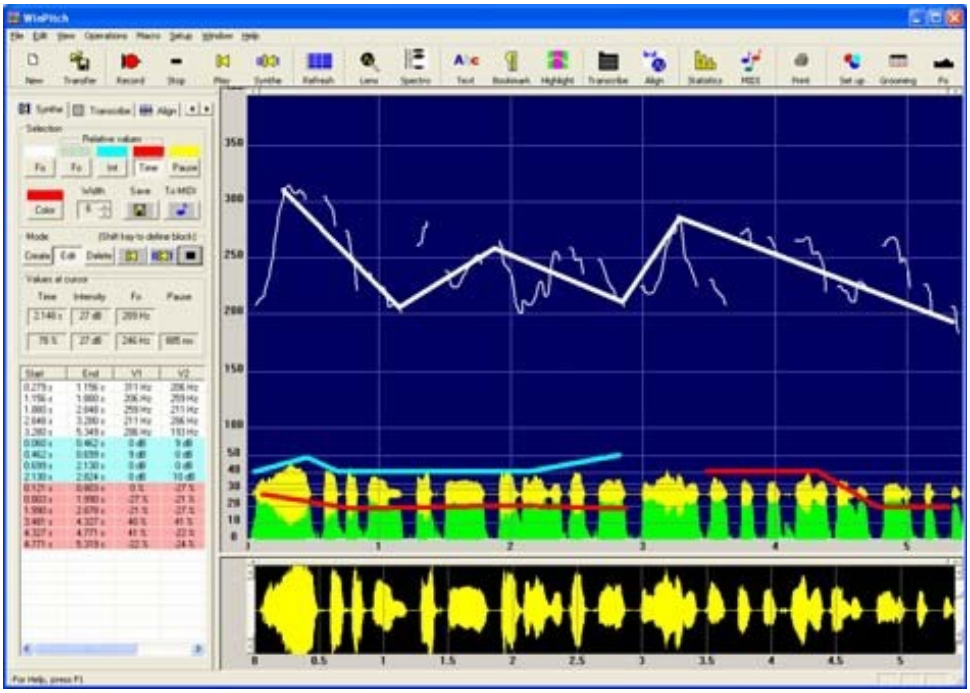


Figure 14 : Fonctionnalités prosodiques du logiciel WINPITCHPRO

Concernant l'alignement texte / signal, qui est ici le point central de notre analyse, WINPITCHPRO présente en théorie deux gros avantages : la gestion de l'audio et de la vidéo sous de nombreux formats ainsi que la compatibilité avec les fichiers de TRANSCRIBER (.trs) et PRAAT (.textgrid), très intéressante pour des échanges de données. Cependant, dans les faits, la réalité est plus nuancée. Au sujet des fichiers audio, WINPITCHPRO gère en effet sans souci les formats .wav, .mp3 et .wma, qui sont probablement les trois plus répandus à l'heure actuelle **;** signalons par ailleurs qu'aucun autre outil testé dans cette étude ne supporte le .wma. Bien sûr, des outils de conversion gratuits existent pour pallier ce problème, mais cela sous-entend toujours une étape supplémentaire dans le processus et donc une perte de temps, aussi brève fût-elle. Concernant la vidéo, les fichiers .avi, .mpg et .mp4 n'ont pas posé de souci pour être ouverts, mais les .mov (format natif de Quicktime) sont en revanche traités comme des fichiers audio, sans possibilité d'obtenir le flux d'images correspondants. Là encore, la conversion peut résoudre ce problème, mais avec les mêmes réserves que celles évoquées précédemment, auxquelles il faut ajouter un possible décalage entre le son et l'image.

A vrai dire, le seul vrai problème que nous ayons rencontré concernant les vidéos est leur vitesse de traitement. Plutôt à l'aise sur les petits fichiers, WINPITCHPRO est beaucoup

plus laborieux sur ceux avoisinant les 30 minutes, et refuse de charger les vidéos dépassant les trois quarts d'heure en indiquant un problème de mémoire insuffisante (nos tests ont été réalisés avec des fichiers .mp4 basse résolution encodés en 320x240, sous Windows XP avec 1 go de RAM). Il faudra donc être attentif si l'on souhaite travailler sur une séance de travail en milieu scolaire par exemple, qui peut parfois excéder la demi-heure.

Nous l'avons dit, WINPITCHPRO propose en formats d'entrée les extensions .trs et .textgrid. Si la seconde ne pose comme seul souci majeur que le fait de perdre le nom des locuteurs, la première connaît de réels problèmes de traitement. Toutes les informations concernant les locuteurs (leurs noms, mais aussi le sexe, la langue maternelle, la qualité du signal, etc.) disparaissent, ainsi que celles à propos des données transcrites. Par ailleurs, les balises ne sont nullement prises en compte et, élément beaucoup plus gênant, des bouts de texte sont purement et simplement éliminés, sans raison apparente. Dans le cas d'une transcription initiale relativement précise et annotée, la différence est considérable et il est en conséquence difficile d'évoquer le terme « compatibilité » ici. L'alignement texte / signal en lui-même est certes conservé, mais le contenu des segments diffère trop largement d'un logiciel à l'autre.

Mais WINPITCHPRO propose bien sûr de réaliser ses propres transcriptions, dans un format spécifique (.wp2). La saisie du texte est agrémentée de possibilités très intéressantes, dont l'une offre un confort de travail sans précédent : la gestion de la vitesse du flux audio. Afin de permettre à l'utilisateur de saisir le texte dans un rythme proche du temps réel, il est possible de ralentir le signal à la vitesse souhaitée. Jusqu'à cinquante pour cent de la vitesse initiale, le résultat demeure excellent en termes de restitution des voix : aucun mot n'est déformé, et tout reste parfaitement audible. Il s'agit là d'une fonctionnalité particulièrement performante, qui autorise, dans de nombreux cas, un gain de temps considérable, et qui permet en outre de mieux percevoir les segments bruités (parole superposée, bruits de fond...). Précisons, mais c'est plus anecdotique dans le cadre de notre étude, qu'il est également possible d'accélérer le signal audio et que là encore, la qualité demeure très bonne pour peu que l'on reste dans des limites raisonnables (en dessous de cent cinquante pour cent de la vitesse initiale).

Au niveau de la transcription proprement dite, WINPITCHPRO se veut également novateur, puisqu'il met à disposition de l'utilisateur toutes les polices Unicode existantes, ainsi que de nombreuses fonctions d'un traitement de texte classique (caractères gras, italiques ou

soulignés ; taille de la police ; possibilité de créer une liste des symboles les plus utilisés...). D'autre part, il permet d'aligner facilement une transcription déjà existante (sous réserve qu'elle soit dans un format compatible) avec un fichier son, en proposant là encore une lecture ralentie de celui-ci, afin de permettre un alignement en temps réel. Ce processus permet d'offrir une seconde jeunesse aux données transcrites il y a quelques années, quand les logiciels d'alignement n'existaient pas encore.

Néanmoins, il est dommage que la gestion des locuteurs soit finalement apparentée à une gestion des tires, en ce sens qu'il n'est même pas possible de leur donner un nom. De même, se présentant sous la même forme que celles de PRAAT, il est très difficile d'organiser une transcription avec de nombreux locuteurs, dans la mesure où l'ensemble devient difficilement lisible : le moindre propos d'un nouvel intervenant, même s'il ne parle plus jamais par la suite, fait l'objet d'une tire supplémentaire. On imagine alors ce que peuvent donner des transcriptions longues de plusieurs heures...

Enfin, les fichiers de transcription ainsi générés peuvent être sauvegardés aux formats XML ou EXCEL, ce qui assure une bonne interopérabilité des données. On ne pourra que regretter, dès lors, que WINPITCHPRO soit payant (100 $) et que ses sources ne soient pas librement diffusées. La pérennisation du logiciel n'est en conséquence nullement assurée, puisque personne sinon l'auteur ne peut en (re)prendre le développement. On notera à ce propos que la dernière version de WINPITCHPRO date désormais de plus de trois ans.

Il est donc dommage qu'un tel logiciel, offrant quelques possibilités inédites, ne soit pas plus entretenu. Il ne faudrait sans doute que peu de modifications pour le rendre réellement attractif, les deux premières étant certainement le passage du côté des logiciels libres et une distribution qui ne soit pas limitée à Windows. Cependant, son interface proche de celle de PRAAT et sa gestion limitée des vidéos l'oriente plutôt vers une utilisation similaire, et non vers de la transcription ou de l'annotation de masses de données. Les utilisateurs de TRANSCRIBER auraient pourtant pu y voir une passerelle intéressante au vu de la compatibilité annoncée avec les fichiers .trs, mais celle-ci engendre trop de pertes d'informations pour être réellement intéressante.

3.4 EXMARaLDA

Page Web : http://www.exmaralda.org/
Auteur(s) : Thomas Schmidt, Kai Wörner
Date de création du logiciel : 2001
Dernière mise à jour du logiciel : 29 novembre 2010
Plates-formes : Windows, Unix, Macintosh
Statut : logiciel libre
Langue de l'interface : allemand, anglais, espagnol, français, suédois, tchèque, turc
Format natif : .xml
Autres formats supportés : .textgrid (Praat), .eaf (Elan), .xml (TEI, graphes d'annotation), .txt, .dat (fichier HIAT-DOS)
Formats d'entrée audio : .aif, .aiff, .mp3, .wav
Formats d'entrée vidéo : .avi, .divx, .mp4, .mov, .mpg, .wmv
Formats d'exportation : .textgrid (Praat), .eaf (Elan), .xml (TEI, graphes d'annotation, fichiers TASX), .cha
Publications : [Schmidt, 2009], [Schmidt, 2010]

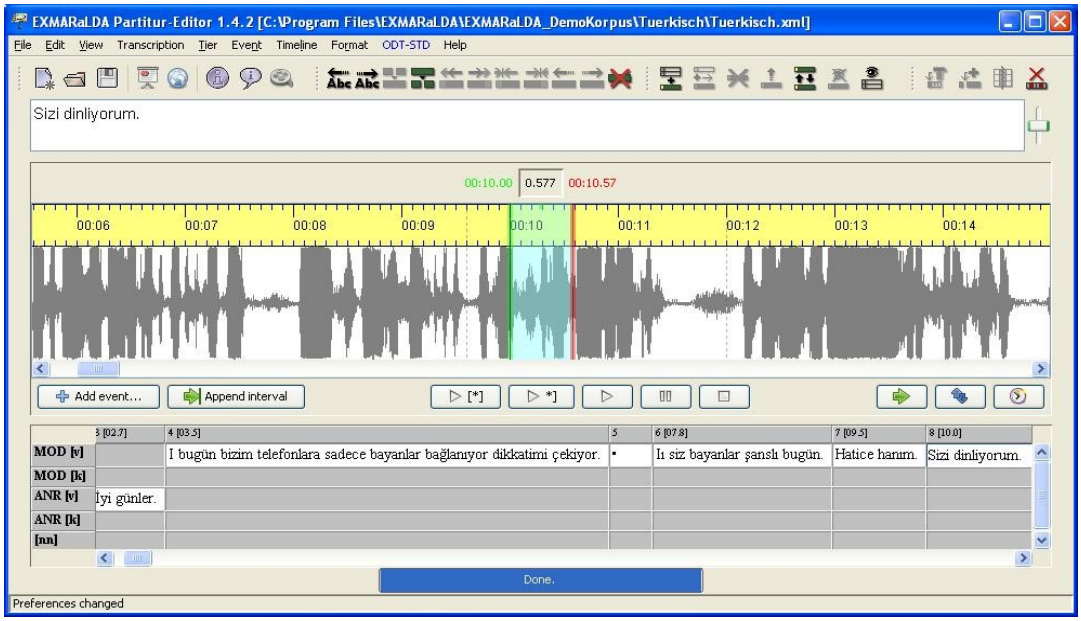


Figure 15 : Interface générale du module PARTITUR-EDITOR

EXMARaLDA se différencie de tous les autres logiciels présentés ici, dans la mesure où il est en fait un système de transcription sub-divisé en trois outils :

- COMA, qui permet de gérer l'archivage des corpus ;
- EXAKT, sorte d'explorateur qui permet d'effectuer des recherches lexicales ;
- PARTITUR-EDITOR, qui concerne l'alignement et la transcription à proprement parler.

C'est à ce dernier que nous allons ici nous intéresser mais par commodité, nous l'appellerons par son nom générique : EXMARaLDA. Il permet donc d'aligner et/ou de transcrire des fichiers audio ou vidéo avec des données textuelles. L'une des particularités revendiquées par son créateur est que EXMARaLDA se présente sous la forme d'une partition musicale, avec différentes « portées » qui sont en fait des tires correspondant soit à des locuteurs, soit à des annotations spécifiques (bruits de fond, prononciations particulières...). En cela, ce logiciel ne se différencie pas vraiment d'ANVIL, de WINPITCHPRO ou de PRAAT. En effet, la disposition des strates est identique et, s'il est possible d'en ajouter à l'infini, cela pose les mêmes problèmes que pour les logiciels précités. L'affichage est surchargé au-delà d'un certain nombre de locuteurs, et devient ainsi peu lisible. Par exemple, nous avons été confrontés dans le corpus EPAC à des fichiers contenant 35 locuteurs différents. Voici ce que donnerait la représentation de l'un d'entre eux avec EXMARaLDA, simplement au niveau de la mise en place des tires :

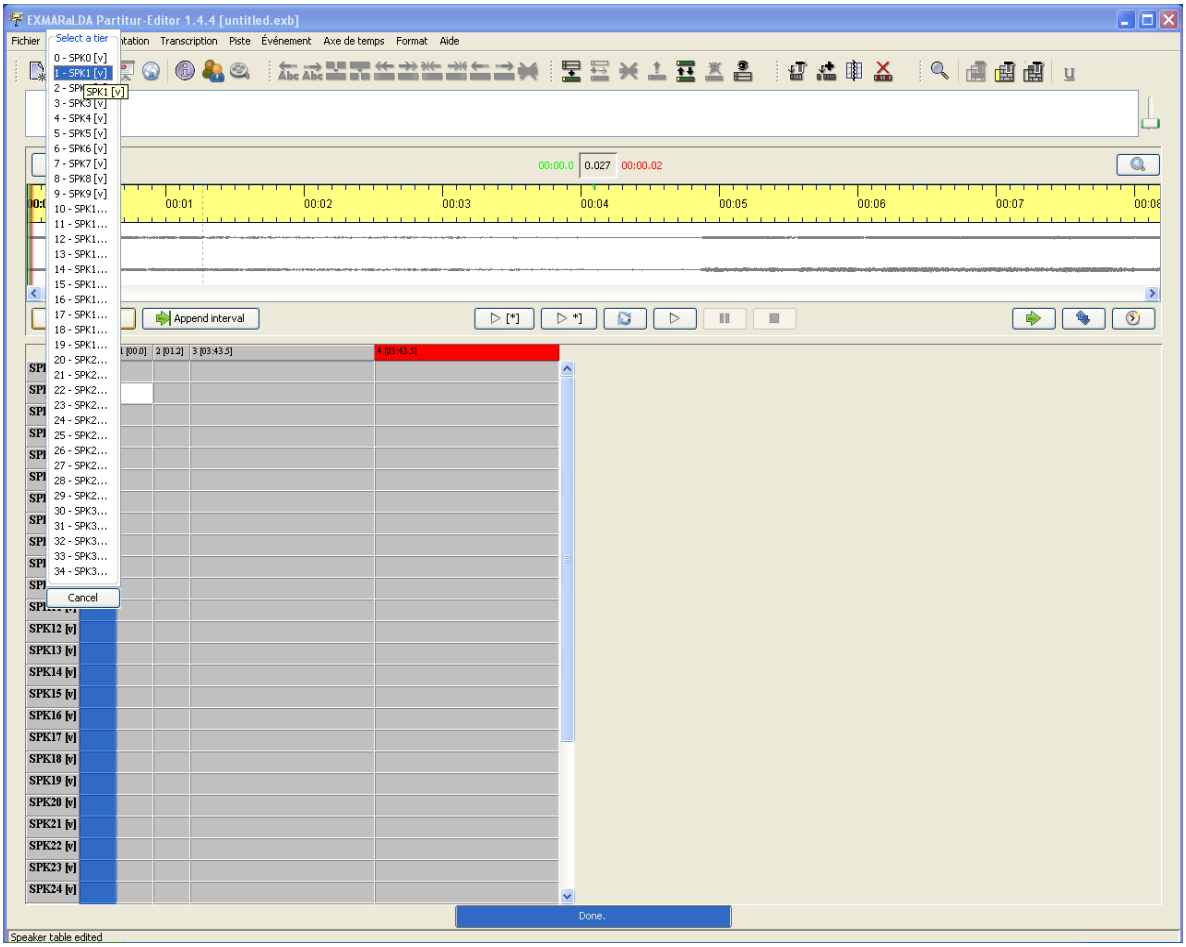


Figure 16 : Représentation des locuteurs avec le module PARTITUR-EDITOR

La figure 16 montre bien les difficultés relatives à ce genre de représentation verticale. Il est techniquement impossible de faire apparaître tous les locuteurs à l'écran. Il faudra donc régulièrement utiliser la barre de navigation pour aller de l'un à l'autre. Ce choix est d'autant plus discutable que, comme on le voit ici, EXMARaLDA impose une étape supplémentaire pour choisir le locuteur à qui on souhaite attribuer un tour de parole. Outre une perte de temps significative si l'on ramène ce procédé à l'échelle d'un fichier d'une heure, on s'aperçoit que le logiciel n'affiche que les premiers caractères de chaque nom. Lorsque plusieurs locuteurs ont le même prénom (ce qui est relativement fréquent), il faut mettre en surbrillance chacun d'entre eux pour faire apparaître les patronymes complets et s'assurer que le choix est le bon. Malgré cela, les possibilités d'annotation concernant les locuteurs sont ici très complètes, encore plus qu'avec TRANSCRIBER, puisqu'il est possible de créer sa propre nomenclature, ce qui permet d'établir de véritables bases de données socio-linguistiques.

Les mêmes remarques peuvent être faites au sujet du texte transcrit. Des possibilités

spécifiques sont offertes, notamment celle qui permet d'insérer en un clic n'importe quel symbole de l'API, et pas seulement concernant les phonèmes :

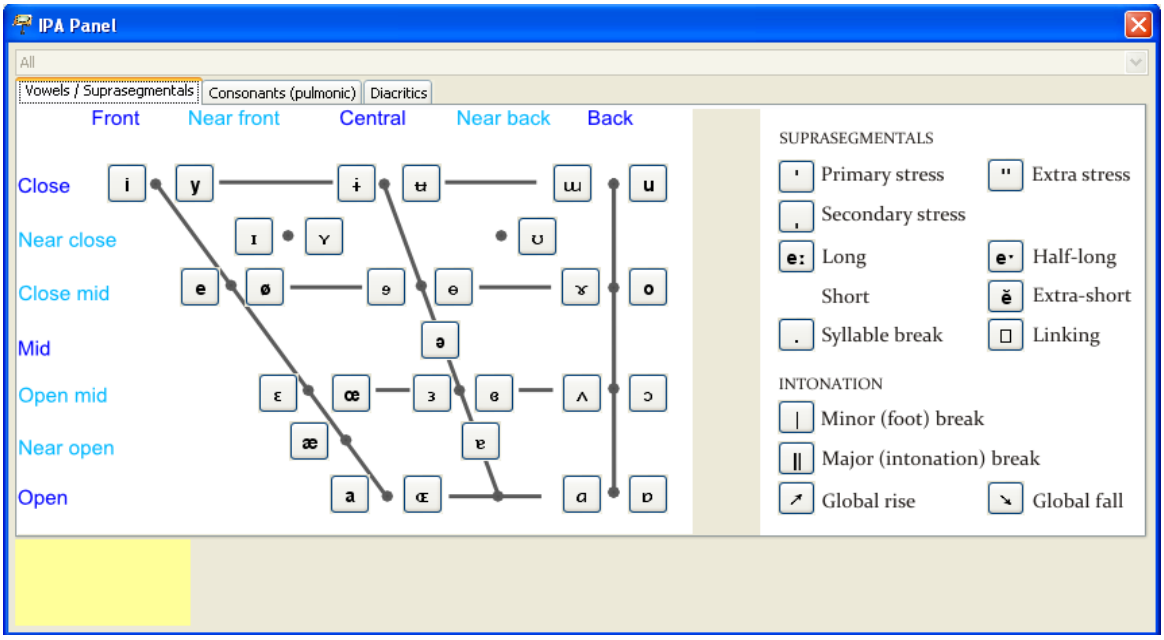


Figure 17 : Gestion de l'API sous EXMARaLDA

De même, la présence d'un clavier virtuel autorise tout type d'insertions dans le corps du texte, et notamment des balises de la TEI relatives à la parole. Cette norme étant encore très peu intégrée aux logiciels de transcription, il est appréciable de la trouver ici facilement exploitable, même si c'est de manière simplifiée.

La segmentation en tours de parole peut se faire en deux temps. On peut tout d'abord en s'appuyer sur la fin d'un segment pour en générer un nouveau à sa suite, ce qui est le mode de segmentation classique de TRANSCRIBER. Cette méthode, nous l'avons dit, est cependant contraignante pour traiter la parole superposée. A ce titre, EXMARaLDA propose donc une autre option qui permet d'aligner librement un segment sur le flux audio, sans tenir compte des autres segments déjà créés, ce qui permet de découper précisément toutes les situations d'énonciation complexes.

Pour ce qui est des formats supportés, EXMARaLDA se montre plutôt capricieux : le .mp3 et le .wav sont par exemple supportés en ce qui concerne l'audio, mais seul ce dernier permet d'afficher le spectrogramme, élément pourtant quasiment indispensable pour avoir un minimum de repères visuels. Cette sélectivité est d'autant plus étrange qu'aucun autre logiciel

ne se montre aussi restrictif à ce niveau. EXMARaLDA supporte également la vidéo sous plusieurs formats mais là encore, il faut ajouter de nombreuses restrictions, à tel point qu'une documentation officielle existe pour les référencer. Cela dépend :

- du système d'exploitation (Windows, Linux, Mac),
- de la carte son et vidéo utilisées,
- des codecs installés,
- du format traité,
- du lecteur vidéo choisi (EXMARaLDA n'en possède pas de spécifique, et en propose trois différents : JMF, DirectShow ou QuickTime)

Le fichier peut être lu parfaitement, ne pas être lu du tout, être lu partiellement (simplement le flux audio) ou avec des problèmes d'affichage, de synchronisation... Pour tenter de pallier ce problème, EXMARaLDA peut en fait être lancé selon deux configurations, qui influent grandement sur le taux de succès de lancement d'un fichier vidéo. Lors de nos tests, cela est presque allé du tout au rien. Gageons que lors d'une prochaine version du logiciel, celui-ci comportera un lecteur intégré comme TRANSANA ou WINPITCHPRO, qui offrent ainsi une compatibilité maximale à ce niveau.

Cependant, s'il est un domaine dans lequel EXMARaLDA se démarque, c'est sans nul doute dans celui de l'interopérabilité. D'une part ce logiciel s'appuie sur XML, langage synonyme de pérennité et d'échange simplifié d'une plate-forme à l'autre, puisqu'il hiérarchise avec précision les données et méta-données d'un document. D'autre part, EXMARaLDA supporte en entrée les formats natifs de PRAAT (.textgrid) et d'ELAN (.eaf), et peut également exporter des transcriptions vers ces mêmes formats, ainsi que vers les standards de CLAN et de la TEI. S'il est dommage que TRANSCRIBER, pourtant très utilisé, ne vienne pas s'ajouter à cette liste (pour l'instant), l'initiative est d'autant plus appréciable qu'elle n'est pas si fréquente à un tel niveau. Cela assure à EXMARaLDA, outil encore relativement « jeune », une utilisation de plus en plus répandue, d'autant qu'il est totalement gratuit et que son site et sa documentation sont actuellement trilingues (allemand, anglais et français).

3.5 ELAN

Page Web : http://www.lat-mpi.eu/tools/elan/
Auteur(s) : Birgit Hellwig, Diter van Uytvanck, Micha Hulsbosch
Date de création du logiciel : février 2002 (version 1.1)
Dernière mise à jour du logiciel : 27 décembre 2010
Plates-formes : Windows, Unix, Macintosh
Statut : logiciel libre
Langue de l'interface : catalan, anglais, allemand, espagnol, français, chinois, suédois, portugais, néerlandais
Format natif : .eaf
Autres formats supportés : fichiers Shoebox, fichiers Flex, chat, .trs, .csv, .textgrid
Formats d'entrée audio : .wav
Formats d'entrée vidéo : .mov, .mpeg, mpeg-4, .mp4, .mpg, .qt
Formats d'exportation : Shoebox, toolbox, chat, textgrid, tiger-xml, html, texte quicktime, SMIL...
Publications : [Hellwig *et al.*, 2009], [Berez, 2007]

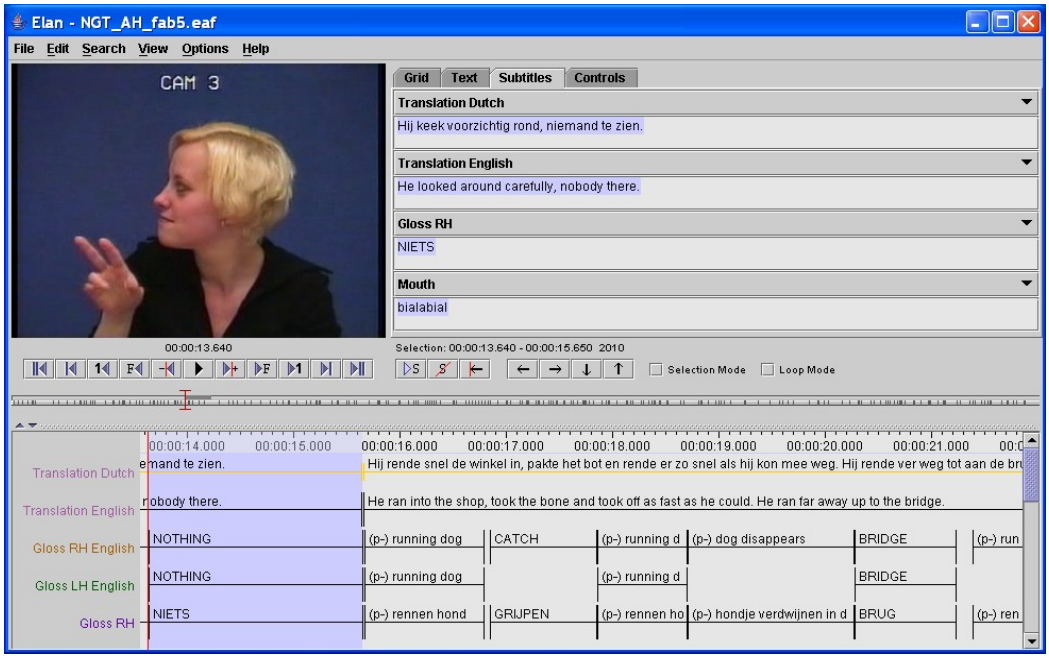


Figure 18 : Interface générale du logiciel ELAN

ELAN fait également partie de ces logiciels qui, en plus des fichiers audio, sont capables de gérer la vidéo. Théoriquement, une grande compatibilité est offerte à ce niveau (mov., .mpeg, .mp4, .qt...) mais dans le cas d'annotation vidéo multi-angles, le programme a énormément de mal à gérer certains formats (le mp4 notamment), et rend presque impossible toute tentative de transcription précise. Qui plus est, au-delà de deux vidéos synchronisées, l'affichage en privilégie une et une seule, affichant les autres dans des fenêtres assez petites qui en réduisent la lisibilité (voir figure 19).

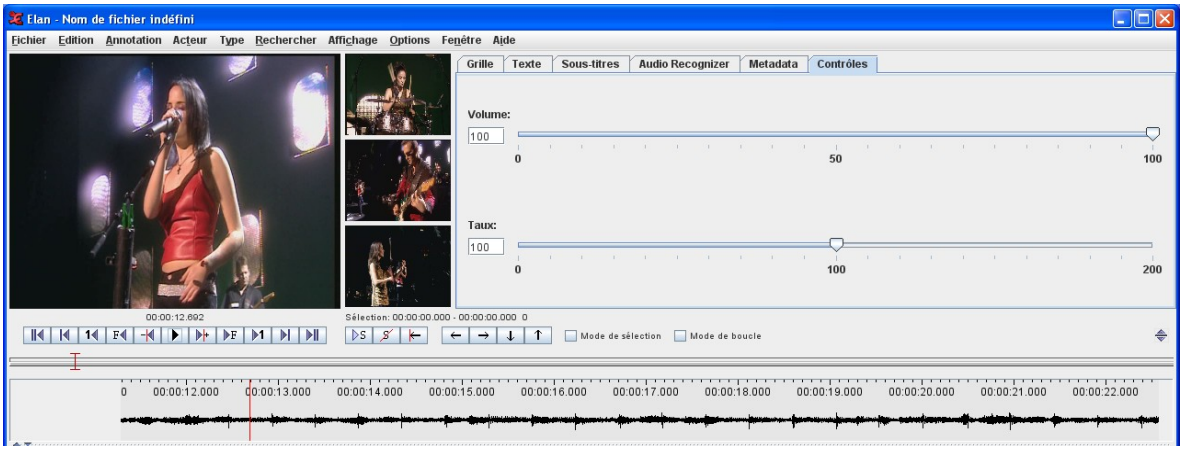


Figure 19 : Gestion de la transcription multi-vidéos sous ELAN

Au niveau des formats audio, ELAN est plutôt sélectif, puisqu'il n'accepte que le .wav. Précisons qu'à la différence de TRANSANA notamment, transcrire une vidéo sous ELAN signifie lui fournir en entrée le fichier vidéo ET le fichier audio correspondant. Il faut donc penser à le générer au préalable si l'on souhaite voir apparaître le spectrogramme dans l'interface. De son côté, TRANSANA s'en charge automatiquement.

Pour ce qui est de la transcription proprement dite, ELAN s'articule autour d'une interface à défilement horizontal, qui n'est pas sans rappeler celle de PRAAT, d'autant que le principe des tires est également identique. Il est possible d'en créer une quantité illimité, ce qui offre de vastes possibilités d'annotation. La gestion de l'alignement temporel est très modulable, puisqu'il n'y a aucune restriction quant à la disposition des ancres qui indiquent les limites des segments ou des tours de parole. Néanmoins, l'utilisateur ne peut pas s'appuyer sur la fin d'un segment aligné pour précisément aligner le suivant. C'est-à-dire qu'à chaque nouveau segment, il convient de délimiter son terme (ce qui est somme toute logique), mais également son commencement. Il faut pour y parvenir faire soigneusement glisser le

marqueur temporel, jusqu'à ce qu'il ne fasse plus qu'un avec le précédent. Loin d'être évidente, cette démarche est relativement coûteuse en temps, bien qu'elle puisse être utile lorsque les interventions des locuteurs sont entrecoupées de pauses significatives. Par ailleurs, elle permet de gérer de façon totalement indépendante les segments superposés, même si les locuteurs concernés sont en nombre important.

Pour ce qui est de la transcription elle-même, celle-ci se présente sous la forme d'une zone de texte apparaissant une fois que l'on a délimité le segment temporel correspondant, et moyennant un double-clic à l'intérieur de celui-ci. Une fois le texte saisi, il faut le valider avec la combinaison CTRL-ENTREE, et il apparaît alors au-dessus du segment (voir figure 18). Aucune balise ou possibilité d'annotation particulière n'est prévue dans ce cadre, mais au même titre que PRAAT, la possibilité d'ajouter de nouvelles strates d'annotation permet finalement une granularité considérable, pour peu que l'on parvienne à maîtriser l'utilisation de l'alignement temporel.

L'un des principaux atouts d'un logiciel comme ELAN est sa gestion des formats, que ce soit pour l'importation et l'exportation. Il est notamment possible d'y ouvrir des fichiers issus de TRANSCRIBER, PRAAT, CHAT, et également d'exporter une transcription en .eaf (le format natif d'ELAN) vers la plupart de ces logiciels, ainsi qu'en .html ou en .txt par exemple. Ajoutée au fait qu'ELAN soit un outil libre et multi plate-formes, cette plus-value en fait un choix intéressant à bien des égards, même si la synchronisation temporelle demande un vrai temps d'adaptation pour qui est davantage familier d'un autre logiciel.

3.6 TRANSANA

Page Web : http://www.transana.org/
Auteur(s) : Chris Fassnacht, David K. Woods
Date de création du logiciel : 5 octobre 2001
Dernière mise à jour du logiciel : 2 septembre 2010
Plates-formes : Windows, Macintosh
Statut : logiciel payant (50 $) ; version d'essai gratuite (possibilités bridées)
Langue de l'interface : anglais, allemand, chinois, danois, espagnol, français, italien, allemand, norvégien, russe, suédois
Format natif : .xml
Autres formats supportés : .rtf, .txt
Formats d'entrée audio : .mp3, .wav
Formats d'entrée vidéo : .avi, .mov, .wmv, .mpeg-1, .mpeg-2, .mp4.
Formats d'exportation : .xml
Publications : [Afitska, 2009], [Silver et Lewins, 2009]

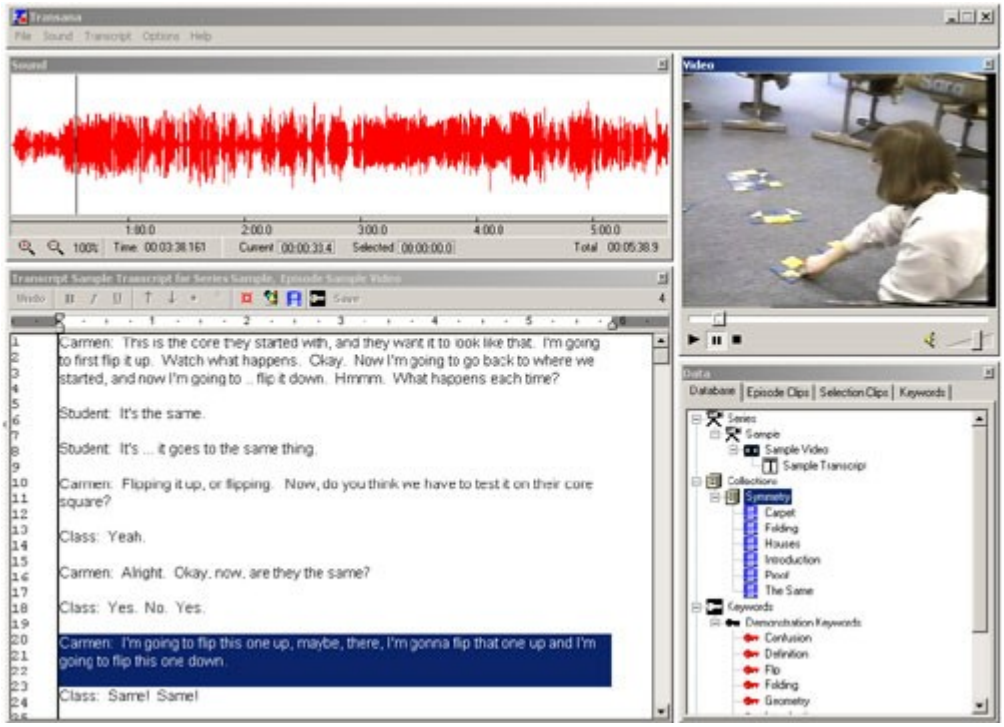


Figure 20 : Interface générale du logiciel TRANSANA

TRANSANA se démarque des autres logiciels que nous avons testés jusqu'à présent pour une raison majeure : sa version complète est payante. Plus précisément, une version d'évaluation est accessible gratuitement mais avec des fonctionnalités moindres. Force est de reconnaître que celles-ci ne sont pas des plus pénalisantes (ce sont notamment les possibilités d'archivage qui sont bridées), et qu'elles n'imposent pas, par exemple, une durée d'utilisation limitée ou une taille de fichier réduite. Cependant, sur le fond, ce détail mérite d'être signalé, d'autant que dans ses premières versions - jusqu'en avril 2007 -, TRANSANA était totalement gratuit. Mais étrangement, une certaine ambiguïté demeure sur le site internet du logiciel. La page d'accueil semble n'avoir jamais été mise à jour depuis ce changement, et il y est toujours écrit (au 10 décembre 2009) : *Transana is inexpensive and Open Source*. Et ce n'est pas le moindre des paradoxes, puisque toutes les anciennes versions gratuites du logiciel ont disparu de la section « Download »...

Néanmoins, TRANSANA est un outil très intéressant sous bien des aspects, au premier rang desquels se situe sa capacité à gérer la vidéo. À ce titre, il ne se contente pas d'un service minimum puisque l'immense majorité des formats existants est supportée (.avi, .mov, .mpg, .mp4, .wmv...). Il est donc possible d'envisager avec ce type de logiciel un genre d'annotation beaucoup plus complet que celles qui s'appuient simplement sur l'audio, pour peu que les données que l'on souhaite exploiter le demandent.[21]

Qui plus est, il faut préciser que la toute dernière version de TRANSANA (2.40) offre une possibilité similaire à celle de ELAN, à savoir le traitement de la vidéo multi-angles. Ainsi, si une scène a été filmée de façon synchronisée par plusieurs caméras et avec différents points de vue, le logiciel est capable d'afficher simultanément les fichiers vidéos correspondants (jusqu'à 4), offrant à l'annotateur une matière première d'une grande richesse. Nous avons expérimenté cette nouvelle fonctionnalité, en extrayant d'un DVD du groupe *The Corrs* un extrait de concert filmé sous plusieurs angles (un par membre du groupe). Voici ce que donne alors l'interface de TRANSANA, pour peu que l'on ajuste les différentes fenêtres (il faut pour cela désactiver préalablement leur auto-ajustement en se rendant dans le menu « Options ») :

21 Des utilisateurs nous ont toutefois indiqué que des problèmes de précision demeuraient lorsque l'on souhaitait interrompre le lecteur vidéo à un endroit précis. Nous n'avons pour notre part pas constaté ce souci.

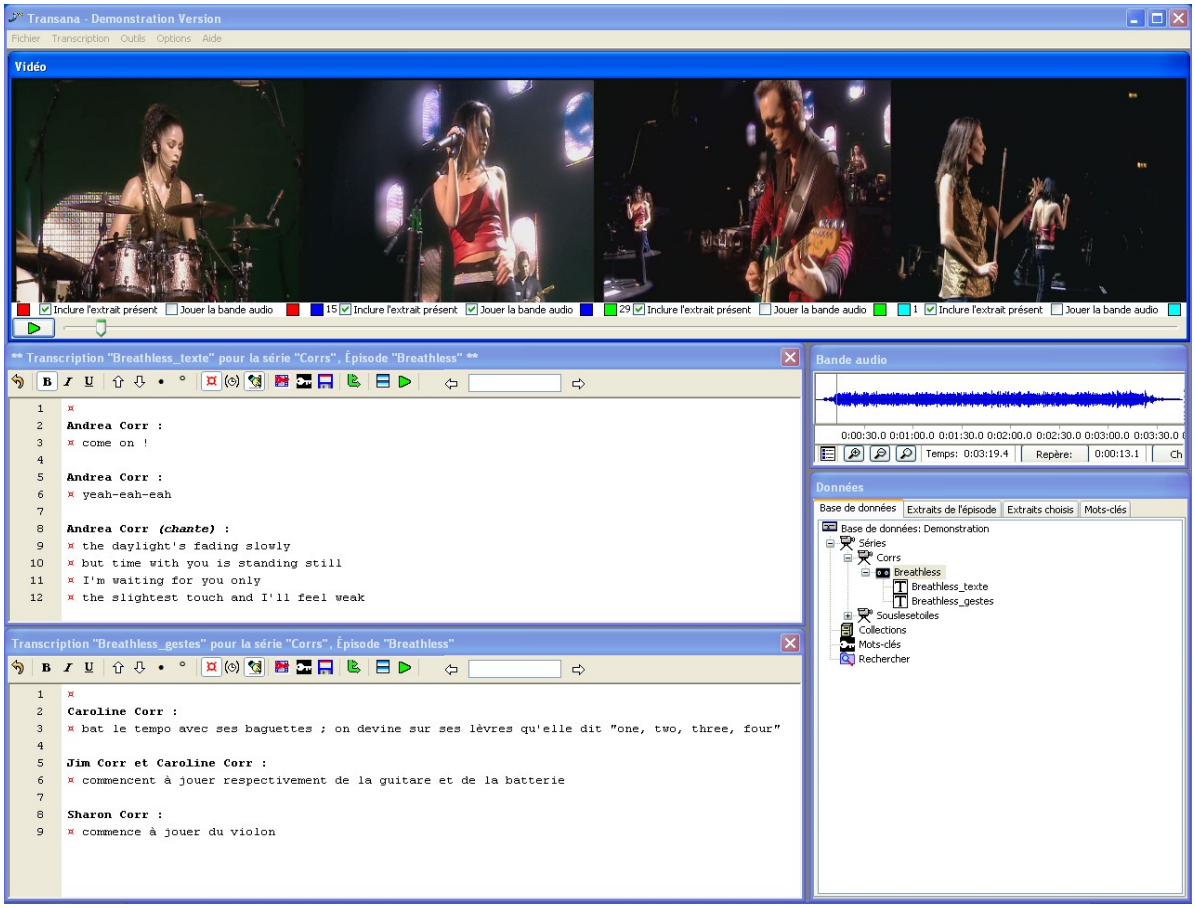


Figure 21 : Gestion de la transcription multi-vidéos sous TRANSANA

Comme on le voit, il est possible d'obtenir un ensemble tout à fait lisible, avec les quatre vidéos synchronisées dans la partie supérieure, le signal audio et la ou les fenêtres de transcription. Pour notre exemple, nous en avons utilisé deux : l'une pour la transcription textuelle « classique », en l'occurrence les paroles du titre interprété, et l'autre pour indiquer les principaux mouvements des musiciens. Il est toutefois possible de multiplier ces couches d'annotation, et même de permettre à plusieurs annotateurs de travailler sur un même fichier. Quelques options sont également disponibles concernant les vidéos en elles-mêmes (notons entres autres la possibilité de supprimer les flux sonores de chacune d'entre elles, bien utile pour éviter un écho désagréable lorsque ces flux sont strictement identiques).

Bien sûr, une telle granularité d'annotation demandera un temps précieux et surtout des données disponibles (les corpus vidéo multi-angles sont encore très marginaux à l'heure actuelle), mais le fait d'avoir dès à présent un outil à disposition est une excellente nouvelle en vue de projets futurs.

Bien que plus complexe que TRANSCRIBER, TRANSANA se prend assez facilement en main, et il est possible d'être rapidement en mesure de maîtriser son fonctionnement de base. La différence principale se situe dans la gestion des fichiers audio ou vidéo. Dès que l'utilisateur souhaite en utiliser un, il lui est demandé de le classer dans une *série*, en tant qu'*épisode*. Ce processus de classification, qui peut sembler de prime abord fastidieux, est en fait très pratique pour archiver ses transcriptions et retrouver facilement l'une d'entre elles au sein d'un corpus volumineux. Par ailleurs, si plusieurs personnes travaillent sur un même projet de transcription, il est ainsi possible de bien classer et étiqueter le travail de chacun.

Une fois ces opérations liminaires effectuées, TRANSANA va se présenter sous la forme d'une interface divisée en quatre fenêtres (voir figure 20) :

- le spectrogramme en haut à gauche, généré si besoin à partir du fichier audio ou vidéo ;
- la vidéo elle-même à côté (si un fichier vidéo est utilisé, il sera synchronisé automatiquement avec le spectrogramme lui faisant face, puisque c'est TRANSANA lui-même qui aura pris soin de le générer) ;
- la zone de transcription dans la partie inférieure gauche ;
- une arborescence de classification avec les séries et épisodes, les collections, les mots-clés, etc.

Pour évoquer la phase de transcription en elle-même, il faut commencer par préciser que la fenêtre consacrée à cette tâche est un peu particulière. Par défaut, elle est totalement désynchronisée du signal audio et/ou vidéo. Par exemple, le fait d'aller à la ligne ne créera pas temporellement un nouveau segment de parole, pas plus que le fait d'ajouter un locuteur ne générera automatiquement un nouveau tour de parole (comme dans TRANSCRIBER notamment). L'utilisation de la touche ENTREE est ainsi rigoureusement identique à celle qu'elle a dans un traitement de texte classique, ce qui est déroutant pour les utilisateurs de TRANSCRIBER ou de PRAAT. Mentionnons à ce titre la présence des icônes qui permettent l'usage de caractères gras, italiques ou soulignés.

Cette esthétique rappelle finalement celle des transcriptions saisies avec l'aide d'un support informatique, mais sans possibilité d'alignement avec la source audio. Un tel choix a en quelque sorte les avantages de ses inconvénients. Il est possible d'obtenir à l'écran un

résultat très proche d'un texte ordinaire dactylographié, en terme de lisibilité et de présentation : on peut ajouter toutes les informations que l'on veut, sous la forme que l'on souhaite, sans jamais perturber la synchronisation temporelle, qui demeure une étape indépendante à part entière. La figure 22 donnera une idée de la présentation qu'il est possible de donner à une transcription :

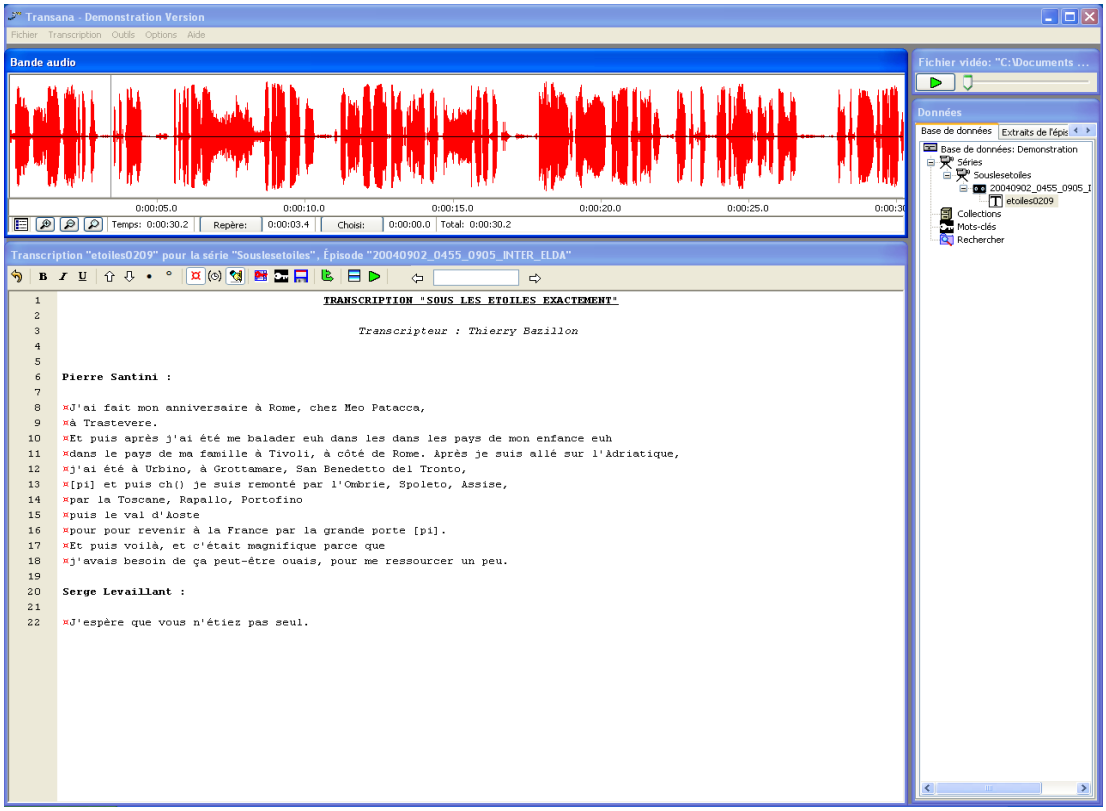


Figure 22 : Exemple de transcription réalisée avec TRANSANA

Un tel résultat est impossible à obtenir avec les autres logiciels présentés jusqu'à maintenant. Cela se fait toutefois au détriment de quelques fonctionnalités moins intuitives, dont la plus importante d'entre elles est l'alignement temporel. Là où la grande majorité des outils concurrents se contentent d'une pression sur une touche, qui fait apparaître de façon explicite la segmentation, TRANSANA propose, *via* la combinaison CTRL-T, d'insérer un petit repère temporel de couleur rouge (voir figure 22). Outre le fait de solliciter deux touches et de ne pas faire ressortir de façon très nette la segmentation (sauf si un retour à la ligne est ajouté manuellement à chaque ancrage, comme dans la figure ci-dessus), ce balisage n'est surtout visible QUE dans la fenêtre de transcription. Ce processus se révèle très perturbant

pour qui a l'habitude de beaucoup utiliser la représentation du spectrogramme. Ce dernier, en dehors de son curseur, ne fait nullement apparaître l'alignement temporel, tout en demeurant par ailleurs difficilement exploitable. Il est notamment impossible de réduire son échelle, si bien qu'il représente soit la totalité de la durée de fichier (ce qui le rend illisible pour les fichiers de plusieurs heures), soit une durée « zoomée » par l'utilisateur, comme c'est le cas ici. Néanmoins, lorsque le zoom est appliqué, il n'est pas dynamique : une fois le curseur arrivé au bout de l'échelle choisie (dans notre exemple, 30 secondes), il continue à avancer au-delà des limites de la fenêtre, sans que l'on puisse voir sa progression. Il faut alors dézoomer puis zoomer à nouveau sur les trente secondes suivantes, ce qui est peu confortable et constitue une réelle perte de temps.

Un point de repère précieux manque donc ici au transcripteur avec, pour conséquence, qu'il est impossible de réaligner manuellement les marqueurs temporels sans les supprimer. Dans une interface comme celle de TRANSCRIBER ou PRRAT, l'utilisateur a simplement à faire « glisser » les délimitations afin de les disposer à sa convenance. Sous TRANSANA, il faut soit être très rigoureux lors du premier ancrage temporel (ce qui loin d'être évident, notamment avec de la parole superposée), soit supprimer le marqueur erroné puis tâtonner une ou plusieurs fois pour qu'il se trouve finalement à l'endroit voulu. Toutefois, un utilisateur confirmé doit pouvoir s'affranchir de ces contraintes et ne perdre que peu de temps, mais avec des données relativement « propres ». Il nous semble qu'une telle interface rend par exemple très difficile la retranscription de débats particulièrement animés, où il faut écouter plusieurs fois de nombreux segments pour distinguer et séparer précisément les tours de parole de chacun.

Cependant, pour le traitement de données vidéo où il n'est requis ni un alignement trop précis ni un alignement trop régulier, TRANSANA semble tout à fait approprié. Il offre à ce titre quelques possibilités de balisage proches de celles de TRANSCRIBER, pour des phénomènes tels que les respirations ou les soupirs, et également dans le domaine prosodique (intonation), ce qui est plutôt rare - à l'exception notable de PRAAT. De même, il est possible d'assigner à certains mots un statut de « mot-clé », permettant par la suite d'effectuer facilement des recherches lexicales. Comme le fait apparaître la figure 22, il n'y a en revanche pas d'interface spécifique aux locuteurs : leurs noms doivent être réécrits à chaque tour de parole, de même que chaque information que l'on souhaiterait leur attribuer (âge, sexe, langue maternelle, canal utilisé...).

Enfin, nous terminerons cette analyse en précisant que TRANSANA propose une interface multilingue particulièrement avancée (plus de dix langues sont proposées, dont le chinois ajouté tout récemment) et une documentation animée en plusieurs volets très efficace, disponible sur le site web du logiciel. Le programme est mis à jour plusieurs fois par an, avec des détails très précis sur les évolutions apportées à chaque version. Si son format de sortie, en .xml, offre aux données transcrites l'assurance de pouvoir être exploitées aisément, une portabilité vers les systèmes LINUX permettrait sans doute à TRANSANA de disposer d'une communauté d'utilisateurs encore plus importante. Cependant, son système de segmentation et de transcription assez marginaux semblent le destiner aux corpus vidéos, domaine ou TRANSANA offre de nombreuses options, dont une représentation à plusieurs flux très bien pensée pour le multi-angles notamment. Cependant, s'il n'est question que de données audio, TRANSANA aura sans doute du mal à convaincre face à l'ergonomie de TRANSCRIBER.

3.7 ANVIL

Page Web : http://www.anvil-software.de/
Auteur(s) : Michael Kipp, Quan Nguyen
Date de création du logiciel : 2000
Dernière mise à jour du logiciel : 21 mars 2011
Plates-formes : Windows, Macintosh, Linux, Solaris
Statut : logiciel libre, mais sur demande à l'auteur
Langue de l'interface : anglais
Format natif : .anvil
Autres formats supportés : Praat (.textgrid, Pitchtier)
Formats d'entrée audio : .wav
Formats d'entrée vidéo : .avi, .mov
Formats d'exportation : .txt, .html
Publications : [Kipp, 2008], [Kipp, 2010]

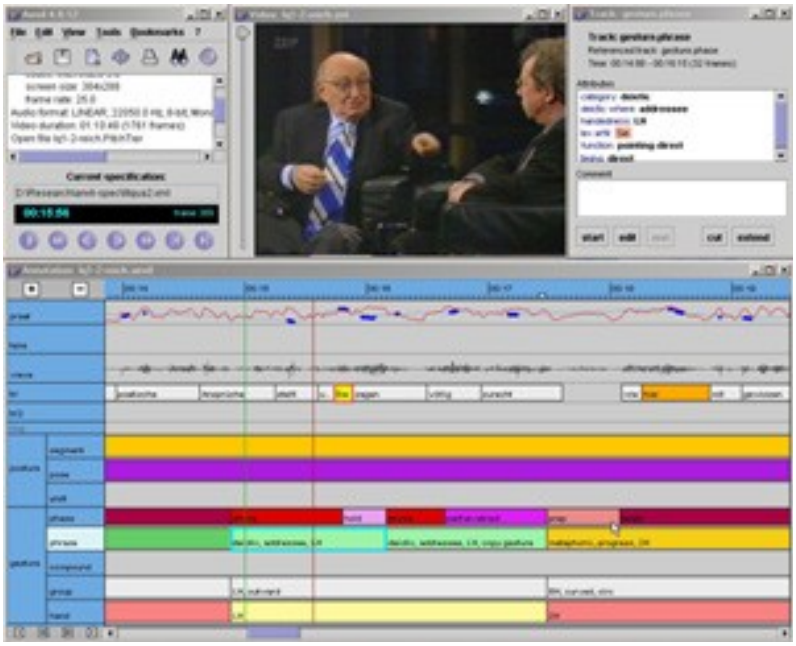


Figure 23 : Interface générale du logiciel ANVIL

ANVIL est un logiciel que l'on acquiert de façon peu commune : il faut, non pas le payer, mais faire une demande par mail à l'auteur qui, en retour (dans un délai plus ou moins rapide), envoie un lien pour en télécharger une copie. Une fois installé, ANVIL se présente sous la forme d'une interface qui évoque celle de ELAN, avec le cadre vidéo au-dessus d'une fenêtre de traitement à déroulement horizontal. La vidéo est ici supportée sous deux formats (.avi et .mov), tandis que du côté audio, il faudra nécessairement utiliser des fichiers .wav. A la différence de TRANSANA ou de ELAN, ANVIL ne gère qu'une seule source vidéo, et ne permet donc pas l'annotation synchronisée avec différents angles de vue. Par ailleurs, les fichiers vidéos de grande taille semblent lui poser quelques soucis de traitement, ce qui le réserve (mais nous aurons l'occasion d'y revenir) à des tâches plutôt spécifiques.

Outre les deux fenêtres préalablement décrites, ANVIL affiche également, dans le coin supérieur gauche, des informations relatives aux formats utilisés (codecs utilisés, résolution, framerate, débit, durée...). De l'autre côté de la fenêtre vidéo se trouvent l'interface, qui permet entre autres de gérer les marqueur temporels, ainsi que l'annotation des données, qui apparaîtront donc au sein des différentes tires générées par l'utilisateur, dans la partie inférieure.

Le terme « annotation » que nous avons employé n'est dans le cas présent pas anodin. Les possibilités dans ce domaine sont avec ANVIL très vastes, et en font davantage une loupe pour s'intéresser aux détails qu'un outil de transcription à proprement parler, tel que l'est TRANSCRIBER par exemple. ANVIL est en cela plutôt à rapprocher de PRAAT.

La gestion des tires par exemple, ici baptisées « tracks » illustre cette caractéristique. Le premier réflexe, lorsque l'on ouvre un logiciel, est d'entreprendre un début de segmentation en tours de parole et en locuteurs. Or, il est impossible d'ajuster le nombre de locuteurs de façon dynamique, pas plus qu'il n'est possible d'ajouter une couche d'annotation de façon « spontanée ». Toute cette nomenclature doit être définie avant, en créant un document XML (appelé DTD) à l'intérieur duquel l'utilisateur spécifiera le nombre de tires consacrées aux locuteurs, l'affichage ou non du spectrogramme, la présence de tires supplémentaires... Il est certes possible de faire ce travail liminaire une fois, et d'utiliser par la suite la même nomenclature pour chaque fichier. Souvent néanmoins les besoins varient au gré des données (surtout le nombre de locuteurs) et des objectifs de l'annotateur. En conséquence, générer un nouveau fichier de spécifications est quasiment indispensable à chaque fois, même si les modifications peuvent y être mineures.

Un tel processus présente toutefois un avantage majeur : il est réellement possible de paramétrer ses annotations comme on le souhaite, qu'il s'agisse d'établir des catégories (par exemple « gestes », comme sur la figure 24) et d'attribuer une sous-catégorie spécifique à chaque locuteur, ou encore de prévoir des strates d'annotation pour des phénomènes non-alignés temporellement (les entités nommées par exemple).

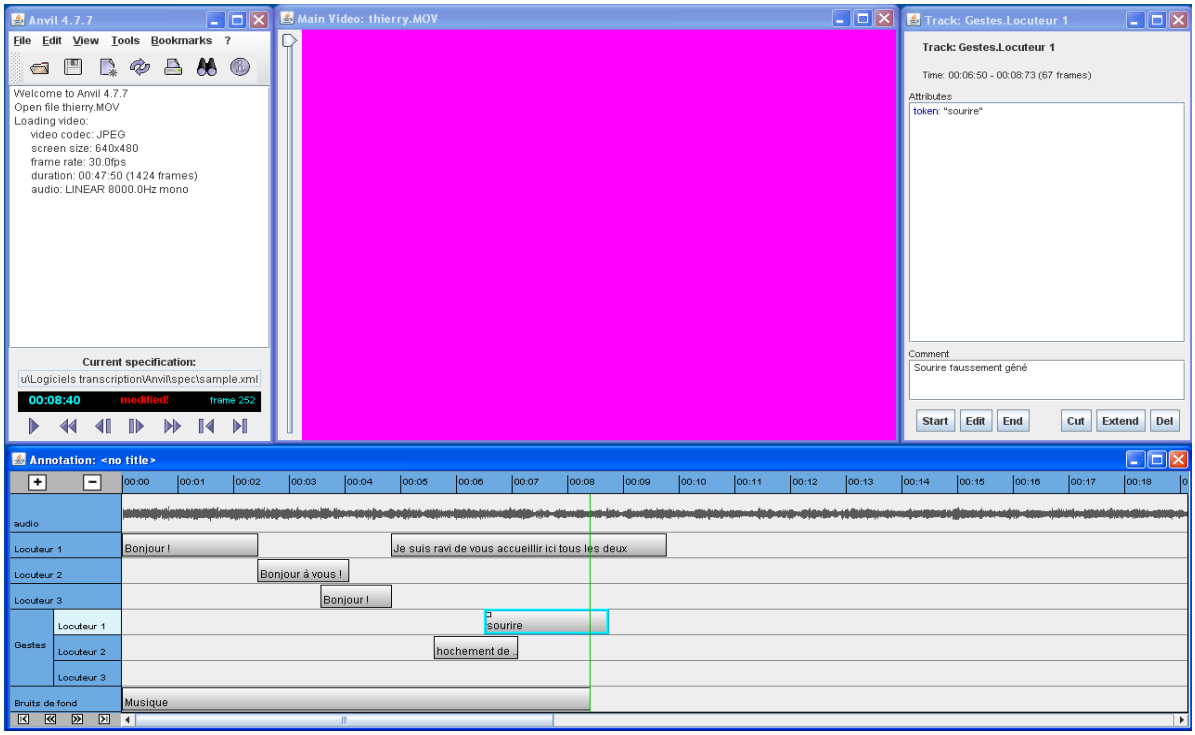


Figure 24 : Représentation des différents niveaux d'annotation sous ANVIL

Il est visible ci-dessus que l'alignement temporel est totalement désynchronisé d'une tire à une autre, et que l'on peut traiter efficacement les phénomènes de superposition (parole, mais aussi gestes, regards...). Notons que si aucune nomenclature prédéfinie n'est disponible (pour indiquer des phénomènes extra-linguistiques par exemple), il est en revanche possible, pour chaque annotation effectuée, d'ajouter un commentaire, ce que nous avons fait dans notre exemple (« sourire faussement gêné »).

Mais, nous l'avons dit, ANVIL nous semble davantage être un logiciel d'annotation que de transcription. Il se rapproche en cela de PRAAT, et il n'est pas étonnant d'y trouver la possibilité d'importer les fichiers natifs de ce dernier (.textgrid ou .pitchtier) pour afficher, dans l'interface même d'ANVIL, des informations relatives à la prosodie. Là encore, cette

option supplémentaire ne sera visible à l'écran que si elle est spécifiée dans le fichier .xml de configuration.

Le statut un peu particulier de la distribution d'ANVIL fait qu'il est difficile de savoir s'il est réellement utilisé, et dans quel(s) cadre(s). Le fait qu'il soit uniquement en anglais est sans doute un frein à sa diffusion, à l'heure ou certains outils font de gros efforts pour être accessible de façon aussi large que possible.

3.8 XTRANS

Page Web : http://www.ldc.upenn.edu/tools/XTrans/
Auteur(s) : Haejoong Lee, Stephanie Strassel
Date de création du logiciel : *information non disponible*
Dernière mise à jour du logiciel : *information non disponible*
Plates-formes : Windows, Linux
Statut : logiciel libre
Langue de l'interface : anglais
Format natif : .tdf
Autres formats supportés : Transcriber (.trs), .sgml
Formats d'entrée audio : .wav, .sph + une dizaine d'autres formats (mais pas de mp3)
Formats d'entrée vidéo : aucun
Formats d'exportation : .trs, .txt, .sgml
Publications : [Lee et Strassel, 2007]

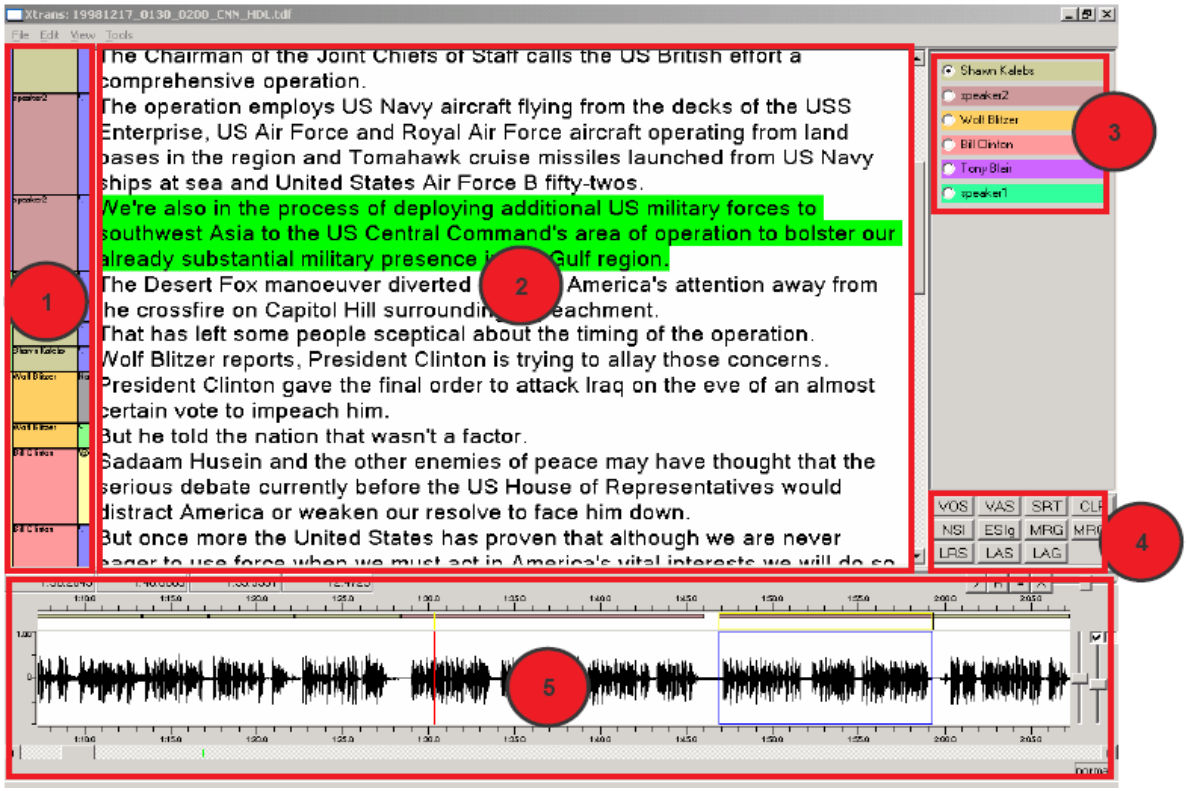


Figure 25 : Interface générale du logiciel XTRANS

XTRANS est un logiciel qui a été développé par le Linguistic Data Consortium (LDC). Le LDC a beaucoup utilisé XTRANS lors de ses propres travaux (plus de 3500 heures de transcription réalisées), avec des résultats très probants : jusqu'à 50% de temps de gagné sur certaines expériences.

Comme l'illustre la figure 25, l'interface de XTRANS est assez proche de celle de TRANSCRIBER, avec une fenêtre de texte à défilement vertical (2) et un spectrogramme occupant toute la partie inférieure (5). Cependant, les locuteurs ont ici un espace dédié, sur le côté gauche (1). À ce propos il est également possible, dans ce même espace, d'indiquer des informations concernant le type de données transcrites (*report*, *conversational* et *nonnews*). Cette option était également disponible sous TRANSCRIBER, mais XTRANS apporte également la possibilité d'annoter chaque segment selon le type d'énoncé qu'il contient. Les questions peuvent ainsi être facilement isolées et analysées, même s'il convient d'adopter une segmentation en phrase (et non en « tour de parole », par exemple) pour que ce paramètre soit efficace. Le parallèle avec TRANSCRIBER ne s'arrête pas là : XTRANS permet en effet d'ajouter quelques informations sur les locuteurs (sexe, langue maternelle ou non), de même qu'il propose lui aussi un jeu de balises pour annoter des événements paratextuels (rires, toux, éternuements...). Qui plus est, l'utilisateur peut créer ses propres balises, et en nombre illimité.

Paradoxalement, si XTRANS a les qualité de TRANSCRIBER, il lui emprunte aussi un défaut majeur : l'impossibilité de lire les vidéos. Cette restriction est d'autant plus dommageable qu'une autre vient s'y ajouter : le format .mp3, pourtant largement répandu aujourd'hui, n'est pas pris en charge. Cependant, les fichier .wav, .sph, .aiff, .flac ou encore .ogg sont eux parfaitement supportés. Puisque nous évoquons les formats supportés par XTRANS, il est tout de suite remarquable de voir que le logiciel supporte (officiellement du moins) les fichiers .trs (TRANSCRIBER) en entrée comme en sortie. Il est à notre connaissance le seul logiciel de transcription à afficher de telles possibilités. Mais une nouvelle fois, la pratique vient fortement nuancer cette affirmation.

S'il est en effet possible d'importer des fichiers .trs sous XTRANS, le résultat laisse à désirer sur plusieurs points. Tout d'abord, les balises de TRANSCRIBER ne sont pas gérées en tant que telles, et se retrouvent donc affichées en mode texte. Une transcription richement annotée devient donc illisible, puisqu'il est presque impossible de distinguer le texte de son enveloppe paratextuelle :

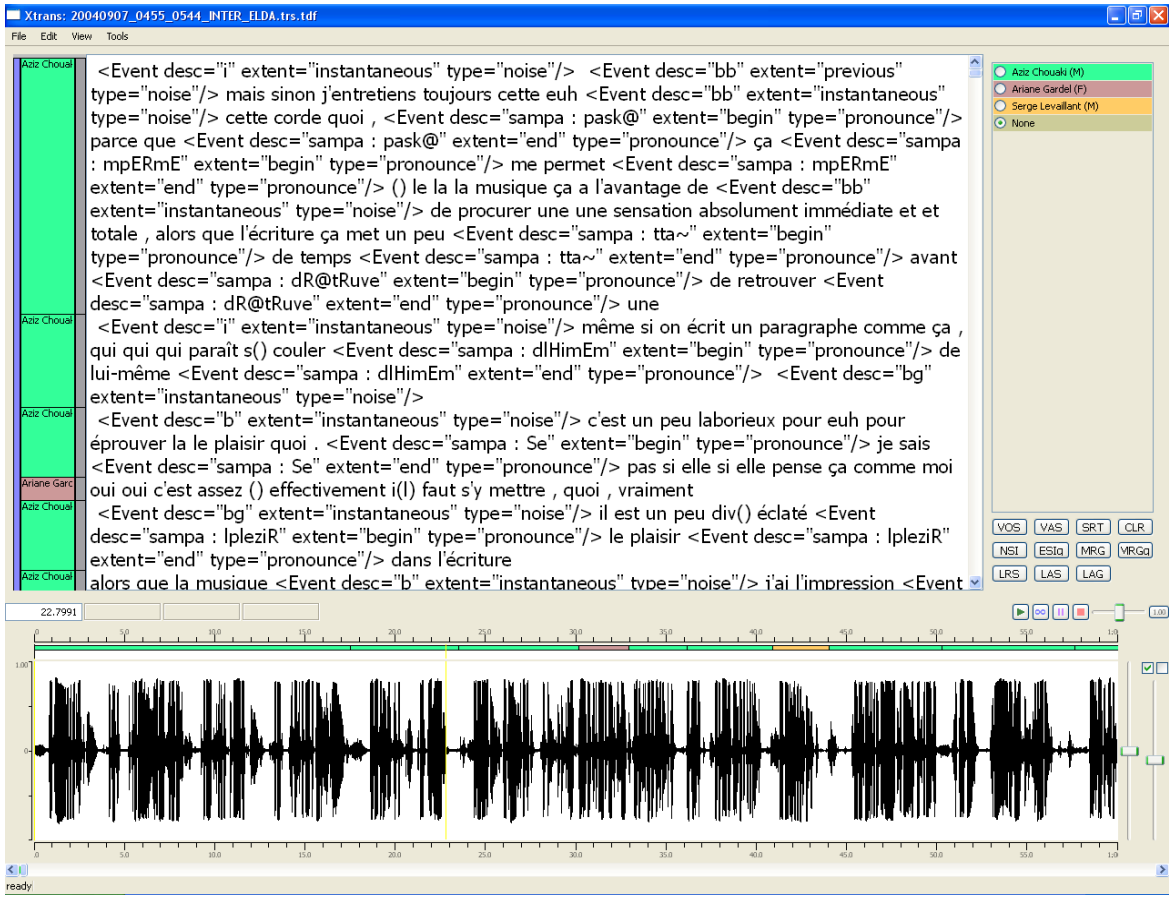


Figure 26 : Affichage d'un fichier .trs importé sous XTRANS

Si, comme nous l'avons vu, la gestion des locuteurs de XTRANS permet de conserver un minimum d'informations à leur égard lors de l'importation (beaucoup sont cependant perdues lors de l'importation, comme avec les autres logiciels), celle-ci souffre néanmoins d'un gros défaut. Pour une raison inconnue, lorsqu'un fichier .trs et son fichier audio correspondant sont importés, il est impossible d'aligner automatiquement le signal à partir du texte. Pour le dire autrement, lorsque l'utilisateur clique sur un segment de parole dans la fenêtre textuelle, le curseur du spectrogramme ne s'aligne pas en conséquence. Il est donc impossible d'écouter le segment en question, sauf à déplacer manuellement le curseur sur l'échelle temporelle. Ce problème n'apparaît pas lors de la démarche inverse : en cliquant sur l'échelle située au dessus du spectrogramme, la fenêtre de texte « défile » jusqu'au segment de parole concerné. Cependant, procéder de la sorte est très inconfortable, et aligner le moindre segment audio avec le texte correspondant relève alors de la gageure.

L'exportation vers le format .trs souffre elle aussi de soucis majeurs, révélateurs des

problèmes d'interopérabilité dont nous faisons état dans cette thèse. Par exemple, au niveau de la gestion des locuteurs, il est possible sous XTRANS de spécifier si un intervenant parle en langue maternelle ou non. Pour ce faire, il faut choisir entre *native* et *other*. Or, sous TRANSCRIBER, les deux possibilités sont *native* et *nonnative*. Ainsi, si un locuteur a été étiqueté *other* avec XTRANS, le fichier exporté en .trs refusera de s'ouvrir sous TRANSCRIBER car l'entité *other* lui est inconnue. Le problème est similaire avec l'annotation des sections : *report*, *filler* et *nontrans* sont disponibles sous TRANSCRIBER, tandis que XTRANS laisse le choix entre *report*, *conversational* et *nonnews*. Seul le choix systématique de *report* permet donc de traiter un fichier XTRANS sous TRANSCRIBER, ce qui contraint le transcripteur à produire une annotation volontairement erronée.

Enfin, XTRANS présente d'étonnants défauts au niveau de la gestion audio. Le premier est la qualité globale de son lecteur intégré, bien en dessous de Windows Media Player par exemple. La différence sur un même fichier est flagrante, sans qu'aucune explication ne soit donnée (réduction du taux d'échantillonnage, de la fréquence, etc.). Ensuite, si XTRANS présente une fonction de lecture ralentie très intéressante *a priori*, celle-ci déforme nettement le signal dès que l'on diminue la vitesse, ne serait-ce que de 20%. WINPITCHPRO s'en sort à bien meilleur compte, d'autant qu'il manque également dans l'interface de XTRANS quelques fonctions très utiles, notamment l'avance ou le retour rapide. Il est possible d'utiliser des raccourcis clavier pour pallier quelques-uns de ces manques, mais l'ensemble reste assez peu pratique. Entre autres, le curseur du signal audio n'est pas totalement dynamique, si bien que la touche « lecture / pause » fait partir le signal à partir d'un endroit qui n'est pas forcément le segment où l'on se trouve. Pour lire le segment précisément sélectionné, il faut passer par un autre raccourci clavier. Enfin, il est dommage que par défaut, la lecture activée par le bouton « play » s'arrête systématiquement à la première frontière qu'elle rencontre. Il n'est donc pas possible d'écouter deux ou trois segments consécutifs sans faire une manipulation préalable.

3.9 Conclusion

Que conclure à l'issue de cette étude consacrée à une dizaine de logiciels d'aide à la transcription ? Tout d'abord, que le choix de l'utilisateur doit avant tout être conditionné par ses besoins. En effet, la simple volonté de traiter de la vidéo élimine par exemple un tiers des outils précités : TRANSCRIBER, PRAAT et XTRANS.

Ensuite, sans même avoir à tenir compte de cette distinction liminaire, il convient d'évoquer le problème des plates-formes de disponibilité. Là encore, tous les logiciels ne se ressemblent pas : WINPITCHPRO est le seul à ne fonctionner que sous un seul OS (Windows), TRANSANA ne peut s'installer sous Linux tandis que XTRANS n'a pas de version prévue pour Mac. Cela étant, tous les autres logiciels présentés ici sont multi plates-formes, ce qui représente un effort de standardisation bienvenu.

La disponibilité de tous ces outils est un autre critère qu'il convient de prendre en compte rapidement. En effet, si l'utilisateur envisage d'utiliser WINPITCHPRO ou TRANSANA par exemple, il devra payer pour pouvoir les utiliser en version complète. Bien sûr, il est toujours possible d'utiliser librement des versions dites d'évaluation, mais leurs possibilités sont plus ou moins limitées. Ainsi la version libre de TRANSANA est, comme nous l'avons vu, beaucoup moins restrictive que celle de WINPITCHPRO. En effet, il n'est officiellement possible de l'utiliser gratuitement que pendant quinze jours ; passé ce délai, le logiciel exigera l'achat d'une licence (100 $) pour démarrer. Signalons enfin le cas d'ANVIL qui, bien que gratuit, doit faire l'objet d'une demande à l'auteur avant de pouvoir être téléchargé.

Pour ce qui est de l'utilisation et de la prise en main de ces logiciels, il semble que TRANSCRIBER soit le plus accessible, bien que ce jugement soit assurément biaisé par l'usage intensif que nous en avons fait. Cela étant, nous avons également confronté des néophytes à TRANSCRIBER, et il a suffi d'une petite heure d'explications pour qu'ils deviennent autonomes et maîtrisent même quelques fonctions avancées du logiciel. Une expérience semblable avec l'un des autres programmes présentés ici aurait très probablement demandé un apprentissage beaucoup plus long. En effet, TRANSCRIBER présente un avantage incontestable au niveau de l'ergonomie : les touches TAB et ENTREE permettent instantanément de réaliser deux actions essentielles, à savoir lire (ou mettre en pause) un flux

audio et y insérer une frontière. Ce détail peut paraître anecdotique, mais il permet en fait de s'affranchir de la souris pour la tâche de segmentation, qui peut donc être réalisée de façon très aisée. Le fait de pouvoir ensuite facilement affiner cette segmentation avec une seule touche (le bouton central de la souris) permet d'établir très facilement le « squelette » d'une transcription. A ce titre, attribuer des locuteurs à chaque tour de parole se fait également de façon aisée, *via* un raccourci clavier assez bien pensé. Des logiciels tels que XTRANS, TRANSANA ou encore ELAN sont loin d'être aussi intuitifs, peut-être aussi parce qu'ils permettent une segmentation extrêmement rigoureuse, notamment au niveau de la parole superposée. En effet, sur ce point TRANSCRIBER est quelque peu imprécis puisqu'on ne peut y synchroniser que deux flux de parole, et pas de manière indépendante. Il en résulte parfois une segmentation hachée, avec l'obligation de créer un nouveau tour de parole dès que deux locuteurs parlent en même temps, et ce même si cela ne dure que quelques dixièmes de seconde.

En ce qui concerne la transcription à proprement parler, aucun logiciel ne se distingue dans un premier temps, puisque chacun permet de saisir les informations textuelles dans de bonnes conditions. EXMARaLDA et TRANSCRIBER sont néanmoins les plus complets pour la gestion des locuteurs, suivis par XTRANS qui offre également quelques options intéressantes. Les autres logiciels ne proposent que la simple possibilité d'indiquer un nom et / ou un prénom. Pour ce qui est des autres données métatextuelles, TRANSCRIBER est une nouvelle fois au dessus du lot grâce à ses balises, très complètes et surtout entièrement personnalisables. XTRANS propose également plusieurs balises prédéfinies, ainsi que la possibilité d'en créer à l'envi. EXMARaLDA permet quant à lui d'intégrer différents éléments de la TEI relatifs à la parole, ce qui est une excellente initiative dans la mesure où la TEI tend à devenir un standard de la normalisation des données. TRANSANA, pour sa part, permet de créer plusieurs fenêtres de transcription synchronisées, ce qui très utile si le transcripteur souhaite annoter différents éléments en plus du texte (notamment par rapport à la vidéo, gérée de façon efficace sous ce logiciel).

Puisque nous venons d'évoquer le support vidéo, il convient désormais de s'y attarder d'un peu plus près. Nous l'avons dit en préambule de ce bilan, trois des logiciels ici étudiés ne la gèrent pas. Concernant les six autres, tous ne permettent pas les mêmes possibilités avec les fichiers vidéo. En premier lieu, au niveau des formats supportés, TRANSANA est le plus complet. Il supporte en effet toutes les extensions classiques (.avi., .mpeg, .mp4, .wmv...),

bien que quelques problèmes de précision puissent apparaître au niveau de la segmentation. EXMARaLDA présente théoriquement des possibilités similaires, mais nous avons eu l'occasion de voir que dans la pratique, la réalité était tout autre. La remarque vaut également pour WINPITCHPRO, dont le lecteur vidéo n'est pourtant pas à mettre en cause directement ; c'est plutôt la gestion des fichiers volumineux qui est ici déficiente. ELAN est intéressant dans la mesure où il gère le multi-angles, avec quelques difficultés cependant. Toutefois, le fait de ne supporter ni les fichiers .avi ni les fichiers .wmv lui est plutôt préjudiciable, dans la mesure où ces deux formats sont très répandus à l'heure actuelle. Enfin, ANVIL se veut modeste (seuls deux formats sont gérés, et il n'y a pas d'affichage multi-angles possible) mais plutôt efficace, à condition que les fichiers traités ne soient pas trop volumineux.

Certains des outils que nous avons évoqués existent maintenant depuis près de 20 ans, tandis que d'autres sont nés dans les années 2000. Pour autant, le plus « ancien » d'entre eux, PRAAT, est toujours régulièrement mis à jour (plusieurs fois par mois), bénéficiant d'une attention toute particulière de la part de ses auteurs. EXMARaLDA est dans cette même dynamique, avec de nouvelles versions régulièrement proposées et un site internet bien entretenu. A l'inverse, WINPITCHPRO ou TRANSCRIBER sont à l'abandon depuis plusieurs années maintenant, mais probablement pas pour les mêmes raisons.

Le premier cité est, nous l'avons dit, payant. Parallèlement, ses sources n'est pas librement diffusé, et aucun utilisateur ne peut donc modifier le logiciel de sa propre initiative. Pour sa part, l'auteur n'ayant manifestement plus le temps d'assurer le développement de WINPITCHPRO, ce dernier n'a plus connu d'évolution depuis décembre 2006. De même, aucune information récente ne peut être trouvée sur son site internet.

Concernant TRANSCRIBER, le problème est tout autre. Le logiciel est libre et ses sources sont disponibles, mais le langage avec lequel il a été programmé (Tcl/Tk) rend délicate toute modification. Par ailleurs, malgré un carnet de route prometteur sur son site internet, aucune mise à jour avant mai 2008 où il a été question d'une refonte totale de TRANSCRIBER. Celle-ci serait basée sur l'Annotation Graph Toolkit (AGTK), et devait voir le jour au second semestre 2009. Mais à ce jour (11 septembre 2010), aucune nouvelle version du logicielle n'est disponible, et son site internet n'a pas été mis à jour depuis cette annonce qui date maintenant de deux ans.

La principale question qui se pose à propos de ces logiciels, de façon générale, est celle de l'échange des données. En attendant que la TEI (dont nous parlerons longuement dans le

chapitre suivant) ne vienne éventuellement s'intégrer aux formats prioritaires cités ici (.trs, .textgrid, .wp2...), un outil récent se propose d'établir des passerelle entre eux : Transformer[22].

Développé par Oliver Ehmer, Transformer gère en entrée les fichiers issus de TRANSCRIBER, PRAAT, ELAN, TRANSANA ainsi que de quelques autres logiciels non traités ici. Il se propose ensuite de les convertir sous différentes formes : fichiers de transcription pour les logiciels précités, documents texte ou encore base de données.

Par ailleurs, les formats d'entrée et de sorties concernant les fichiers de transcription ne sont pas les mêmes. Si les fichiers .trs peuvent importés, TRANSFORMER ne permet d'en générer en sortie, ce qui est quelque peu dommage au vu des nombreux utilisateurs de TRANSCRIBER. À l'inverse, il est possible d'exporter des données sous la forme de fichiers CLAN (.cha), mais ces derniers ne sont pas supportés en entrée. Enfin, nous signalerons que pour une raison inexpliquée, aucun des fichiers .trs que nous avons tenté d'ouvrir avec Transformer (version 6.0) n'a pu s'afficher correctement à l'écran.

Il semble donc que l'idée d'un format commun comme celui proposé par la TEI soit encore le moyen le plus efficace de pouvoir échanger des données, en ne perdant qu'un minimum d'informations. Car c'est là le principal enseignement de l'étude que nous avons menée dans les pages précédentes : tous les logiciels cités ont leurs défauts, mais le plus important demeure la divergence des formats de chacun d'entre eux. C'est pourquoi notre chapitre suivant sera consacré à ce problème du codage des données, avec en point de mire une solution potentielle : la TEI.

22 http://www.oliverehmer.de/transformer/

Chapitre 4 : Le codage

Des projets tels que EPAC ou VARILING apportent un soin tout particulier à l'accessibilité et à l'interopérabilité de leurs données. Bien qu'étant chacun dans une perspective différente, leurs responsables respectifs tiennent à ce que leur travail puisse être utilisé et exploité dans un maximum d'applications : sciences du langage, sciences humaines, politique linguistique ou didactique... Ces efforts, parce qu'encore assez peu répandus, sont à mettre en avant. S'il est en effet certain que la moindre transcription réalisée par une équipe de recherche a toujours un but précis (étude morpho-syntaxique, prosodique, sociologique, lexicale ; reconnaissance de la parole...), il est regrettable que le logiciel utilisé pour cela « verrouille » parfois son utilisation.

L'exemple de la gestion des locuteurs est très représentatif : nous l'avons dit, il est possible avec TRANSCRIBER d'ajouter de nombreuses informations les concernant (sexe, nationalité, accent, canal utilisé, nature de la parole...). Or, parmi les autres outils qui revendiquent pourtant une compatibilité avec les fichiers de TRANSCRIBER, aucun ne conserve l'intégralité de ces informations, parfois très utiles. Certains, comme PRAAT, vont même jusqu'à ne retranscrire que les noms et prénoms. Ainsi, dans pareil cas de figure, il faut reproduire « à la main » les données que l'on souhaitait conserver lors du passage d'un logiciel à un autre.

Il s'agit ici d'un problème de format, ou plus précisément de codage : les fichiers dans lesquels sont stockés les informations concernant la transcription ne codent pas les mêmes choses de la même façon, la faute à des structures différentes.

Nous nous intéresserons donc dans un premier temps à ces divergences, avant de nous attarder en détails sur ce qui pourrait en être une solution : la TEI (Text Encoding Initiative), qui propose un format d'échange proche du langage XML[23] et ayant pour principaux atouts l'interopérabilité et la pérennité des données.

23 Extensible Markup Language

4.1 Quelques généralités

Si l'annotation des corpus de parole spontanée est déjà source de nombreux soucis d'homogénéité, comme nous avons eu l'occasion de le démontrer, le codage de ces annotations pose également problème. Cependant, il convient tout d'abord de distinguer clairement l'annotation et le codage.

- l'annotation est la représentation à l'écran d'un phénomène textuel ou paratextuel, quel qu'il soit. Ainsi, selon les conventions ESTER, les parenthèses sont l'annotation utilisée pour indiquer une troncation. De même, les balises proposées par le logiciel TRANSCRIBER sont également autant d'annotations mises à la disposition du transcripteur.

- le codage, quant à lui, est la façon dont sont stockées ces annotations au sein du fichier informatique contenant la transcription. Ce codage ne se voit donc pas directement dans l'interface du logiciel utilisé ; il peut devenir visible si l'on ouvre le fichier de sortie avec un éditeur de texte, par exemple. Pour autant, cette partie immergée de la transcription est essentielle pour permettre l'échange et la compatibilité des corpus. Ainsi, si deux transcriptions réalisées avec deux logiciels et deux façons d'annoter différentes ont un format de sortie commun (XML par exemple), il y a de fortes chances pour qu'elles soient réutilisables et échangeables sans devoir faire trop de modifications. A l'inverse, si les formats de sortie sont totalement différents, alors l'interopérabilité sera beaucoup plus délicate et certaines données risquent même d'être perdues, faute d'avoir pu les exploiter pleinement et surtout facilement. Loin d'être un détail, ce problème est aujourd'hui un véritable enjeu pour ne pas rendre caduques des données précieuses, qui sont souvent disponibles en petite quantité.

Par exemple, le logiciel TRANSCRIBER possède un format de sortie (.trs) très proche du XML, ce qui le rend en théorie assez facilement exploitable, notamment par d'autres logiciels de transcription. Cela étant, comme ces derniers n'ont pas la même interopérabilité, les choses se compliquent grandement : PRAAT, entre autres, permet bel et bien une importation de fichiers issus de TRANSCRIBER, mais son propre format de sortie (.textgrid) est malheureusement assez pauvre. Ainsi, de nombreuses informations contenues dans les fichiers .trs, notamment concernant les locuteurs, sont perdues lors du passage d'un logiciel à un autre.

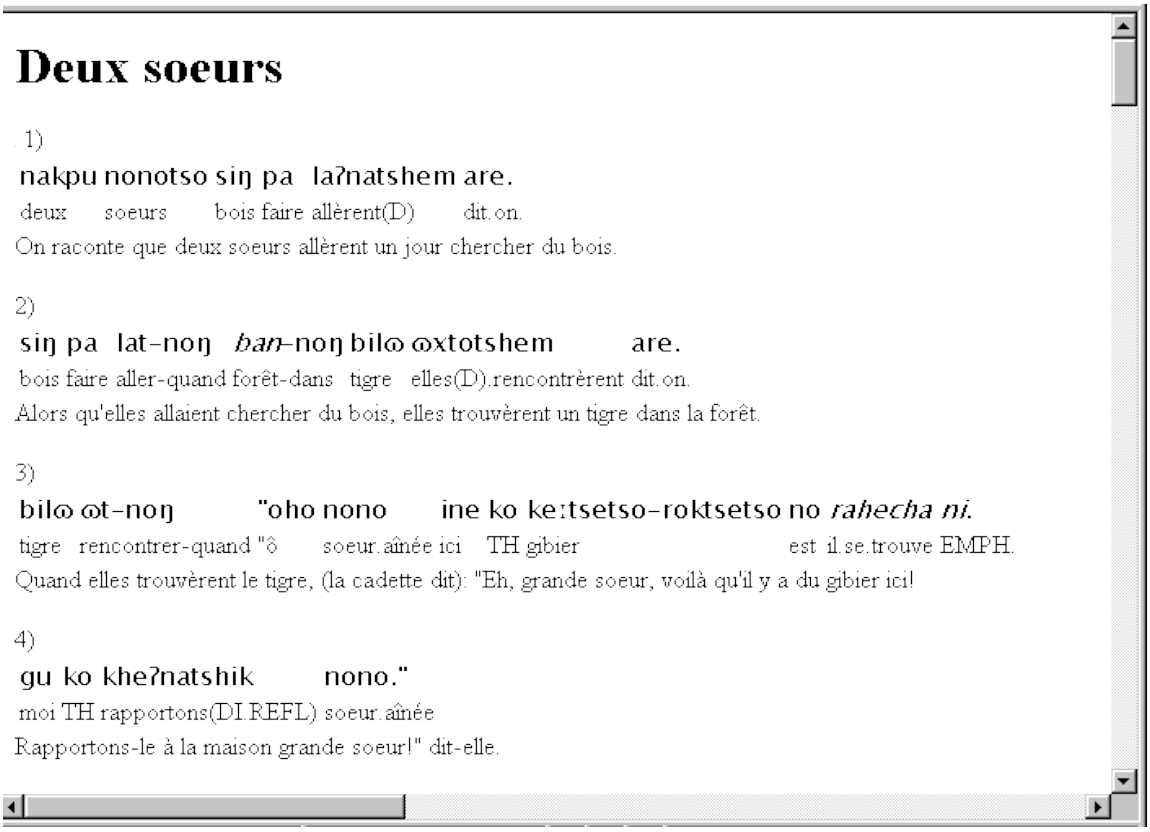
Cependant, prenant conscience de ce problème, quelques développeurs commencent à proposer des plates-formes de conversion performantes, pouvant traiter un maximum de formats de fichiers. Nous avons déjà mentionné dans le chapitre précédent le logiciel TRANSFORMER, sorte de passerelle entre différents logiciels de transcription. Cependant, celui-ci est payant et souffre de quelques défauts pénalisants.

À ce titre, des entreprises comme celle du LACITO[24] (LAngues et CIvilisations à Tradition Orale) ou de CATCOD[25] (CATalogage et CODage des corpus oraux) sont tout à fait salutaires. Le laboratoire de recherches du LACITO se propose en effet d'archiver et de diffuser des enregistrements de parole spontanée concernant des langues rares. 200 documents représentant 44 langues sont ainsi disponibles sous la forme d'enregistrements sonores, accompagnés de leurs transcriptions et traductions en français. Les transcriptions sont balisées en XML et accompagnées de deux feuilles de style[26] qui permettent chacune un mode d'affichage différent : l'un avec la traduction mot-à-mot et l'autre sans.

24 http://lacito.vjf.cnrs.fr/archivage/index.htm

25 http://crdo.risc.cnrs.fr/exist/crdo/catcod.htm

26 Document permettant de fournir des informations spécifiques quant à l'affichage des données



Deux soeurs

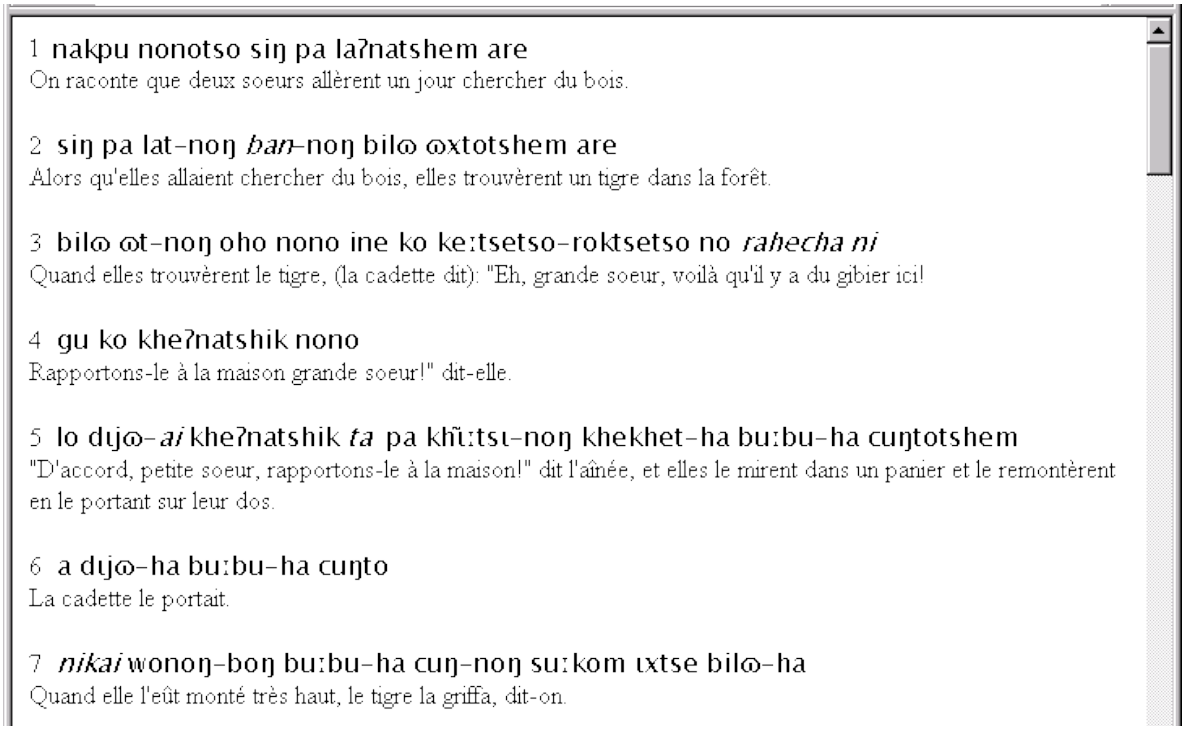
1)
nakpu nonotso siŋ pa laʔnatshem are.
deux soeurs bois faire allèrent(D) dit.on.
On raconte que deux soeurs allèrent un jour chercher du bois.

2)
siŋ pa lat-noŋ *ban*-noŋ bilɷ ɷxtotshem are.
bois faire aller-quand forêt-dans tigre elles(D).rencontrèrent dit.on.
Alors qu'elles allaient chercher du bois, elles trouvèrent un tigre dans la forêt.

3)
bilɷ ɷt-noŋ "oho nono ine ko keːtsetso-roktsetso no *rahecha ni.*
tigre rencontrer-quand "ô soeur.aînée ici TH gibier est il.se.trouve EMPH.
Quand elles trouvèrent le tigre, (la cadette dit): "Eh, grande soeur, voilà qu'il y a du gibier ici!

4)
gu ko kheʔnatshik nono."
moi TH rapportons(DI.REFL) soeur.aînée
Rapportons-le à la maison grande soeur!" dit-elle.

Figure 27 : Exemple de transcription proposée par le LACITO (avec mot-à-mot)



1 nakpu nonotso siŋ pa laʔnatshem are
On raconte que deux soeurs allèrent un jour chercher du bois.

2 siŋ pa lat-noŋ *ban*-noŋ bilɷ ɷxtotshem are
Alors qu'elles allaient chercher du bois, elles trouvèrent un tigre dans la forêt.

3 bilɷ ɷt-noŋ oho nono ine ko keːtsetso-roktsetso no *rahecha ni*
Quand elles trouvèrent le tigre, (la cadette dit): "Eh, grande soeur, voilà qu'il y a du gibier ici!

4 gu ko kheʔnatshik nono
Rapportons-le à la maison grande soeur!" dit-elle.

5 lo dɟɷ-*ai* kheʔnatshik *ta* pa khɯ̃ːtsɪ-noŋ khekhet-ha buːbu-ha cuŋtotshem
"D'accord, petite soeur, rapportons-le à la maison!" dit l'aînée, et elles le mirent dans un panier et le remontèrent en le portant sur leur dos.

6 a dɟɷ-ha buːbu-ha cuŋto
La cadette le portait.

7 *nikai* wonoŋ-boŋ buːbu-ha cuŋ-noŋ suːkom ɪxtse bilɷ-ha
Quand elle l'eût monté très haut, le tigre la griffa, dit-on.

Figure 28 : Exemple de transcription proposée par le LACITO

CatCod, pour sa part, est un groupe de travail coordonné par Serge Heiden, Emmanuel Schang et Michel Jacobson (déjà fortement impliqué au sein des travaux du LACITO), s'intéressant surtout au partage des corpus oraux. Cela signifie bien entendu que le codage est au centre de ses priorités, puisque c'est en passant par sa standardisation qu'un maximum de données pourra être partagé. CatCod vise donc à définir à proposer un ensemble de codages uniforme pour la communauté parole, ensemble qui se veut en lien étroit avec la TEI.

Pour ce faire, ses responsables ont notamment organisé en décembre 2008 le premier *workshop* international CatCod, qui s'est tenu à Orléans[27]. Celui-ci a permis de réunir officiellement tous les membres de la communauté CatCod, afin d'échanger des réflexions sur le partage des données et surtout de mettre à plat tous les travaux antérieurs sur le sujet. Chacun a ainsi pu convenir du rôle essentiel qu'était amenée à jouer la TEI dans les années à venir en tant que standard de codage universel.

Il existe d'ailleurs certains organismes qui proposent déjà, depuis plusieurs années, des possibilités d'interaction avec la TEI. La base de données CLAPI[28] (équipe ICAR, Lyon) offre par exemple une fonction de conversion très intéressante, puisqu'elle permet d'exporter les transcriptions de sa plate-forme au format TEI, *via* un processus entièrement automatisé.

Il est d'ailleurs désormais temps d'analyser en détail ce qu'est la Text Encoding Initiative (TEI), tout d'abord à travers une présentation globale de son consortium. Ensuite, nous nous intéresserons spécifiquement aux codages qu'elle proposent, en nous focalisant sur ceux réservés au domaine de la parole. Enfin, nous essaierons de démontrer que les recommandations de la TEI sont aujourd'hui une étape qu'il est indispensable de franchir : d'une part pour normaliser l'ensemble des corpus disponibles aujourd'hui, et sortir de l'oubli tous ceux qui demeurent inutilisables faute de codage efficace (données manuscrites, tapées à la machine, etc.) ; d'autre part pour s'assurer que ces corpus seront toujours exploitables dans les années à venir, en les associant à un format informatique synonyme de pérennité.

27 http://www.catcod.org/

28 http://corpus.univ-lyon2.fr/

4.2 La TEI

4.2.1 Introduction

La TEI (Text Encoding Initiative) est un consortium dont les bases ont été posées en 1987, en vue de proposer un codage numérique et normalisé de toutes sortes de documents (textes imprimés, documents sonores...). Trois sociétés savantes sont à l'origine de sa création : l'*Association for Computers in the Humanities*, l'*Association for Literary and Linguistic Computing* et l'*Association for Computational Linguistics*. La première matérialisation du codage est proposée sous la forme de *Guidelines* (ensemble de recommandations pour un bon usage de la TEI) portant la numérotation P1, en juin 1990. Cependant, la première version officielle de ces *Guidelines* sera la P3, disponible quatre ans plus tard, en mai 1994.

Mais c'est seulement en 1999 que le consortium TEI est véritablement et officiellement créé, avec pour but premier de promouvoir la TEI à l'échelle internationale. Ce à quoi il va s'évertuer, tout en structurant progressivement son organisation et ses activités : recherche de sponsors, fédération et développement d'une communauté d'utilisateurs, etc.

L'un des atouts majeurs de la TEI est d'ailleurs qu'elle demeure très à l'écoute de cette communauté d'utilisateurs. Ceux-ci peuvent soumettre de nouvelles propositions de codages, signaler l'inutilité de l'un d'entre eux, etc. À ce titre, une site internet collaboratif (*wiki*) a été créé[29], où il suffit de s'inscrire pour proposer de nouvelles évolutions à la TEI. Le conseil scientifique de la TEI, composé de douze membres et présidé par Laurent Romary, examine ensuite ces requêtes lors de ses réunions annuelles. Si certaines sont jugées pertinentes, elles sont alors validées puis intégrées à la version suivante des *Guidelines*.

Jusqu'à sa version P3, la TEI était basée sur le SGML[30], un langage de description à balises qui est né en 1969. Néanmoins, à la fin des années 90, le langage XML (version simplifiée du SGML) s'impose progressivement, et le consortium TEI décide donc de s'appuyer sur ce nouveau standard émergeant. Ainsi la version P4 des *Guidelines* de la TEI, parue en juin 2002, n'est en fait que la version P3 réécrite en langage XML. A la fin de l'année 2007, une version P5 avec de nombreuses améliorations voit le jour, et c'est celle-ci qui fait

29 http://wiki.tei-c.org/index.php/Main_Page

30 Standard Generalized Markup Language

encore référence à l'heure actuelle.

4.2.2 SGML : un langage de balisage

Le SGML (Standard Generalized Markup Language) appartient à la catégorie des langages dits « de balisage ». Les balises sont représentées à l'écran par des chevrons (< et >) qui permettent de structurer un document, le cas échéant selon plusieurs niveaux d'imbrication.

```
<bibliothèque>
  <livre>La Légende des Siècles</livre>
  <livre>Les Misérables</livre>
  <livre>Astérix en Corse</livre>
</bibliothèque>
```

Comme nous le voyons ici, une balise est généralement composée de deux éléments : l'un ouvrant (<...>) et l'autre fermant (</...>) l'élément concerné. Il existe un nombre infini de balises pouvant être utilisées dans un document SGML. La seule condition est qu'elles soient compulsées dans une DTD[31], c'est-à-dire dans une structure définissant la syntaxe des balises utilisées à l'intérieur d'un document (quelle balise peut être utilisée à l'intérieur d'une autre, par exemple). Même s'il existe une certaine nomenclature par défaut, toute balise est considérée comme acceptable à l'intérieur d'un document si elle est déclarée dans sa DTD.

En plus de cette DTD, un document SGML contient également une feuille de style qui, elle, expose les modalités de présentation dudit document. Cette feuille de style permet de spécifier un mode d'affichage précis du document, suivant la forme sous laquelle on souhaite le faire apparaître.

Prenons en exemple le cas une maison d'édition, qui édite des livres au format classique et également au format « poche ». Elle peut créer deux feuilles de style (une pour chaque format) avec différentes informations concernant la taille de police, les espaces entre les paragraphes, etc. Ainsi, pour chaque ouvrage qu'elle souhaite publier, elle n'a qu'à appliquer au texte brut la feuille de style idoine, et le texte sera automatiquement mis en page sur l'écran.

31 Document Type Definition

Le format HTML, très utilisé par Internet pour l'affichage des pages Web, est une application du SGML. Par exemple, lorsqu'un texte apparaît centré sur un site Internet, c'est que se trouve dans le codage de cette page une balise qui le spécifie. A l'inverse du SGML, les balises en HTML ont chacune une signification précise définie par le W3C[32], un organisme chargé de standardiser et de rendre compatibles entre elles les technologies du Web. Ainsi, on utilisera par exemple les balises et pour coder respectivement les caractères italiques et gras.

Voici comment cette nomenclature apparaît dans le code HTML (« invisible » pour l'utilisateur classique), et ce que voit justement cet utilisateur sur son écran :

code HTML :	ceci est un exemple de titre
représentation à l'écran :	ceci est un exemple de titre
code HTML :	ceci est un exemple de titre
représentation à l'écran :	*ceci est un exemple de titre*
code HTML :	ceci est un exemple de titre
représentation à l'écran :	**ceci est un exemple de titre**

4.2.3 XML : le SGML simplifié

Dans les années 1990, comme nous l'avons vu en introduction, le langage XML (Extensible Markup Language) est venu succéder au SGML. Son principal objectif était de simplifier le SGML tout en en conservant les avantages, et notamment l'interopérabilité. Ainsi, on note quelques différences remarquables entre les deux langages :

- la DTD n'est plus indispensable en XML : plusieurs vocabulaires de balises peuvent donc cohabiter, sans nécessiter la présence d'un document spécifique pour le justifier.

- en SGML, les caractères sont encodés selon la norme ASCII[33] [André et Goossens, 1995], c'est-à-dire en alphabet latin sans lettres accentuées. Ainsi, les caractères français

32 World Wide Web Consortium

33 American Standard Code for Information Interchange

« é » , « è », « à »... n'y sont par exemple pas directement représentés. Il faut pour les représenter à l'écran passer par un encodage particulier : &eacute pour « é », &egrave pour « e », &agrave pour « à », etc. À l'inverse, le langage XML s'appuie sur la norme Unicode [André et Goossens, 1995], qui considère tous les caractères de tous les systèmes d'écriture comme autant d'entités spécifiques. Chacun de ces caractères est ainsi nommé et identifié numériquement, et cet identifiant reste le même quel que soit le contexte d'utilisation.

- XML exige que toute balise ouverte soit fermée, alors qu'il est possible en SGML d'utiliser quelques balises à fermeture optionnelle.

Ainsi le langage XML, en s'appuyant sur les préceptes de son aîné SGML, apparaît comme étant une base répondant de façon efficace à la problématique posée par le codage des corpus de parole : en effet, l'objectif explicitement revendiqué est l'interopérabilité sous la forme d'un langage simplifié autant que possible, et assurant un échange et une compatibilité facilités. C'est précisément ce que recherchait le consortium de la TEI lorsqu'il s'est créé, et il est donc tout à fait compréhensibles que les *Guidelines* de la TEI se soient appuyés tout d'abord sur le format SGML, et ensuite sur XML. Nous allons donc à présent nous intéresser précisément au contenu des recommandations de la TEI, et plus particulièrement à celles concernant le codage de la parole.

4.2.4 La TEI : un codage fait pour les corpus oraux ?

Dans ses *Guidelines*, la TEI présente donc ensemble de recommandations permettant et/ou facilitant la préservation et l'échange électronique de données, notamment textuelles. Reposant sur un système de balises proche du format XML, elle permet ainsi de « coder » (ou de « baliser ») un grand nombre d'informations ayant trait à tout document électronique. Ces recommandations se doivent d'être simples et accessibles à tous, de façon à ce qu'une large communauté d'utilisateurs puisse y recourir. Les principaux codages doivent pouvoir être compris par tous, et ceux plus spécifiques être supprimés ou ajoutés au « squelette » existant sans que cela pose de difficultés. Ainsi, suivant l'usage qu'il souhaite en faire, un utilisateur pourra en quelque sorte adapter la TEI à ses besoins.

Par ailleurs, la TEI a certes pour vocation de créer des nouveaux documents suivant ses normes, mais aussi de coder ceux qui existent déjà. Cette précision revêt une importance toute particulière aujourd'hui, à l'heure où la numérisation des textes entre - parfois avec fracas - dans les mœurs. Pour ne pas se retrouver confronté à des masses de données inexploitables

parce que désolidarisées les unes des autres, il convient de proposer un principe d'archivage simple mais complet, qui permette aux futurs chercheurs d'exploiter pleinement le matériau mis à leur disposition. Et à ce titre, des initiatives comme la TEI semblent tout à fait appropriées, puisque cette dernière prend en considération des problèmes allant jusqu'à la distinction entre guillemets et apostrophes, fréquents notamment dans la langue anglaise. Tel l'un des intérêts majeurs de la TEI : permettre des échanges de documents allant au-delà de la barrière du langage, et dont le codage permettra au moins de comporter de nombreux éléments structurels ou paratextuels, à défaut de maîtriser parfaitement la langue source concernée.

Pour illustrer ces propos, nous reproduisons ci-dessous un exemple en français ainsi que sa normalisation selon la TEI, tous deux récupérés à l'adresse suivante : http://www.tei-c.org/Guidelines/Customization/Lite/teiu5_fr.html

- Texte classique :

« C'est un mois d'octobre... exceptionnel » , dit Gisèle Dufrène; ils acquiescent, ils sourient, une chaleur d'été tombe du ciel gris-bleu - Qu'est-ce que les autres ont que je n'ai pas ? - ils caressent leurs regards à l'image parfaite qu'ont reproduite ***Plaisir de France*** et ***Votre Maison*** : la ferme achetée pour une bouchée de pain - enfin, disons, de pain brioché - et aménagée par Jean-Charles au prix d'une tonne de caviar. (`` je n'en suis pas à un million près '', a dit Gilbert), les roses contre les murs de pierre, les chrysanthèmes, les asters, les dahlias `` les plus beaux de toute l'Ile-de-France '', dit Dominique; le paravent et les fauteuils bleux et violet - c'est d'une audace ! - tranchent sur le vert de la pelouse, la glace tinte dans les verres, Houdan baise la main de Dominique, très mince dans son pantalon noir et son chemisier éclatant, les cheveux pâles, mi-blonds, mi-blancs, de dos on lui donnerait trente ans. …

- Même texte codé selon la TEI :

```
<p> <q rend=frdqo> C'est un mois d'octobre ... exceptionnel
</q>, dit Gisèle Dufrène;
ils acquiescent, ils sourient, une chaleur d'été
tombe du ciel gris-bleu-
<q>Qu'est-ce que les autres ont que je n'ai pas? </q>
```

```
- ils caressent leurs regards á l'image parfaite
qu'ont reproduite <title>Plaisir de France</title> et
<title>Votre Maison</title>: la ferme achetée pour une
bouchée de pain -  enfin, disons, de pain
brioché - et aménagée par
Jean-Charles au prix d'une tonne de caviar. (<q rend=endqo>je
n'en suis pas á un million près</q>, a
dit Gilbert), les roses contre les murs de pierre, les
chrysanthèmes, les asters, les  dahlias <q rend=endqo>
les plus beaux de toute l'Ile-de-France</q>, dit Dominique;
le paravent et les fauteuils bleux et violet - <q>c'est
d'une audace!</q> - tranchent sur le vert de la pelouse,
la glace  tinte dans les verres, Houdan baise la main de
Dominique,  très mince dans son pantalon noir et son
chemisier éclatant, les cheveux pâles, mi-blonds,
mi-blancs, de dos on lui donnerait trente ans.
…
```

Si l'on observe la version « TEI » du texte, on s'aperçoit que des balises ont été utilisées pour coder différents phénomènes qui peuvent être ambigus, lors du passage d'une langue à une autre par exemple :

- les guillemets employés dans le texte original sont un mélange de typographies français et anglaise. Afin de ne pas perdre cette information dans le codage, deux types de balises « citation » ont été utilisées : <q rend=frdqo> pour les guillemets français (**fr**ench **d**ouble **q**u**o**te), et <q rend=endqo> pour les guillemets anglais (**en**glish **d**ouble **q**u**o**te). Précisons qu'un seul codage a été établi pour ces derniers, bien qu'ils apparaissent sous deux formes distinctes dans le texte original.

- les titres « Plaisir de France » et « Votre Maison » sont en gras et italiques dans le texte original, mais cela est plus un choix éditorial qu'une convention. Ainsi, en TEI, pour ne pas perdre la notion de titre tout en laissant le libre choix de la représentation visuelle à adopter, des balises <title> seront utilisées.

Également, et bien que ces choix n'aient pas été mis en œuvre dans l'exemple ci-dessus, les caractères particuliers d'une langue (accents, ligatures...) sont en général annotés de façon spécifique en TEI, afin de proposer un ensemble qui respecte la norme ASCII. Il s'agit de

rendre les fichiers facilement échangeables d'un utilisateur à un autre, quelle que soit la langue source. Ainsi, le « é » français sera codé « &eacute », et « œ » apparaîtra sous la forme « &oelig ».

La TEI a donc pour but d'uniformiser au mieux tout type de données auxquelles elle se trouve confrontée, par une structuration très précise des documents concernés. Nous ne pourrons naturellement pas ici détailler les *Guidelines* de la TEI, dont la dernière version (datant de novembre 2009) « pèse » quelque 1420 pages. Cependant, nous essaierons d'en présenter les grandes lignes, et notamment celles qui ont trait à la parole, puisque ce genre de données nous intéresse ici au premier chef

4.2.4.1 La TEI en quelques balises

Afin d'assurer une hiérarchisation précise, chaque document codé avec la TEI comporte un en-tête, intitulé *header*. Celui-ci est divisé en quatre parties principales :

- une description bibliographique du document concerné, que l'on représente avec la balise <fileDesc>. Celle-ci contient toutes les informations générales que l'on trouverait, par exemple, en tête d'un livre : titre, auteur, édition, nombre de « pages », etc.
- une description de la façon dont le document a été codé, signalée par la balise <encodingDesc>. Elle précise quelles sont les balises utilisées, le but du codage, la façon dont le texte source a été traité...
- une description du profil (<profileDesc>), qui est cette fois un élément optionnel. S'il est activé, il permet d'apporter des détails sur différents aspects descriptifs du document : langue utilisée, thème(s) abordé(s), détails sur sa création...
- un historique des révisions apportées au document, symbolisé par la balise <revisionDesc>. A nouveau, cette division est optionnelle, mais elle est toutefois fortement recommandée. A défaut, aucune information sur les modifications ou mises à jour apportées au document n'est disponible, ce qui rend difficile tout échange de données.

En ce qui concerne la structure du texte, la TEI comporte quatre balises de découpage liminaires :

- <front>, qui concerne tout ce que nous appellerons « avant-texte » : titre, préface, dédicace, etc.
- <group>, qui permet de regrouper plusieurs textes ou groupes de textes.
- <body>, c'est-à-dire le corps du texte proprement dit.
- <back>, qui référence les annexes suivant le texte.

Une fois que l'on arrive au niveau du texte lui-même, il va s'agir de le découper. Pour ce faire, deux balises seront principalement utilisées : <p>, pour délimiter les paragraphes, et , pour indiquer des divisions plus marquées (saut de ligne, de page ou de chapitre notamment). Suivant le type de texte codé, il est possible d'utiliser des balises additionnelles pour indiquer des séparations plus spécifiques. Dans le cas d'un texte poétique, il conviendra par exemple de segmenter à chaque fin de vers avec des marqueurs bien identifiés.

De nombreux autres éléments autour du texte peuvent aussi être annotés, afin de conserver le maximum d'informations par rapport au document original, principalement quand celui-ci est issu d'un livre. Ainsi, les changements de pages peuvent être codés à l'intérieur du texte, même s'ils n'ont bien entendu plus aucune valeur « physique », à l'aide de la balise <pb>. Il est même possible de lui adjoindre les balises <n> et <ed> pour indiquer respectivement le numéro précis de la page, ainsi que l'édition dans laquelle cette numérotation est effective. Ces données peuvent toutefois être lues et servir par exemple à une présentation ou une édition du texte, en fonction des besoins des utilisateurs.

Comme dans un traitement de texte conventionnel, la TEI permet également de mettre en valeur les caractères gras ou italiques ou soulignés (balise <rend> suivi de l'attribut souhaité), ainsi que de spécifier des polices de caractères précises *via* le marqueur <hi>. Des mises en valeur rhétorique ou linguistique peuvent également être codées, évoquant ainsi certaines balises du logiciel TRANSCRIBER, déjà évoquées précédemment en 3.1.

La TEI permet également de rapprocher le codage d'un travail éditorial à venir, en mentionnant les erreurs (supposées ou réelles) d'un texte original, les oublis, les corrections ou ajouts apportés par l'annotateur. Ces annotations, sous formes de balises, constituent des « couches » qui peuvent être conjointement codées, un peu comme dans les éditions d'ouvrages qui proposent des versions manuscrites d'un texte. On peut par exemple y voir « en temps réel » le travail d'un écrivain, jusqu'à la version définitive de son travail. En voici un exemple concret, avec la phrase suivante :

« … for his nose was as sharp as a pen and a' table of green feelds » (Shakespeare).

Selon l'éditeur Clive Gifford, cette version du texte de Shakespeare comporte deux erreurs : *table* est substitué de façon erronée à *babbl'd*, et les orthographes *a'* et *feelds* doivent être normalisées en *he* et *fields*. Si l'on code ces modifications en TEI, voici ce que cela donnera :

... for his nose was as sharp as a pen and <reg sic="a'">he</reg>
<corr sic='table' ed=Gifford>babbl'd</corr> of green <reg sic='feelds'>fields</reg>

<reg sic> indique ici une forme qui a été régularisée (à savoir *a'*, devenu *he*, et *feelds* qui se transforme en *fields*). <corr sic> donne quant à lui la forme originelle d'une erreur supposée (*table*), ainsi que la forme jugée correcte (*babbl'd*).

La TEI comporte encore bien d'autres attributs généraux, qui dépassent parfois le simple cadre textuel. On peut par exemple coder également des graphiques ou des figures, moyennant quelques restrictions d'usage (formats compatibles, précisions à apporter au niveau de la DTD). Après ce rapide survol explicatif, nous allons nous intéresser à l'application de la TEI dans le cadre de notre étude, à savoir le codage de la parole.

4.2.4.2 Le codage de la parole selon la TEI

Nous avons déjà longuement évoqué la transcription de la parole (et notamment la parole spontanée), mais finalement assez peu la manière dont elle pouvait être codée. Il s'agit en quelque sorte de la partie immergée de l'iceberg. Bien qu'étant invisible durant tout le processus de transcription en lui-même, celle-ci revêt la plus grande importance une fois l'annotation terminée. Si l'on souhaite pouvoir échanger, compléter ou réutiliser ces données, il faut qu'il existe entre elles une « passerelle » d'unification puisque, nous l'avons vu, il n'y a pas *une* façon universelle de transcrire et d'annoter. La TEI, que nous venons de présenter, semble donc tout à fait correspondre à cette tâche eu égard à ses vertus normalisatrices. Qui plus est, un chapitre spécifique des *Guidelines* de la TEI est consacré au codage de la parole, et c'est sur celui-ci que nous allons nous appuyer pour en détailler les particularités.

En premier lieu, tout document codé avec les normes de la TEI obéit à des règles

communes, et la parole n'y échappe pas. Ainsi, le *header* dont nous avons parlé sera toujours présent ici, mais avec quelques précisions supplémentaires concernant les conditions d'enregistrement des données, leur durée, les locuteurs présents... Beaucoup de balises « fines » sont prévues pour être en mesure de coder nombre de, notamment au sujet de l'enregistrement : le matériel utilisé, la fréquence du signal ou encore le responsable de l'enregistrement peuvent ainsi être mentionnés.

Naturellement, coder de la parole transcrite, c'est-à-dire une représentation codifiée de la langue parlée, ne revient pas à coder un document-texte, c'est-à-dire de l'écrit, et la codification utilisée n'est pas de même nature. Si nous évoquions plus haut la balise <p> pour coder un paragraphe, elle ne trouve plus sa justification ici puisque les données traitées relèvent du code oral. On parlera donc plutôt d'*utterance* (segment de parole), terme auquel correspond la balise <u>, ce qui reflète qu'à l'oral il n'y a plus ni phrases, ni paragraphes, ni pages, et qu'il convient de référer à des unités d'une autre nature (ce qui n'interdit pas de les utiliser, par exemple pour de la parole préparée, manifestement ponctuée comme à l'écrit, comme sur un prompteur). Comme en utilisant un logiciel de transcription, il est possible d'assigner un locuteur à ce segment de parole. A cet effet, il convient d'ajouter l'attribut « who », puis d'indiquer l'identité souhaitée sous la forme suivante :

<u who="Nicolas Demorand">

Le type de transition entre deux tours de parole peut être indiqué avec l'attribut « trans », qui trouvera également sa place à l'intérieur de la balise <u>. Il existe quatre façons principales de coder la transition :

<u trans="smooth"> : transition naturelle, sans pause significative

<u trans="pause"> : transition avec pause significative

<u trans="latching"> : transition brutale, où un locuteur en interrompt un autre ; il n'y a cependant pas chevauchement des deux flux de parole

<u trans="overlap"> : chevauchement avec le flux de parole précédant cette balise

Ainsi, si l'on imagine un segment de parole dont l'auteur est Nicolas Demorand, et à supposer que ce dernier marque une pause avant de parler, cela donnera la représentation

suivante :

<u trans="pause" who="Nicolas Demorand">eh bien merci pour ce beau témoignage</u>

Il est même possible, pour peu que l'annotateur souhaite un niveau de granularité supérieur (à l'intérieur même d'un segment de parole), d'utiliser la balise <seg>. Cette dernière permet ainsi des annotations qui portent sur des composants syntaxiques ou des phénomènes prosodiques : isoler des syntagmes suivant l'intonation qui les conclut, les classer selon la vitesse à laquelle ils sont prononcés, etc.

Cette balise <u> et ses différents attributs (que nous n'avons pas tous référencés ici) est un des chaînons essentiels pour coder de la parole selon la TEI. Il existe également tout un jeu de balises relatives aux phénomènes linguistiques ou extra-linguistiques d'une situation énonciative :

- <pause/> : pause entre deux segments de parole
- <vocal> : bruits de bouche et autres phénomènes vocalisés non lexicaux
- <kinesic> : gestes et autres phénomènes de communication, pas forcément vocalisés
- <incident> : bruits de fond et autres phénomènes perturbant la communication
- <shift/> : point de départ d'une modification énonciative, essentiellement prosodique.

On retiendra notamment :

- la vitesse d'élocution (rapide, très rapide, lente, très lente, progressive, dégressive...)
- le volume sonore (fort, très fort, faible...)
- l'intonation (montante, descendante, monotone...)
- le rythme (régulier, saccadé...)
- le type de voix (essoufflée, tremblante, pleurante...)

Par ailleurs, la plupart des phénomènes que nous avons nommés « disfluences » [Adda *et al.*, 2005 ; Adda-Decker *et al.*, 2004] dès l'introduction de cette étude trouvent également dans la TEI des solutions de codage :

- troncation : <del type="truncation">pour</del>pourtant
- répétition : <del type="repetition">mais mais</del> mais vous savez
- faux départ : <del type="falseStart">c'est vrai que</del>enfin, je dois dire

Le choix que préconise la TEI pour ces trois cas de figure est intéressant. Outre la balise « del » et son attribut « type », on observe systématiquement que la disfluence en elle-même est isolée, permettant ainsi de visualiser facilement la version « propre » de l'énoncé. Pour autant, *isolée* ne signifie pas *inexploitable*. Grâce aux balises, chacune d'entre elles est aisément repérable, que ce soit pour en produire une représentation, pour permettre de parcourir les données, pour créer un sous fichier....

Une fonctionnalité similaire, mais qui utilise la balise <choice>, permet de mentionner un mot incorrectement prononcé par un locuteur, et d'indiquer conjointement la prononciation exacte (un peu à la façon de TRANSCRIBER avec la balise « pron »). Celle-ci se rapproche de l'exemple du texte de Shakespeare que nous avons mentionné en 4.2.4.1, même s'il n'est cette fois nullement question de choix éditoriaux :

<choice>
  <sic>omnibulé</sic>
  <corr>obnubilé</corr>[34]
</choice>

Enfin, pour terminer cette description de la TEI, nous allons nous intéresser en détail au codage de la parole superposée, et par là même à celui de la temporalité. En effet, l'exemple que nous avons donné tout à l'heure avec la balise <overlap> ne permet d'indiquer précisément quel part du segment premier segment est recouvert par le deuxième. Pour pallier ce problème, il existe différentes possibilités. L'une d'entre elles consiste à ajouter dans le *header* un module autorisant le codage de liens « visuels » entre les différents énoncés superposés. Dans ce cas, il n'est donc toujours pas question de temporalité. En voici un exemple :

<u xml:id="utt1" who="Nicolas Demorand"> et vous pensez vraiment <anchor

34 *sic* représentant la prononciation « telle quelle », et *corr* sa version académique

synch"#utt2" xml:id="a1"/>que ce livre</u>

<u xml:id="utt2" who="Invité">ah oui</u>

Une partie du propos de Nicolas Demorand (utt1) est chevauché par un « ah oui » (utt2). On insère donc une balise <anchor> pour isoler la partie chevauchée dans utt1 (« que ce livre »), en y indiquant la référence à utt2. Enfin, on donne également un code d'identification à ce chevauchement : ici, a1.

Lorsque des chevauchements plus complexes sont à coder, il convient alors d'utiliser une balise temporelle que l'on appelle, dans la nomenclature TEI, <timeline>. Il faut ensuite y adjoindre différentes balises <when>, correspondant chacune à un point-temps précis. Ensuite, des informations indiquant l'emplacement temporel des balises <when> les unes par rapport aux autres permettent de synchroniser l'ensemble comme on le souhaite :

<**timeline** unit="s" origin="#TS-P1">
<**when** xml:id="TS-P1" absolute="12:20:01"/>
<**when** xml:id="TS-P2" interval="4" since="#TS-P1"/>
<**when** xml:id="TS-P6"/>
<**when** xml:id="TS-P3" interval="1" since="#TS-P6"/>
</**timeline**>

Dans l'exemple ci-dessus, inspiré de celui des *Guidelines*, on repère tout d'abord que l'unité temporelle pour les intervalles est la seconde (s), et qu'il y a quatre points-temps : TS-P1, TS-P2, TS-P6 et TS-P3. TS-P1 est le point de base, situé à 12:20:01. TS-P2 est placé 4 secondes plus loin, donc à 12:20:05. On ne dispose d'aucune information temporelle concernant TS-P6, sinon qu'il est situé quelque part entre TS-P2 et TS-P3, qui lui se trouve une seconde après TS-P6.

Dans la pratique, un tel codage peut donner le résultat suivant :

<**timeline** unit="s" origin="#TS-P1">
<**when** xml:id="TS-P1" absolute="07:24:18"/>
<**when** xml:id="TS-P2" interval="2" since="#TS-P1"/>
</**timeline**>

<**u** who="Nicolas Demorand">et vous pensez vraiment
<**anchor** synch="#TS-P1"/>que ce livre
<**anchor** synch="#TS-P2"/>est tout public</**u**>
<**u** who="Invité">
<**anchor** synch="#TS-P1"/>ah oui<**anchor** synch="#TS-P2"/>
</**u**>

Il est donc ici possible, grâce au codage, d'indiquer avec précision quels sont les segments qui se coupent (« que ce livre » et « ah oui »), ainsi que le point de départ de la superposition (situé à 07:24:18) et son point d'arrivée (situé à 07:24:18 + 2 donc 07:24:20). Bien sûr, en multipliant les balises, il est possible d'apporter des informations temporelles sur tous les segments souhaités, et pas seulement ceux qui sont superposés.

Il existe encore bien d'autres possibilités de codage pour représenter le phénomène de la superposition, mais celles que nous avons envisagées nous semblent les plus efficaces, et surtout les moins complexes à mettre en place.

4.3 Conclusion

La parole nous semble trouver avec la TEI une base de poids pour appréhender son codage. Il est d'ailleurs important de bien garder à l'esprit la différence entre transcription/annotation et codage : notre propos n'est nullement d'encourager les transcripteurs à baliser manuellement leurs transcriptions comme nous l'avons fait dans les quelques exemples ci-dessus. Il consiste plutôt à montrer que, quel que soit le logiciel et les conventions de transcriptions adoptées pour le travail de transcription « à l'écran », il serait sans doute pertinent d'envisager un format commun en ce qui concerne le fichier de sortie. Ce format annihilerait probablement les problèmes d'échange et de lisibilité des corpus auxquels sont actuellement confrontés les transcripteurs désireux d'utiliser d'autres ressources que celles qu'ils ont eux-mêmes créées. Il est sans doute utopique d'envisager une totale homogénéité des données transcrites, au point qu'on puisse les ouvrir toutes sans souci d'affichage, avec n'importe quel logiciel. Le travail de normalisation est en effet énorme, et si certains groupes de recherche comme ICOR ont commencé des démarches allant dans ce sens, la route est encore (très) longue. Il y a en effet un nombre gigantesque de corpus disparates au niveau codage, la faute à leur date de création ou à un codage « maison ». Cela étant, décider d'un instant T à partir duquel tous les projets de corpus s'appuierait désormais sur le même format de sortie serait sans doute une très bonne chose. Le simple fait de pouvoir effectuer de la fouille de données dans des fichiers tous structurés de la même façon, en utilisant n'importe quel éditeur de texte, constituerait déjà une réelle avancée.

Bien que parfois imparfaite – le codage de la parole superposée est ardu à déchiffrer – ou incomplète (aucune balise ne semble par exemple être prévue pour expliciter en API ou en caractères SAMPA une prononciation particulière), la TEI envisage sans doute plus de possibilités de codage qu'un transcripteur n'en utilisera jamais dans ses annotations. À ce titre, elle est finalement bien plus qu'une simple démarche normalisatrice : elle va aussi et surtout beaucoup plus loin, dans le domaine de la parole, que toutes les conventions de transcription existantes. En d'autres termes, la TEI doit permettre non seulement de partager facilement des corpus mais aussi de les enrichir de façon considérable par la suite, au gré des besoins de chacun.

Chapitre 5 : La parole spontanée

Avant 1950, avant le *Français fondamental* [Gougenheim *et al.*, 1964], avant le magnétophone, l'existence du versant oral de la langue n'était pas passée totalement inaperçu. La littérature tout d'abord s'en était faite le témoin et l'acteur. De ce point de vue, on peut commencer par l'ancien français, dont bon nombre de textes sont fortement oralisés, à une époque où l'écrit de référence était le latin. Par la suite, La Fontaine (dont la fraîcheur tient en partie à une forme d'oralité du style) puis quelques grandes pages du 19ème siècle ont contribué à forger notre représentation de l'oral, trop souvent confondu avec l'argot et le registre familier. Les contes normands de Maupassant et *Les Misérables*, avec ses dialogues et son chapitre sur l'argot, en sont les exemples les plus patents. Au 20ème siècle, cette oralité réservée à certains discours rapportés a laissé place à une oralité assumée par le narrateur. *Voyage au bout de la nuit* et *Zazie dans le métro* ont ainsi ouvert la voie, et *Le petit Nicolas* a suivi. Cela étant, dans l'oeuvre de Sempé et Goscinny, l'argument de l'oralité devient la spontanéité de la pensée d'un enfant, et non plus une relation virile à une féroce réalité.

Durant cette première partie du 20ème siècle, comme nous l'avons mentionné dans notre deuxième chapitre, certains linguistes disposaient déjà de quelques corpus, ou plus exactement de quelques images d'usages oraux du français. Les théoriciens, et notamment Saussure, sont venus ajouter à ces travaux précurseurs (Damourette et Pichon, Frei, Bally...) quelques concepts devenus aujourd'hui bases de référence. Citons particulièrement l'opposition saussurienne entre « langue » et « parole », ou encore la théorisation de la fonction interactive du langage

En somme, avant le *Français fondamental*, l'accès à l'oral se faisait par des biais. La théorie en est un, au moins dans une certaine mesure. Les écrits non normés qu'étudie Frei en sont un autre. Les énoncés « à la volée » constituent un corpus à la fois subjectif et fragmentaire (20 secondes de parole au maximum, comme indiqué dans [Blanche-Benveniste, 2010]). La littérature enfin utilise l'oralité, la déforme et la représente pour la dévier.

Avec le français fondamental, les chercheurs ont enfin eu accès à de véritables données,

certes rudimentaires, mais néanmoins réelles. L'objectif à cette époque n'était pas exactement l'« oral », la « parole spontanée », ou le « français parlé ». Il s'agissait d'établir un « basic french », c'est-à-dire une version pour le français du « basic English » [Ogden 1932], à partir de données attestées, et non à partir d'une image abstraite et idéologiquement marquée. La rencontre avec l'oral était néanmoins inéluctable, dont témoigne notamment Aurélien Sauvageot dans son ouvrage intitulé *Français écrit, français parlé* [Sauvageot, 1962].

Un centre de recherche dédié à l'oral s'est ensuite constitué à Aix-en-Provence, notamment animé par Claire Blanche Benveniste : le GARS (Groupe Aixois de Recherches en Syntaxe). Il a érigé l'objet « français parlé » comme objet de recherches, et nombreux sont les chercheurs du domaine qui en sont issus, comme [Bilger et Cappeau, 2004] ou [Deulofeu, 1977 et 2001].

Claire Blanche-Benveniste et Colette Jeanjean [Blanche-Benveniste et Jeanjean, 1987] ont publié en 1987 un premier bilan des travaux réalisés dans le domaine du « français parlé ». À cette époque, aucun état de l'art précis n'avait été réalisé sur le sujet, sans doute parce que, comme elles l'écrivent, « peu de gens y [voyaient] un objet légitime d'étude, même chez les linguistes ». Malgré, entre autres, la grammaire atypique de Damourette et Pichon au début du vingtième siècle [Damourette et Pichon, 1911-1940][35]. Avant l'avènement du magnétophone, dans les années 60, peu de chercheurs s'étaient risqués à collecter des données sur le terrain, pourtant seul véritable juge de paix en la matière. Il en a résulté des études assez incomplètes, dans lesquelles les auteurs « en [arrivaient] vite à dire des sottises », du fait notamment d'une opposition jugée « simpliste » entre *parlé* et *écrit*. En effet, pour Claire Blanche-Benveniste et Colette Jeanjean, ces deux appellations sont à regrouper sous une même bannière : celle de la langue. Et à l'intérieur même du *français parlé*, elles en distinguent même plusieurs types, afin d'éviter les conclusions hâtives et réductionnistes (par exemple la disparition de « vous » à l'oral).

C'est sous cet angle que nous avons voulu envisager notre analyse de la « parole spontanée », terme qui a émergé dans le cadre des données du projet EPAC, par opposition à la « parole préparée ». Ce sont certes des concepts contestables, mais ils reposent sur une base difficilement discutable : comme nous le verrons, selon les zones de parole auquel il est confronté, un système de RAP affiche des taux d'erreurs mot qui peuvent varier du simple au triple. Or, dans le corpus ESTER, de tels écarts ne sont que très rarement la conséquence de

35 Ils n'étaient pas moins atypiques que leur grand œuvre : avant de devenir linguistes, Damourette était architecte et Pichon médecin...

problèmes purement acoustiques (mauvaise qualité d'enregistrement, environnements bruyants, etc.). De façon schématique, il existe donc bien, à l'intérieur même du cadre de l'énoncé, deux types de parole : une qui est correctement reconnue par le SRAP, et une qui ne l'est pas. Et, entre ces deux extrêmes, plusieurs tendances intermédiaires...

La « spontanéité » dont il est question a pour effet, de façon non exclusive, de faire disparaître les schwas, de provoquer des assimilations en cascade, de multiplier les phénomènes de bruit *(fillers* et autres disfluences), les phénomènes de dislocation (énoncés fragmentaires), et les phénomènes d'interaction (énoncés brefs, co-construits en ce sens qu'ils sont construits sur plusieurs tours de parole). Souvent le contenu est faiblement référentiel, avec des une faible proportion de mots signifiants de 2 syllabes ou plus. Nous aurons d'ailleurs l'occasion d'illustrer cette affirmation dans la première partie du présent chapitre.

La « parole préparée », à l'inverse, obtient des taux de reconnaissance bien supérieurs[36]. Elle a tendance à limiter ces phénomènes, et s'apparente donc davantage à l'écrit. Il n'est qu'à voir les nombreux liens qu'entretiennent écrit et parole préparée : lors d'un journal d'informations, le présentateur lit soit un prompteur, soit des notes qu'il a devant lui. Dans un cas comme dans l'autre, il y a bien un support *écrit* sur lequel le locuteur s'appuie. Il en va de même pour les acteurs de cinéma ou de théâtre, qui eux aussi ont souvent recours à la parole préparée : même si lors d'un tournage ou d'une représentation il n'ont pas de support écrit « visible » avec eux, ils s'en sont servi en amont pour apprendre et mémoriser leurs textes. On peut même considérer que quand ils produisent de la parole « spontanée », fortement disfluente, ils la jouent davantage qu'ils ne l'improvisent.

Il arrive bien sûr que la parole préparée « dérape » : journaliste qui bégaie, acteur qui oublie son texte... Dans ces situations, le discours devient à son tour « bruité », des disfluence apparaissent et les performances des SRAP s'en ressentent.

Cela étant, ces performances sont parfois dépendantes de phénomènes extra-linguistiques. L'un des plus fréquents est la mauvaise qualité du signal audio. Nous avons ainsi parfois été confrontés, dans le corpus EPAC, à des journaux d'informations où un envoyé spécial intervenait à distance *via* une ligne téléphonique dont la qualité n'est pas optimale. Le discours du journaliste est certes fluide (le plus souvent, il lit un texte déjà écrit), mais il est à nouveau « bruité », cette fois par des interférences ou des coupures de signal. Les scores obtenus par les SRAP dans ces circonstances sont alors proches de ceux de la parole

36 Voir l'expérience menée dans le chapitre 2.2.

spontanée, alors qu'aucune disfluence n'est venue perturber le propos.

Précisons toutefois que ces cas demeurent marginaux et que, nous aurons l'occasion de le voir dans en 5.1, la mesure du taux d'erreur mots par les SRAP demeure un indice fiable de la spontanéité ou non d'un segment de parole.

Certes, cette opposition binaire va à l'encontre des théories de [Blanche-Benveniste, 1987] ou [Biber, 2006] notamment. C'est donc là que se situe la frontière entre reconnaissance automatique de la parole et linguistique. Là où une perspective théorique dégage diverses variations du discours, les résultats d'un SRAP peuvent en général se résumer comme suit : un débat (spontané) obtient des scores d'erreur deux à trois fois plus élevés qu'un journal d'informations (préparé). Les raisons en sont clairement identifiables et en grande partie quantifiables, comme nous l'avons vu quelques paragraphes plus haut (disparition des schwas, assimilations, *fillers*...).

En tout état de cause, il est clair que *parole spontanée* et *parole préparée* constituent deux versants parmi d'autres du *français parlé* tel qu'il est défini par C. Blanche-Benveniste et C. Jeanjean. Chaque extrait de corpus relève des deux, dans des proportions variables, et chaque prise de parole s'inscrit dans une palette de variations [Gadet, 2007]. Chaque locuteur les pratique en utilisant sa *langue du dimanche* ou sa *langue de tous les jours*. Chaque auditeur peut trouver en fonctions de critères qui lui sont propres que tel ou tel extrait de *parole spontanée* comme de *parole préparée*, peut relever d'un discours familier ou d'un discours soutenu, rejoignant en cela Bally pour qui un syllogisme socratique pouvait parfaitement relever d'un discours « populaire » :

Un fait de langage qui reflète un état social supérieur ou une forme d'activité ou de pensée plus haute que celle du commun, appartient à la langue dite écrite ; mais il faut s'entendre tout de suite sur ce terme. Une expression n'a nullement besoin d'être écrite pour porter la marque de cette forme générale ; elle conserve ce caractère et même le montre mieux encore quand elle est employée dans le parler… Si un fait de langage permet de constater l'absence de toute différence sociale entre les sujets parlants, ou bien si ce fait de langage révèle un état social inférieur, ou une forme de pensée et d'activité au-dessous de la mentalité commune, dans tous les cas, on a affaire à la langue dite parlée ou langue de la conversation. Ce mode d'expression comporte à son tour de nombreuses variétés, depuis le ton simplement familier, qui en est le caractère essentiel, jusqu'à la langue dite populaire,

l'expression vulgaire et l'argot le plus grossier …

La langue parlée dirait : « les homme sont mortels, n'est-ce pas ? bon, ! Mais Paul n'est-il pas un homme ? Oui ? eh bien ! Alors il doit être mortel. » [Bally, 1929]

Cette diversité d'appréciation tient également aux buts poursuivis : description morphosyntaxique pour Claire Blanche-Benveniste [Blanche-Benveniste, 2010], analyse prosodique et syntaxique pour Mary-Annick Morel et Laurent Danon-Boileau [Morel et Danon-Boileau, 1998] ou A. Lacheret [Lacheret, 2007], approche sociolinguistique pour F. Gadet [Gadet, 2007]... Les données que l'on invoque sont enfin issues de deux champs disciplinaires souvent étrangers, à l'exception notamment de [Luzzati, 2004] : la linguistique, qui poursuit une entreprise de description, et l'informatique, *via* la reconnaissance de la parole, qui poursuit un objectif technologique. Les données les plus récentes, recueillies dans le cadre de projets ANR par exemple, sont clairement fléchées : soit il s'agit de projets corpus (VARILING, RHAPSODIE...), soit il s'agit de projets masse de données (EPAC, PORT MEDIA...). Pour le dire autrement, les uns relèvent des sciences humaines et les autres, des sciences de l'information.

Notre travail est quant à lui un travail d'ingénierie linguistique. Il s'inscrit dans ce dernier cadre, et les concepts retenus (parole spontanée / parole préparée) sont pertinents dans ce cadre, même s'il peut être intéressant de les discuter d'un point de vue linguistique. Il s'agit alors d'un autre travail, et il faut prendre des précautions avant de se livrer à des extrapolations qui deviennent vite sommaires.

Ainsi, le simple fait d'employer l'adjectif *spontanée* est déjà une prise de position, ou tout du moins une orientation, une tendance. Cette terminologie n'est en effet pas universelle, puisqu'autour d'elle gravitent des termes comme *conversationnelle*, *co-construite*, *non préméditée, familière, interactive*...

Pour les besoins de cette thèse et des travaux qui ont permis sa réalisation, notre choix s'est tourné vers l'expression *parole spontanée* non pas parce qu'elle serait plus « vraie » qu'une autre mais parce que le terme *spontanée* sous-entend deux distinctions pour nous essentielles :

- le caractère naturel de l'énoncé. Ce dernier n'a pas été pensé, suggéré ou même lu avant sa production par un locuteur. Plus concrètement, dans les données radiophoniques (et parfois télévisuelles) que nous avons traitées, les animateurs n'avaient en général ni textes, ni

notes, ni prompteurs.

- la distance prise avec la notion d'interaction. Là où le mot *conversationnelle* suppose qu'il y ait systématiquement une *conversation* entre au moins deux interlocuteurs, *spontanée* nous a semblé être plus à même de caractériser certaines situations d'énonciation « naturelles » du corpus EPAC, où l'interaction entre participants demeure pourtant très discrète. Nous aurons l'occasion de développer ce point dans la dernière partie du présent chapitre.

Afin de proposer une vision globale de notre démarche, nous avons choisi de présenter dans un premier temps une expérience qui aura pour principal atout de confronter les critères linguistiques et informatiques que nous avons évoqués dans cette introduction. Pour ce faire, nous avons demandé à deux annotateurs d'évaluer le « degré de spontanéité » d'une partie du corpus ESTER. Un protocole d'annotation précis a été défini, et l'accord inter-annotateurs a été mesuré au terme de cette expérience. Les annotations ont ensuite été mises en relation avec les performances d'un système de reconnaissance automatique de la parole, LIUM RT, avec un enjeu simple : voir si les performances des SRAP sont proportionnelles au nombre de disfluences[37] contenues dans un tour de parole. Ce type d'approche descendante nous offrira la possibilité, données à l'appui, de vérifier si les définitions linguistiques et informatiques peuvent cohabiter, et surtout si elles recouvrent bien un même objet.

Cette étude liminaire sera également une passerelle idéale pour s'intéresser plus largement aux spécificités et définitions qui permettent de circonscrire la parole *spontanée*. Qu'elles soient morpho-syntaxiques, lexicales, phonologiques ou prosodiques, toutes permettent de distinguer, d'une façon ou d'une autre, un pan de ce que recouvre cette appellation – et, par opposition, de la distinguer de la parole *préparée*.

Nous poursuivrons notre analyse en présentant une étude morpho-syntaxique de la parole spontanée, réalisée grâce aux transcriptions et annotations du corpus EPAC. Elle portera sur l'analyse des disfluences, en comparant les discours de plusieurs « interviewés » placés dans une même situation énonciative, et confrontés au même intervieweur. Il s'agira notamment de repérer où se trouvent les pics de disfluences, et de voir si celles-ci sont un critère systématiquement pertinent pour caractériser la parole spontanée. Nous proposerons à cette occasion une définition précise de ce que recouvre pour nous le terme « disfluence ».

Enfin, nous évoquerons le lien entretenu par la parole spontanée avec l'interaction. En

37 Voir les définitions du terme « disfluence » données en 5.1

effet, si la parole spontanée est aussi souvent appelée « parole conversationnelle », c'est à notre avis parce qu'il s'y trouve une notion sous-jacente de *conversation*, c'est-à-dire d'échange entre un locuteur et (au moins) un interlocuteur. Il serait réducteur, et sans doute même fallacieux, de placer sous une même bannière une interview proche du monologue et un débat particulièrement animé. C'est ce que nous tenterons de démontrer, chiffres et statistiques à l'appui, en comparant plusieurs émissions radiophoniques répondant pourtant à de nombreux critères « spontanés ». Nous nous intéresserons tout particulièrement au programme *Sous les étoiles exactement* qui, au gré des invités reçus, offre un exemple patent de variations interactionnelles.

Par ailleurs, nous inclurons dans cette étude un corpus de transcriptions d'émissions télévisées, afin de voir si le support visuel modifie d'une quelconque façon la gestion de l'interaction entre les intervenants.

5.1 La parole spontanée : des chiffres pour établir une théorie ?

Depuis l'aventure du français fondamental [Gougenheim *et al.*, 1964], les approches linguistiques pour circonscrire la *parole spontanée* (ou toute autre appellation lui faisant écho) ne manquent pas, allant du *français parlé* à la *parole interactive*. Les points de vue divergent souvent : énonciation, morpho-syntaxe, phonologie, prosodie... sont autant d'angles d'approche possibles. Un concept récent, issu de chercheurs qui travaillent en reconnaissance de la parole, nous semble être une approche intéressante et nouvelle en vue de caractériser la *parole spontanée* : les disfluences.

D'après [Blanche-Benveniste *et al.*, 1990], une disfluence est un endroit dans un énoncé où « le déroulement syntagmatique est brisé ». [Guénot, 2005] ne dit pas autre-chose : selon elle, lorsqu'il y a disfluence, « on occupe une même place syntaxique avec plusieurs objets ». La figure 29 se propose de représenter cette définition sous une forme librement inspirée des « grilles » de [Blanche-Benveniste, 1987].

à travers les siècles et à travers | **le**
à travers les siècles et à travers | **la** critique en passant | **de**
à travers les siècles et à travers | la critique en passant | **de r/**
à travers les siècles et à travers | la critique en passant | **par**
à travers les siècles et à travers | la critique en passant | **par**
à travers les siècles et à travers | la critique en passant | **par** Rousseau, par tous les gens | **qui**
à travers les siècles et à travers | la critique en passant | par Rousseau, par tous les gens | **qui** ont pu | **pa/**
à travers les siècles et à travers | la critique en passant | par Rousseau, par tous les gens | qui ont pu | **parler** | **de**
à travers les siècles et à travers | la critique en passant | par Rousseau, par tous les gens | qui ont pu | parler | **de** ce personnage

Figure 29 : Représentation en relief du phénomène de disfluence

Les caractères gras indiquent les zones de disfluence en elles-mêmes. Comme on le voit, celles-ci peuvent ici être des répétitions (*de, qui*), des troncations (*pa/*, *r/*) ou encore ce que nous appelons dans cette étude des faux départs (*le* / *la*). Le message que souhaite transmettre le locuteur est le suivant :

« bon à travers les siècles et à travers la critique, en passant par Rousseau, par tous les gens qui ont pu parler de ce personnage »

Or on s'aperçoit, pour paraphraser Blanche-Benveniste et Guénot, que le déroulement syntagmatique est effectivement brisé, et qu'il y a jusqu'à cinq objets (l'accumulation des prépositions *de* et *par)* qui occupent une seule et même place syntaxique. Au final, il faut donc cinq places syntaxiques occupées par treize objets « disfluents » pour que le message soit intégralement transmis.

Pour en revenir à la caractérisation des disfluences, [Guénot, 2005] emploie pour sa part une terminologie binaire, regroupant les répétitions et faux départs sous l'appellation *bribes* (reprises à partir de syntagmes inachevées), qu'elle oppose à *amorces* (synonyme de *troncations* ici). Ce terme *amorces* est également employé par Berthille Pallaud [Pallaud, 2004 et 2006], dont nous évoquons les travaux en 5.2.2.2.

[Adda-Decker *et al.*, 2004] ajoute à cette liste les notions d'*hésitations* et de *lapsus*, mais s'appuie aussi et surtout sur les conventions d'annotation du Linguistic Data Consortium[38], qui distingue entre autres la catégorie des *filler words* (des « remplisseurs » tels que *euh* et *ben* en français).

Pour les besoins de cette thèse, nous avons choisi de mettre sous l'appellation « disfluences » quatre des concepts cités ci-dessus (répétitions, troncations, *fillers* et faux départs, terme dont nous préciserons notre définition en 5.2.2.3), auxquels nous avons ajouté :

- les assimilations découlant de l'élision des schwas.
- les mots mal prononcés, à bien distinguer des phénomènes d'assimilation : pour nous, un mot *mal prononcé* n'est pas identifiable en tant que tel. A l'inverse, les prononciations [Sfal] pour « cheval » ou [Spa~s] pour « je pense » ne sont pas considérées comme ambiguës, car solidement ancrées dans l'usage de la langue.
- la parole superposée

Une telle diversité nous a ainsi permis, tout au long de nos recherches, d'aborder différentes thématiques :

- énonciation (parole superposée, ou parfois simple interruption d'un locuteur par un

38 http://projects.ldc.upenn.edu/MDE/Guidelines/SimpleMDE_V6.2.pdf

autre)

- morpho-syntaxe (troncations, faux départs, répétitions)
- phonologie (élisions des schwas et assimilations)
- lexique (*fillers* comme « ben » et « euh »)

Les remarques précédentes pourraient donner au lecteur l'impression que les disfluences ne s'étudient que sur des corpus de langue française : il n'en est rien. Non seulement des francophones étudient les disfluences dans d'autres langues telles que l'arabe [Bahou *et al.*, 2010], mais le terme anglais *disfluency* est également fréquemment employé, dans le même domaine et avec la même acception : Elizabeth Shriberg a par exemple soutenu en 1994 une thèse intitulée « Detecting disfluency in spontaneous speech » [Shriberg, 1994]. Par ailleurs, comme Daniel Luzzati s'était intéressé il y a près de trente ans au *filler* français « ben » [Luzzati, 1982], Corley et Stewart ont récemment publié une étude sur la signification d'un de ses alter ego britanniques, « um » [Corley et Stewart, 2008].

L'étude suivante a donc principalement été fondée sur l'analyse des disfluences, ce qui l'éloigne par exemple de travaux comme ceux de [Morel et Danon-Boileau, 1998], qui reposent pour leur part sur des considérations prosodiques. Dans notre cadre expérimental, une orientation morpho-syntaxique et lexicale était en effet plus simple à mettre en œuvre, notamment par rapport au protocole et au système de notation que nous souhaitions utiliser (voir ci-dessous en 5.1.1). Toutefois, la frontière entre morpho-syntaxe et prosodie étant parfois très mince, nous ferons de temps en temps appel à des notions comme la régularité ou l'irrégularité du débit.

En quelques mots, l'objectif de cette étude était de voir si des terminologies linguistiques telles que *troncations* ou *assimilations*, définissant de façon très ouverte ce que peut être la parole spontanée, pouvaient trouver un écho informatique : celui de la reconnaissance automatique de la parole.

Ainsi, notre hypothèse de départ était la suivante : l'apprentissage de LIUM RT ayant été réalisé avec des données de parole « propre » (en quelque sorte, de la parole *préparée*), son taux d'erreur mot devrait être plus élevé sur des segments comportant de nombreuses disfluences. Il convenait pour cela de trouver une façon de quantifier ces disfluences, ce qui a été fait grâce à la collaboration de deux annotateurs. La section suivante détaille avec précision ce processus d'annotation.

5.1.1 Protocole

Pour les besoins de cette expérience, 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 la transcription elle-même en a été supprimée. Deux annotateurs étaient donc chargés de noter chaque segment de parole suivant une échelle numérique allant de 1 à 9. 1 était la note attribuée à un segment sans aucune disfluence, avec une élocution parfaitement claire. 9 indiquait un segment difficilement interprétable tant les hésitations, répétitions, troncations, *fillers* et faux départs 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 être sûrs qu'ils utilisaient bien les mêmes critères de notation. Ensuite le coefficient Kappa [Cohen, 1960] a été calculé pour valider le processus. Il s'agit d'une formule mathématique permettant de mesurer l'accord inter-annotateurs. Un score de 0.852 a été obtenu, sachant que les scores dépassant 0.81 sont considérés comme étant excellents. Cela prouve que, sans avoir au préalable défini précisément ce qu'était la *parole spontanée*, les deux annotateurs étaient pour autant capables de tomber largement d'accord sur ce qu'elle représentait. Plus important encore, le coefficient Kappa a montré qu'au delà d'une simple distinction spontanée / préparée, les annotateurs étaient en mesure de s'accorder également sur les différents *degrés de spontanéité*, véritable fondement de l'expérience. Autrement dit le concept de *parole spontanée*, tout en demeurant à bien des égards flou, n'en est pas moins parfaitement objectivable, du moins dans le contexte d'un laboratoire d'informatique travaillant sur la reconnaissance de la parole. Dans un tel contexte, le problème qui se pose est la difficulté à conserver des taux de reconnaissance satisfaisants dès lors qu'un ensemble de phénomènes discrets prennent une importance suffisante pour être identifiables par divers annotateurs.

Afin de vérifier si ces annotations conjointes allaient 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 d'abord été

mesuré segment par segment. Ensuite, une moyenne de ces résultats a été calculée pour chaque note (de 1 à 8), afin d'obtenir un taux d'erreur mot pour chacun des huit degrés de spontanéité.

5.1.2 Résultats généraux

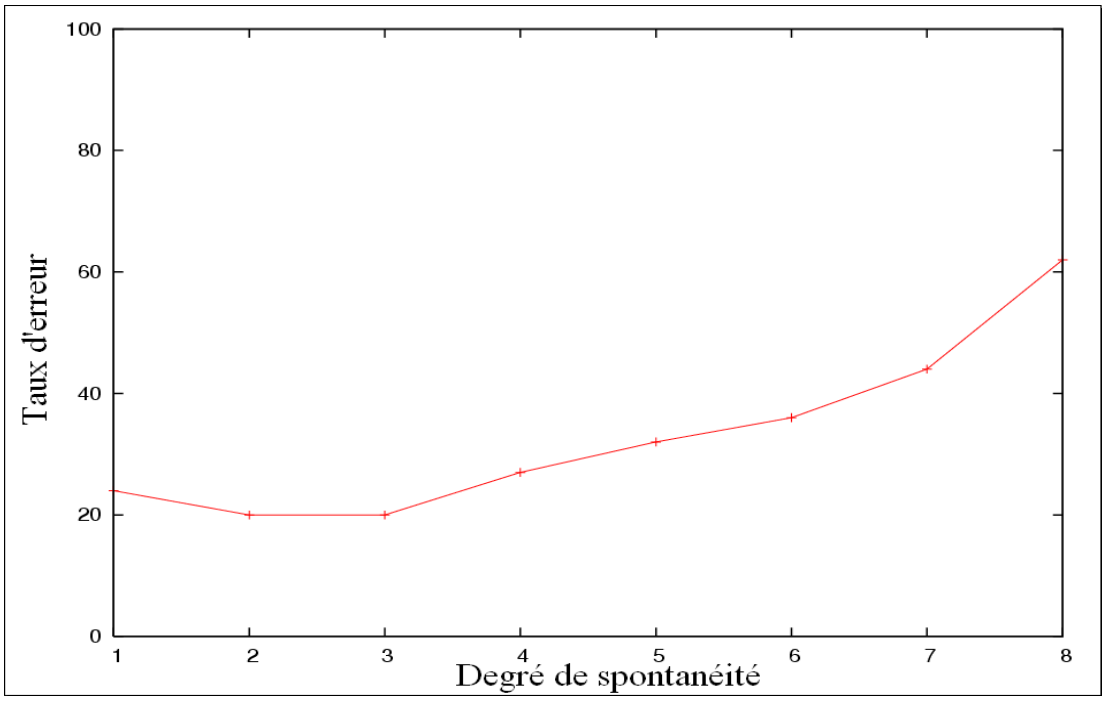


Figure 30 : Taux d'erreur de LIUM RT en fonction du degré de spontanéité

Comme l'indique la courbe ci-dessus, il existe une corrélation entre degré de spontanéité et taux d'erreur mot. Si l'on excepte les segments auxquels ont été adjointes les notes 2 et 3, légèrement mieux reconnus que ceux auxquels 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 deux grandes catégories :

- De 1 à 3, la parole est fluide voire très fluide, avec peu ou pas de disfluences, et le taux d'erreur se situe aux alentours de 20%. Il s'agit très majoritairement de segments issus de journaux d'informations.

- De 4 à 8, la parole devient de moins en moins fluide. Les hésitations, répétitions, déformations... deviennent fréquentes, et le taux d'erreur croît en conséquence. Il va jusqu'à atteindre 60% sur les segments notés 8 : dans ces cas précis, la parole devient difficilement compréhensible tant les disfluences sont nombreuses et accentuées.

5.1.3 Résultats détaillés

Afin de donner un aperçu du travail et des réflexions menés par les deux annotateurs, nous avons choisi de représenter ci-dessous la transcription de trois segments extraits de notre corpus. Le premier a obtenu la note 1, le deuxième la note 5 et le troisième la note 8. Ils sont ainsi représentatifs des principales tendances de notre barème d'annotation puisque, nous le rappelons, la note 9 n'a jamais été attribuée au cours de l'expérience. Nous avons utilisé les caractères gras pour mettre en évidence les disfluences ainsi que les troncations et ruptures de syntaxe, symbolisées par l'usage de parenthèses.

- Exemple 1 : « et on commence avec **euh** la déception des scientifiques européens dans leur expédition martienne la sonde Mars Express n'a pas réussi à prendre contact avec Beagle deux »

- Exemple 2 : « **euh** de majorité qualifiée ce que les peuples ne peuvent pas forcément comprendre et davantage de politique commune politique économique politique structurelle politique de développement je pense que pour un pays comme la Pologne son adhésion doit **être** davantage fondée sur **euh euh la possibilité de de** connaître **euh euh** de l'expansion »

- Exemple 3 : « alors oui on **pour() on on on () si on prend le si on prend le ()** la fable **euh de de ()** la pièce sur **le le** plan du fait divers on peut penser à *Oranges Mécaniques* »

A la simple lecture de ces énoncés, on comprend tout de suite que le premier est clairement un propos journalistique, tandis que les deux autres évoquent plutôt des réponses données lors d'une interview. L'énoncé n°1 n'a posé aucune difficulté aux annotateurs : le propos est limpide, clair (il est très vraisemblablement lu sur un prompteur) et ne comporte qu'une seule marque discrète de disfluence en la présence du filler « euh ». En conséquence, excepté une petite saccade à cet endroit précis, le débit du locuteur est ici très fluide.

On notera également une forte proportion de mots porteurs de sens d'au moins deux syllabes (*déception*, *scientifiques*, *européens*, *expédition*, etc.) : comme nous l'indiquions dans l'introduction du présent chapitre, cet indice est révélateur de la spontanéité ou non d'un propos.

Enfin, la réalisation de certains schwas, et notamment celui de l'infinitif « prendre », corroborent la note qui a été attribuée au segment, et donc son caractère préparé Dans un cadre de parole spontanée, un locuteur francophone (sans accent particulier, notamment

méridional) prononcera neuf fois sur dix la séquence « prendre contact » de la façon suivante : [pRa~dko~takt]. Dans le même ordre d'idée, la liaison entre « scientifiques » et « européens » est elle aussi typique d'un propos préparé. À la différence de celles entre le déterminant et le nom (« mon ami » = [mo~nami], c'est une liaison facultative. En conséquence, elle « tombe » facilement à l'oral, et la prononciation « sja~tifik**z**2Rope9~] est donc ici un indice pertinent du faible degré de spontanéité.

Outre l'élision du schwa, on observe donc également celle du deuxième « r » de « prendre », très fréquente à l'oral avec les syllabes finales[39] en *-bre*, *-dre*, *-tre*... Or, dans notre énoncé n°1, la syllabe *-dre* est très distinctement prononcée dans son ensemble, attestant d'un soin de diction spécifique aux journalistes notamment.

L'exemple n°2 est plus délicat à traiter. D'une part le segment est relativement long, et d'autre part les disfluences y sont principalement concentrées en un seul et même endroit. En effet, si le segment commence par un filler « euh », le débit du propos est ensuite très fluide jusqu'à l'expression « son adhésion ». Puis brusquement, des disfluences apparaissent : la forme *être* est prononcé [Et], donc avec élisions successives du « r » et du « e » muet final, et des répétitions apparaissent soudain. Celles-ci sont d'autant plus intéressantes qu'elles concernent en partie le filler « euh », donc traduisant sans doute plus une véritable hésitation qu'un bégaiement. Enfin, le groupe nominal « la possibilité » est déformé par le locuteur : il subsiste un doute sur la forme exacte du déterminant (« la » ou « les » ?), et le mot « possibilité » est compressé au niveau de ses troisième et quatrième syllabe, donnant une forme que l'on pourrait approximativement retranscrire ainsi : [posibiite]. Nous pouvons en somme faire l'hypothèse qu'il y a une évolution de la « spontanétié » entre le corps de cet énoncé et sa partie terminale.

L'attribution de la note a donc ici été plus complexe : les disfluences sont certes peu représentatives si on les compare au nombre total de mots contenus dans l'énoncé, mais elles n'en demeurent pas moins présentes. Par ailleurs, comme nous venons de le voir, certaines influent véritablement sur la compréhension du propos. L'argument des mots porteurs de sens d'au moins deux syllabes vient quant à lui quelque peu pondérer ces remarques : *majorité*, *qualifiée*, *politique*, *économique*, *structurelle*, *développement*... Le lecteur aura tôt fait de comprendre que le locuteur ici concerné est un politicien et que son discours, en dépit des disfluences mentionnées, demeure sémantiquement riche. A ce titre, le début du segment est

39 Nous reviendrons longuement sur ces phénomènes d'élision en 5.2.1.1

un modèle de cohérence morpho-syntaxique et, à bien y regarder, il comprend autant de disfluences que le segment n°1, c'est-à-dire une seule : le filler « euh ». Ainsi, il aurait sans doute suffi que le locuteur marque une pause significative pour que le logiciel LIUM RT segmente cet énoncé en deux parties, dont les notes auraient assurément été radicalement opposées. Dans le cas présent, l'annotateur a justement choisi, pour toutes les raisons que nous venons d'évoquer, de pondérer son jugement. Il en a résulté un score de 5, équivalent à un degré de spontanéité moyen.

Enfin, l'exemple n°3 fait partie des segments les plus disfluents que les annotateurs aient eu à traiter. La prédominance des caractères gras et des parenthèses fait apparaître un nombre considérable de répétitions, faux départs, *fillers* et troncations. L'idée directrice que veut faire passer le locuteur (« on peut penser à *Oranges Mécaniques* ») est sans cesse retardée, reformulée ou bégayée, et l'énoncé ne devient fluide qu'après de nombreuses saccades dues à ces disfluences. On est ici bien loin du débit presque métronomique de l'exemple n°1.

Par ailleurs, au jeu des mots porteurs de sens d'au moins deux syllabes, cet énoncé est atypique : il n'y en a quasiment aucun, excepté le titre du film de Stanley Kubrick. À l'inverse, le nombre de monosyllabes « parasites » est pour sa part impressionnant : *on*, *si*, *prend*, *le*, *de*, *euh*... Et certains d'entre eux, comme *on*, apparaissent à cinq reprises.

L'une des conséquences presque logique de ce constat est la non-pertinence de l'élision des schwas, dans un contexte pourtant fortement disfluent. Et ce pour une bonne raison : la très faible présence de mots plurisyllabiques provoque une faible proportion de « e » muets. Nous en dénombrons ainsi seulement trois : *fabl**e***, *pièc**e*** et *Orang**es***. Le premier n'est pas élidé, parce que suivi d'un filler « euh » qui pourrait, selon le point de vue dont on se place, n'être qu'un « e » final allongé. Nous expliquerons en 5.3.2. pourquoi nous avons choisi, dans nos transcription, de considérer cet allongement comme un « mot » à part entière.

Le schwa de *pièce* est pour sa part élidé, presque naturellement puisque suivi d'une consonne [s] (« sur ») qui vient se lier à celle de *pièce*. Ainsi, la prononciation de la séquence devient [pjEssyR], mais celle-ci n'a rien de spécifique à l'oral spontanée. Elle est également la plus répandue dans un cadre préparé, tout simplement parce qu'elle est beaucoup moins contraignante d'un point de vue articulatoire[40].

Enfin, si le « e » de *Oranges* est élidé, c'est à nouveau pour une raison n'ayant pas trait à une quelconque spontanéité de la parole. *Oranges Mécaniques* est un titre de film, une

40 À nouveau, nous ne prenons pas en compte ici les accents (notamment méridionaux)

expression figée que tout à chacun prononce [oRa~Zmekanik], que ce soit dans une interview ou un journal d'informations.

Remarquons enfin la faible proportion dans ce segment des clitiques avec schwa (*que, le, te, me, se, ce*), parfois très nombreux et alors systématiquement élidés.

En conséquence, cet exemple n°3 fait bel et bien partie des segments avec le plus haut degré de spontanéité de notre étude, mais la note 8 a surtout été motivée par l'incohérence syntaxique du propos, qui ne se décante véritablement qu'avec les derniers mots du locuteurs. Paradoxalement, des indicateurs généralement fiables comme les schwas élidés ne sont pas pertinents ici, entre autres parce que les mots susceptibles d'en comporter sont noyés parmi les *fillers* et autres « parasites ».

Ainsi cette expérience tend-elle à prouver que *disfluences* (avec l'acception que nous lui avons donné) et *taux d'erreur mot* sont deux concepts étroitement liés. Plus précisément, on pourrait dire que certains éléments morpho-syntaxiques, lexicaux ou prosodiques influent sur les performances d'un SRAP. Ces éléments sont quantifiables et identifiables par deux annotateurs avec des écarts de résultats minimes, pour peu que ces deux personnes se soient mises d'accord en amont sur la notion de « disfluence ».

Peut-on pour autant affirmer que le rapport entre nombre de disfluences et taux d'erreur mot est systématiquement révélateur de la spontanéité ou non d'un énoncé ? Il conviendra d'envisager au moins trois autres questions avant de répondre précisément à celle-ci :

- peut-on envisager la parole spontanée sous d'autres angles que celui des « disfluences » mentionnées dans cette étude ?
- les disfluences sont-elles un critère de spontanéité toujours pertinent ?
- un propos spontané ne doit-il pas aussi être vu *via* son degré d'interaction avec le public auquel il s'adresse ?

Les trois sous-parties qui vont suivre tenteront de répondre successivement à chacune de ces questions. Dans un premier temps, nous allons donc nous intéresser aux différents angles sous lesquels il est possible d'envisager la parole spontanée. Nous en avons déjà évoqué quelques-uns au cours de cette expérience ; ceux-ci seront repris et précisés dans les pages qui suivent, où nous proposerons également d'autres approches pour définir la notion de « parole spontanée ».

5.2 La parole spontanée : approches et analyses croisées

5.2.1 La morpho-syntaxe

Nous avons déjà eu l'occasion d'aborder, dans la première partie de ce chapitre, une particularité morpho-syntaxique propre à la parole *spontanée* : les disfluences [Adda-Decker *et al.*, 2004] [Guénot, 2005]. Nous reviendrons donc dans un premier temps sur celles qui nous semblent les plus significatives dans le cadre de cette étude : l'élision, la troncation, le faux départ et la répétition. Ensuite, nous nous intéresserons aux constructions interrogatives, autre indice qui permet régulièrement de différencier la parole spontanée de la parole préparée.

5.2.1.1 L'élision[41]

En morpho-syntaxe, l'élision est un phénomène de réduction, catégorie sous laquelle on peut également ranger la troncation. Pour sa part, l'élision réduit la longueur d'un message en l'amputant d'un phonème, d'une syllabe ou même d'un mot entier (*je* [ne] *viendrai pas !*). Contrairement à la troncation, dont nous parlerons ensuite, l'élision n'induit pas de notion de correction ni d'erreur du message. Il n'est nullement question d'une réduction *volontaire* d'un mot, dans le but d'en employer un autre à la place (*mais j'exi- je préférerais qu'on fasse comme ça*). Il ne s'agit pas non plus d'une réduction dû à un bafouillage, invitant le locuteur à reprendre un mot ou une expression dans son intégralité ensuite (*ces gens-là fabri- euh fabriquent nos vies*).

En fait, l'élision est une démarche plus ou moins naturelle du locuteur lorsqu'il s'exprime de façon spontanée. Celui-ci s'affranchit de certaines combinaisons consonantiques et/ou vocaliques, pour des raisons qui peuvent aller de la socio-culture à la phonologie, en passant par la prosodie (plus la vitesse d'élocution est élevée, plus il est difficile de prononcer certaines syllabes dans leur intégralité). [Luzzati, 2010] s'appuie sur la conjugaison des auxiliaires *être* et *avoir* pour mettre en avant les différences entre « langue de tous les jours » et « langue du dimanche », reposant sur l'élision ou non des pronoms personnels. Il suffit d'ailleurs d'enregistrer quelques bribes d'une conversation au détour d'une rue ou même d'une

41 L'élision spécifique du schwa (ou « e » muet), simplement envisagée ici, sera largement abordée en 5.2.4.

interview (radiophonique ou télévisée) pour s'apercevoir que cette « langue du dimanche » est finalement très peu utilisée, que ce soit par un anonyme ou un homme politique. On croisera ainsi régulièrement les prononciations suivantes :

- *je veux* : [Zv2]
- *tu as* : [ta]
- *il vient* : [ivje~] / *elle va* : [Eva]
- *vous pouvez pas* : [puvepa] / *vous êtes sûrs* : [zEtsyR]
- *ils savaient pas* : [isavEpa] / *elles voulaient chanter* [EvulESa~te]

D'un point de vue discursif, il n'est pas certain que le message soit perçu différemment selon qu'il est prononcé avec ou sans ce genre d'élisions. Les monosyllabes concernés sont utilisés de façon permanente dans le discours, et leurs formes élidées sont presque une nouvelle nomenclature à elles toutes seules. Ainsi [Luzzati, 2010] a-t-il regroupé ce genre de phénomènes « oralisés » sous l'appellation *grammaire α*, qu'il oppose à la *grammaire β*, c'est-à-dire celle des manuels scolaires[42].

D'un point de vue théorique, tout fonctionne en somme comme si aux personnes 1, 2, 4 et 5, le sujet était marqué par une enclise consonantique gauche, et comme si aux personnes 3 et 6 demeurait surtout une opposition masculin / féminin entre [i] et [E]. Nous avons volontairement fait abstraction ici du pronom personnel « nous » puisque celui-ci, dans la langue orale spontanée, n'est pas élidé : il est plus fréquemment remplacé par la forme « on », sur laquelle nous nous attarderons en 5.2.3.3.

Outre ces cas spécifiques qui concernent des monosyllabes omniprésents dans tout discours, la principale élision que nous ayons observée dans le corpus EPAC 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 prononcera plutôt la séquence [akatpat], beaucoup plus simple à articuler dans un discours à débit relativement rapide. Dans le corpus ESTER [Galliano, 2005], on trouve ainsi les prononciations suivantes :

42 Nous reviendrons dans les pages suivantes sur l'opposition entre ces deux concepts.

- *novembre* : [nova~b]
- *convaincre* : [ko~ve~k]
- *descendre* : [desa~d]
- *peut-être* : [p2tEt]
- *survivre* : [syRviv]
- *mettre* : [mEt]

De même, les sons [e] et [E] disparaissent dans certains cas, que ce soit en voyelle initiale ou en interconsonantique initiale :

- *c'était* : [stE]
- *c'est-à-dire* : [stadiR]
- *déjà* : [dZa]
- *écoutez* : [kute]

Un cas assez unique de ce genre d'élision transforme indirectement un schwa final en voyelle accentuée : les démonstratifs « cet » et « cette » sont ainsi parfois prononcés [st9].

Le son [l], dans deux cas précis, peut également être élidé : « plus » ou « je lui » sont ainsi parfois réduits à [py] ou [ZHi] (notons alors que le schwa de « je » est lui aussi élidé).

Il existe également quelques cas isolés d'élisions : « puis », « parce que », « enfin » et « celui » deviennent très souvent [pi], [pask@], [fe~] et [sHi], avec un sens sans doute différent d'un emploi sans élision. De même, l'adjectif indéfini « tout » tend à s'élider partiellement ou totalement dans certaines expressions figées telles que « tout à l'heure » ou « de toute façon ». Ainsi, il n'est pas rare d'entendre dans une interview ou un débat les séquences [ttal9R] et [tfaso~]. Pour ce dernier cas, en plus de l'élision totale du mot « toute », il faut également relever celle du schwa de la préposition « de », entraînant l'assimilation régressive de la consonne « d » qui se transforme alors en « t ». Nous reviendrons en détails sur ces élisions spécifiques avec assimilation en 5.2.4.

Enfin, le pronom démonstratif neutre « ça » est à lui seul un cas d'élision très particulier[43]. Étymologiquement, il est une réduction de la forme « cela », amputée de ses deux phonèmes centraux [@] et [l] et devenant ainsi une abréviation considérée comme

43 Nous ne nous intéresserons pas ici à la forme « ça » en tant qu'interjection (*ah, ça ! Si je m'y attendais...*).

« familière » par les grammaires. Ce qui signifie que « ça » est reconnu en tant que mot faisant partie d'un lexique, avec sa propre orthographe et ses propres usages qui le différencient de « cela ».

Cette dualité est assez unique dans la langue française, et il conviendrait de se demander si la pratique reflète effectivement la différence de registres de langue entre les deux formes. Par ailleurs, il serait intéressant de se demander si une forme élidée comme [stadiR] n'est pas au moins aussi employée que « ça » dans un contexte d'oral spontané. Et dans ce cas, pourquoi ne pas envisager une entrée telle que *stadir* dans le lexique français ?

5.2.1.2 La troncation

La troncation[44] 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, l'hésitation ou l'auto-correction), ne finit pas. En cas de bégaiement par exemple, le mot tronqué est ensuite complètement prononcé (*vous av- avez raison*). Lorsqu'il y a ce que nous appelons auto-correction, le mot tronqué n'est pas repris, mais remplacé par un autre que le locuteur juge plus adéquat (*c'est une pers- une femme remarquable*). On notera dans ce dernier exemple que le « corrigeant » n'est pas forcément accolé au « corrigé » : il peut y avoir un déterminant, un adverbe comme « enfin » ou même un syntagme entier qui les sépare (*elle a abus- enfin disons qu'elle a bien insisté quoi*).

Cela donne des séquences telles que celles présentées ci-après (la troncation est symbolisée ci-dessous par l'emploi de parenthèses) :

26) « des idées ré() révolutionnaires »
27) « il était aussi passionné d'avia() d'aviation »
28) « elle ne sera pas premier secrétai() euh secrétaire »
29) « et ça rebaisse régulière() régulièrement

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 (30), soit il le reprend partiellement (31 et 32) ou en totalité, créant ainsi une rupture de syntaxe (33).

44 Certains linguistes, et notamment Berthille Pallaud, préfèrent dans le cas présent utiliser le terme « amorce ». Nous renvoyons le lecteur à ses travaux [Pallaud, 2004 ; Pallaud, 2006] pour une analyse détaillée de ce phénomène.

30) « alors auj() le starsystem s'est emparé de la télé »

31) « c'est t() vraiment une lettre très émouvante »

32) « c'est-à-dire que pou() sur un repas que vous vendez sept à huit euros »

33) « oui et c'est un k() il vient du Canada »

Enfin, il nous semble important d'établir une distinction entre la troncation telle que nous venons de la définir et les phénomènes d'apocope. La troncation, comme nous l'indiquions quelques paragraphes plus haut, est connotée de façon négative en ce sens qu'elle sous-entend une « erreur » au sens large. Qu'il s'agisse d'un bégaiement ou d'une auto-correction, le point de vue est le même : le locuteur ne se lance pas dans un discours avec la *volonté* d'y tronquer des mots. Lorsqu'elles apparaissent, les troncations sont donc bien à considérer comme des disfluences. Comme nous l'avons illustré en 5.1, elles participent donc des phénomènes de « grille », et chacune d'entre elles représente un *objet disfluent* qui occupe une *place syntaxique*.

Les apocopes s'opposent aux troncations sur un point majeur : ce sont des phénomènes de réduction *voulus* par le locuteur qui les emploie. Il s'agit essentiellement de diminutifs de prénoms (*Alex*, *Fred*, etc.) ou de mots d'usages réduits pour des raisons parfois obscures, mais toujours avec un souci essentiel : celui de gagner du temps. Ainsi *photo(graphie)*, *déco(ration)*, *après-m(idi)* ou encore *imper(méable)* font désormais partie intégrante de la langue française, et le premier cité pourrait même être employé dans un journal d'informations sans que cela ne choque quiconque – pour ainsi dire, qui même le remarquerait ?

Ces exemples n'entrent donc pas dans le cadre de notre étude, sauf à considérer qu'ils puissent être la conséquence d'un bégaiement par exemple ; auquel cas, la confusion entre phénomène voulu ou non par le locuteur peut être totale.

5.2.1.3 Le faux départ

La rupture de syntaxe que nous évoquions ci-dessus avec l'exemple 33 nous amène à parler du faux départ, phénomène en effet assez proche de certaines troncations. Cela étant, nous prendrons soin de distinguer ces deux éléments.

De notre point de vue, le faux départ désigne une interruption à l'intérieur d'un énoncé, et non à l'intérieur d'un mot. Cet énoncé ne connaît pas de fin puisque le locuteur en

commence un autre immédiatement, d'où la notion explicite de « faute ». La conséquence est toutefois la même que dans l'exemple 33, à savoir l'apparition d'une rupture de syntaxe :

34) « ça a été lu et c'est () on a la photo »
35) « il y a dix mille () mais c'était mal »

Dans ces deux exemples, la notion de rupture est totale puisque la deuxième partie de l'énoncé (située après la parenthèse) fonctionne de façon autonome. Elle ne reprend pas les termes de la première, pas plus qu'elle n'y fait explicitement référence. C'est une proposition indépendante qui se suffit à elle seule.

Il arrive cependant que l'on rencontre des « semi faux-départs », où le second énoncé est en fait un complément (36), une reformulation (37) ou une reprise (38) du premier. Comme les faux-départs, il s'agit donc de phénomènes de télescopage, mais sans rupture syntaxique à proprement parler.

En conséquence, les semi faux-départs présentent quelques différences majeures : les segments à droite des parenthèses ne fonctionnent pas toujours comme des propositions indépendantes (37), et surtout ils ont tous un lien lexical ou syntaxique avec leur contexte gauche :

- dans l'exemple 36, l'infinitif *passer* dépend du verbe *voulais*.
- dans l'exemple 37, l'expression *les autres* est explicitement reprise, mais devient sujet et non plus agent.
- dans l'exemple 38, la proposition *vous semblez être* est intégralement reprise, précédé de l'adverbe *pardon* et de la conjonction *mais*.

36) « je voulais vous dire aussi () passer un gros coup de gueule »
37) « donc c'est les autres () c'est le regard que vous renvoient les autres
38) vous semblez être () pardon mais vous semblez être

Précisons enfin que les faux-départs (et semi faux-départs) sont, dans un cadre d'oral spontané, beaucoup moins fréquents que les autres types de disfluences. Nous aurons ainsi l'occasion de montrer dans la troisième partie de ce chapitre que, quel que soit le « degré de disfluence » d'un locuteur, les faux-départs demeurent toujours assez marginaux par rapport à

l'importance que peuvent revêtir les *fillers* ou les répétitions par exemple.

5.2.1.4 La répétition

Comme nous venons de l'évoquer, la répétition d'un même mot ou d'une même séquence de mots est une autre disfluence fortement représentée dans le discours spontané. Sandrine Henry y a consacré une étude [Henry, 2002] sur laquelle nous nous appuierons pour construire ce paragraphe.

Tout d'abord, il convient d'établir une distinction claire entre ce que [Henry, 2002] appelle les répétitions « de compétence » et les répétitions « de performance ». Une répétition de compétence est due à des règles spécifiques à la langue, qui font se juxtaposer deux représentations graphiques du même mot. Par exemple :

- une juxtaposition de deux pronoms, l'un sujet et l'autre objet : *vous vous êtes vus ?*
- une apposition syntaxique : *nous, nous viendrons demain*
- un phénomène d'emphase : *pouvoir, pouvoir, pouvoir... si seulement je pouvais !*

A l'inverse, une répétition de performance participe du « travail de formulation », et est donc à considérer comme une succession d'objets (au moins deux) occupant une même place syntaxique à l'intérieur d'un énoncé (voir 5.1) :

39) « je ne peux pas les donner **dans dans** une biographie classique »

40) « et par des détournements par d**es des des des des des** accidents presque je suis arrivé à la photo »

C'est donc à ce second type de répétition que nous nous intéresserons ici, sous la simple appellation de *répétition* pour ne pas surcharger inutilement un propos désormais clarifié. Comme l'illustrent les exemples 39 et 40, la répétition peut porter sur un seul mot. Mais il arrive aussi qu'un ensemble syntaxique plus conséquent soit concerné :

41) « et donc à ce moment-là ça devient **une valeur une valeur** marchande »

42) « je crois qu'il y a un piste quand même **pour un pour un** scepticisme positif »

43) « il y a une chose **qui le qui le** fait s'irriter et s'indigner »

44) « **ça serait ça serait** assez intéressant »

45) « mais **ce qui est ce qui est ce qui est** inquiétant c'est qu'est-ce qui va arriver ? »

46) « je crois qu'**on a trop on a trop** intellectualisé les choses »

47) « bon lui **il dit ça parce que il dit ça parce que** il a des problèmes »

Les exemples ci-dessus montrent la diversité des structures pouvant être concernées par la répétition (groupe nominal, groupe verbal, pronom, préposition...). Cela étant, on s'aperçoit qu'une répétition concerne rarement plus de deux mots, du moins dans les données du corpus EPAC sur lequel nous nous appuyons tout au long de cette thèse. Les exemples comme 45 et 46 sont même très marginaux, et il semble encore plus rare de trouver une répétition englobant plus de trois mots telle que dans 47.

Par ailleurs, comme le dit [Henry, 2002], toutes les répétitions ne sont pas *verbatim*. Pour le dire autrement, il arrive que la répétition ne soit pas l'image exacte du terme initial mais l'un de ses dérivés, ou encore qu'une répétition ne soit que partielle. Cela arrive principalement dans les énoncés à répétitions multiples (48 et 49), ou encore lorsqu'une troncation vient s'imbriquer à l'intérieur même du segment concerné (50).

48) « là c'était **le le la le la l'**accusation la plus grave »

49) « enfin **le le fond le fond** de la pièce est extrêmement politique »

50) « **la l() la** lettre de Guy Môquet »

Les deux principaux motifs pouvant être à la source d'une répétition sont l'hésitation et le bégaiement, chacun pouvant grandement être influencé par la corde émotionnelle du locuteur (aptitude à gérer le stress, conditions d'énonciation, influence du ou des interlocuteurs...).

À ce titre, si très peu de répétitions sont observées lors d'un journal d'informations par exemple, c'est entre autres parce que la fonction émotive est tempérée par de nombreux éléments : professionnalisme du locuteur, présence d'un prompteur ou de notes, excellentes conditions d'énonciation, pas ou peu d'interlocuteurs directs, etc.

5.2.1.5 Les constructions interrogatives

Les grammaires de langue française, dans leur grande majorité, indiquent que les

énoncés interrogatifs se construisent selon une syntaxe précise et normée. En général, il s'agit principalement d'une construction :

- avec inversion sujet / verbe : *viens-tu demain ?*
- avec la tournure est-ce que : *est-ce que tu viens demain ?*
- avec un pronom interrogatif, simple ou complexe : *qui est là ? Qu'est-ce que tu veux ?*
- avec un adverbe interrogatif, impliquant le cas échéant une inversion sujet / verbe : *pourquoi raisonner ainsi ? comment vas-tu ?*
- avec un déterminant interrogatif, impliquant le cas échéant une inversion sujet / verbe : *quelle décision prendre ? quel temps fera-t-il demain ?*

Or, si les exemples ci-dessus se rencontrent fréquemment dans un contexte de parole préparée, ils deviennent beaucoup plus rares dès que celle-ci est spontanée. Les inversions y deviennent en effet quasiment inexistantes, et seule l'intonation devient alors un marqueur interrogatif explicite (*tu viens ? on y va ?*). Afin de donner au lecteur une idée de cette ambivalence, nous avons analysé les tournures interrogatives des présentateurs de deux émissions (dix exemples chacun, choisis de façon aléatoire) : *Sous les étoiles exactement* et *Quartiers d'été*. Le ton de la première, qui traite de l'actualité artistique, est résolument décontracté sans pour autant jamais tomber dans le registre familier. *Quartiers d'été* est pour sa part une émission culturelle, avec un registre de langue plus soutenu.

Sous les étoiles exactement :

51) *vous revenez de vacances ?*
52) *un avant-tour ?*
53) *vous êtes donc bassiste, contrebassiste ?*
54) *elle se plie ?*
55) *vous avez vous-même une batterie de voyage ?*
56) *vous êtes membre fondateur de Paris Combo ?*
57) *vous vous doutiez du destin extraordinaire qui vous attendait ?*
58) *vous l'appelez vraiment « Belle », Belle ?*
59) *vous avez appris que des petites filles étaient prénommées Belle ?*

60) *il manque qui ?*

Quartiers d'été :

61) *est-ce qu'il y a quelque chose de plus ?*

62) *une histoire subjective de la cinquième république ?*

63) *pourquoi est-ce que c'est aussi difficile en France d'accéder à ce type d'écriture ?*

64) *la vie politique, comment est-ce qu'elle entre dans nos vies ?*

65) *c'est par exemple la fin de la présidence Pompidou, sur fond de riffs de Jimi Hendrix ?*

66) *Mitterrand / De Lattre ?*

67) *c'est aussi une manière de dire que on peut rompre avec les logiques du monde qui nous entoure tout en étant dedans, tout en réussissant d'une certaine manière ?*

68) *un photographe paysagiste peut parfois aussi se faire paparazzi ?*

69) *vous, quel est votre regard au fond ?*

70) c*omment est-ce que vous réussissez à tenir à ce statut complètement à part ?*

Dans le cas de *Sous les étoiles exactement*, aucun des dix exemples n'est introduit par un marqueur interrogatif particulier. Ainsi, si des points d'interrogation ont été ajoutés lors de la transcription, ce n'est dû qu'à l'intonation montante ou à un contexte explicite. L'exemple 60 fait, il est vrai, quelque peu exception puisqu'on y trouve le pronom interrogatif « qui », mais en position finale (par opposition à la forme classique « qui manque-t-il ? »), c'est-à-dire tonique [Garcia-Fernandez et Lailler, 2008].

Sur le plan syntaxique, nous avons donc toujours affaire à la même structure, c'est-à-dire celle d'un énoncé affirmatif classique (sauf dans l'exemple 60). Les neuf premiers exemples, sans la ponctuation, seraient donc impossibles à ranger dans la catégorie « questions » au sens où les grammaires l'entendent, et ne rentrent dans aucun schéma interrogatif conventionnel.

Il semble donc que l'un des principaux traits morpho-syntaxiques de la parole spontanée telle qu'on la rencontre dans le corpus EPAC soit de ne pas employer les tournures en « est-ce que » ou le principe de l'inversion du sujet.

Cela étant, les exemples issus de *Quartiers d'été* viennent quelque peu nuancer ce constat. En effet, la moitié d'entre eux (61, 63, 64, 69 et 70) sont construits sur les modèles

canoniques de l'interrogation, et notamment sur l'un d'entre eux : la structure avec « est-ce que », parfois précédée d'un adverbe interrogatif (63, 64 et 70). Certes, certaines de ces question présentent un renforcement emphatique (64 et 69) souvent caractéristique de l'oral spontané (*toi, tu y vas demain ?*), mais qui ici ne vient pas perturber une construction interrogative par ailleurs très rigoureuse.

Les cinq autres énoncés (62, 65, 66, 67 et 68) s'apparentent plus, dans leur structures, aux questions issue de l'émission *Sous les étoiles exactement*. Certains ne contiennent aucune forme verbale – conjuguée ou non (62 et 66), creusant encore plus le fossé avec l'autre moitié du corpus *Quartiers d'été*.

Pour autant, il est possible d'utiliser un facteur qui permet de distinguer globalement les deux ensembles de dix occurrences : le nombre moyen de mots par question. L'impression visuelle que l'on peut avoir en regardant simplement les deux listes est ainsi confirmée par les chiffres, puisque *Sous les étoiles exactement* obtient un total moyen de 5,8 mots par question, quand *Quartiers d'été* arrive à 12,6.

Cet écart, supérieur au rapport simple / double, n'est pas le fruit du hasard. Le relier directement aux degrés de spontanéité des deux émissions serait quelque peu hâtif, mais il est incontestable qu'une corrélation existe. Comme nous l'avons dit au début de ce paragraphe, le ton de *Sous les étoiles exactement* n'a pas grand-chose à voir avec celui de *Quartiers d'été*[45]. Les questions des animateurs participent de cette dichotomie, tant sur leur forme que sur les réponses qu'elles sous-entendent. Ainsi, sans rentrer dans une simple opposition « interrogation totale / interrogation partielle » pas toujours pertinente dans ce contexte (*est-ce qu'il y a quelque chose de plus ?* attend une réponse plus fournie qu'un simple *oui*), on parlera plutôt de questions à réponse simple ou complexe.

Envisagées sous cette angle, les questions de *Sous les étoiles exactement* appartiennent toutes à la premières catégories. Les réponses attendues tiennent dans des « oui ou des « non » qui se suffisent à eux-mêmes (51, 54, 55, 56, 57, 58 et 59), des précisions sur un lexique particulier (52 et 53) ou encore sur l'identité de personnes absentes (60).

À l'inverse, les questions posées par Ali Baddou, le journaliste présentateur de *Quartiers d'été*, attendent pour la plupart des réponses complexes, même lorsqu'il s'agit d'interrogations totales (61, 65 ou 68 par exemple). Il ne s'agit plus de questions où un « oui » clôt instantanément le débat (*vous revenez de vacances ? elle se plie ?*), mais de questions où

45 Le chapitre 5.4 explicitera de façon plus précise cette différence de ton.

cet adverbe n'est que le point de départ d'une réflexion dont on attend qu'elle soit solidement développée (*c'est par exemple la fin de la présidence Pompidou, sur fond de riffs de Jimi Hendrix ? est-ce qu'il y a quelque chose de plus ?*).

Quant aux interrogations partielles, outre le fait qu'elles soient bien plus nombreuses dans *Quartiers d'été* (7 contre 2), elles sont aussi et surtout beaucoup plus complexes de par les réponses qu'elles réclament. À nouveau, l'opposition est assez remarquable : il n'est plus question ici de précisions lexicales ou de noms à donner, mais de véritables sujets historiques ou philosophiques auxquels l'interviewé se voit confronté (*pourquoi est-ce que c'est aussi difficile en France d'accéder à ce type d'écriture ? la vie politique, comment est-ce qu'elle entre dans nos vies ?*).

5.2.2 L'énonciation

D'un point de vue énonciatif, l'expression *parole spontanée* peut être résumée par la définition de [Luzzati, 2004], qui évoque des « énoncés conçus et perçus dans le fil de leur énonciation ». Ces énoncés sont produits 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 masquer les imperfections de son discours. De ce point de vue, on comprend qu'on puisse parler également de parole conversationnelle, non préméditée ou co-construite.

L'expression « énoncés conçus et perçus dans le fil de leur énonciation » mérité que l'on reprenne terme à terme son contenu. Elle sous-entend :

- la réalisation d'un **énoncé**, c'est-à-dire d'un produit (oral dans le cadre de notre étude, mais qui peut aussi être écrit) de l'acte d'énonciation. La notion d'*énoncé* est à distinguer de celle de *phrase*. Selon [Riegel *et al.*, 2009], une phrase peut se concevoir « en dehors de toute situation de communication et de tout contexte linguistique ». Et les auteurs de prendre l'exemple de la séquence *comment*, *allez* et *vous* pour étayer leur propos. Chaque fois que cette phrase de trois mots est prononcée (ou écrite), elle l'est à travers différents énoncés, qui varient selon le contexte. Ainsi, en dépit d'une structure lexico-syntaxique identique, la portée

variera selon que la phrase *Comment allez-vous ?* :

- vient compléter *bonjour !*
- est une question sincèrement posée par un médecin à son patient.
- a une visée ironique si le locuteur sait que son interlocuteur ne va effectivement pas bien.

- cet énoncé s'inscrit donc nécessairement dans un cadre d'**énonciation**, c'est-à-dire un « acte de production d'un énoncé par un locuteur dans une situation de communication » [Riegel *et al.*, 2009]. Le locuteur s'adresse donc à un allocutaire (ou interlocuteur) qui peut être réel ou virtuel[46], dans un contexte spatio-temporel particulier. Ce contexte est souvent indispensable pour identifier les acteurs et le cadre d'une situation énonciative : des expressions comme « je », « aujourd'hui » ou « le premier ministre » ne peuvent être comprises que par rapport à une situation référentielle ancrée dans le temps.

- cet énoncé est **conçu** dans le fil de son énonciation, c'est-à-dire que le locuteur n'a pas pré-construit son discours. Il l'élabore au fur et à mesure de son expression, et peut le modifier suivant les réactions de son ou ses interlocuteurs. On touche ici du doigt la théorie des actes de langage [Searle, 1972], qui tend à montrer que « tout locuteur, quand il énonce une phrase dans une situation de communication donnée, accomplit un acte de langage, qui instaure un certain type de relation avec l'allocutaire » [Riegel *et al.*, 2009]. Cette théorie est particulièrement intéressante lorsqu'on l'applique à une situation de parole spontanée, là où l'énoncé est précisément susceptible d'être modifié, raccourci, prolongé, etc.

Cela dépend tout d'abord du locuteur lui-même, et de son aisance face à un interlocuteur, qu'il soit sous la forme de micros, de caméras ou d'un véritable public. Cela dépend aussi des réactions dudit interlocuteur : regards, gestes, soupirs, prises de parole intempestives... Tous ces éléments concernent l'acte de langage perlocutoire, c'est-à-dire l'effet produit par un discours sur l'allocutaire. Certains journalistes se sont fait une spécialité de cette fonction perlocutoire, en sachant toujours « viser juste » de façon à venir perturber la conception du discours de leurs invités.

- car en plus d'être **conçu**, un énoncé de parole spontanée est également **perçu** dans le fil de son énonciation. Cela signifie que l'interlocuteur, virtuel ou non, va le recevoir au fur et

46 Ce qui est assez fréquent dans un corpus radiophonique comme EPAC, où les animateurs sont souvent seuls dans leur studio.

à mesure qu'il lui est délivré. Cette situation est par exemple à opposer au théâtre, où un comédien à qui s'adresse une réplique connaît précisément le début, le contenu et la fin de celle-ci. Il *sait* à quel moment il doit réagir, que ce soit par la parole, le geste ou le regard. Ces situations préparées, sauf exception (improvisations, trous de mémoire...), ne surprennent donc pas ceux qui y prennent part. A l'inverse, dans un cadre spontané, l'interlocuteur ne peut qu'essayer de deviner les tenants et les aboutissants du discours qu'on lui tient, par le biais de quelques indices plus ou moins fiables : intonation, rythme, construction syntaxique... Mais comme le dit [Blanche-Benveniste, 2010], un discours spontané « n'est plus une construction linéaire », et le message peut donc être allongé de façon parfois démesurée (avec des expressions comme *je voudrais enfin rajouter*, *juste une dernière petite chose*, *non mais pour terminer...*)

En conséquence, l'interlocuteur aura toujours tendance à réagir au fur et à mesure de sa perception du discours, et non à la fin effective de celui-ci. Ce qui explique les nombreuses situations confuses qui émaillent régulièrement la parole spontanée (parole superposée, quiproquos, interruptions intempestives...).

De façon plus large, la parole spontanée pourrait, d'un point de vue énonciatif, être synthétisée ainsi. Il s'agit :

- d'un discours **interactif**, matériellement caractérisé par la possibilité de tours de paroles, avec des superpositions, des bribes, des rectifications...

- d'un discours **conversationnel**, explicitement ou implicitement négocié avec des interlocuteurs, qui suppose des ajustements en temps réel, un discours dirigé par un rapport aux autres, dans lequel le sens est une construction commune, même lorsqu'un locuteur tente de transmettre unilatéralement un contenu. Pour autant, nous aurons l'occasion de voir lors d'une expérience menée en 5.4. que la « conversationnalité » ou la « co-construction » ne sont pas toujours mises en exergue dans un cadre que l'on qualifierait pourtant de spontané (interview ou débat par exemple). En effet, lorsqu'un interlocuteur « réel » n'intervient presque pas dans un échange verbal binaire, est-on toujours dans une optique énonciative spontanée ? Quelle profonde différence établir entre ce cas de figure et, par exemple, un journal d'informations co-présenté, où l'interaction entre les deux locuteurs sera finalement assez présente ?

- d'un discours **déictique**, en l'occurrence un discours destiné à être interprété *hic et nunc*, hors de toute distorsion spatiale et surtout temporelle (notamment lors d'échanges téléphoniques). C'est un discours destiné avant tout à provoquer une réaction (éventuellement linguistique) et non une interprétation, une réponse ou une glose.

- d'un discours **erratique,** qui cherche son chemin avec ses mots, dans lequel l'erreur se traduit inéluctablement par un supplément de message, à l'opposé de toute forme de récitation d'un discours préparé par avance, dans lequel ratures et réécritures ont pu jouer leur rôle réparateur.

- d'un discours **spontané**, au sens cognitif du terme, c'est-à-dire d'un discours qui s'improvise en parlant, un discours dont les composants ne sont pas des unités grammaticales conscientes et prédéfinies, et où le concept d'espace de cohérence syntaxique n'a qu'un lointain rapport avec celui de « phrase ».

Ce dernier adjectif (*spontané*) justifie un commentaire particulier, dans la mesure où il convient d'en faire une interprétation non romantique. Ce n'est pas le cœur qui parle, dans une spontanéité du parlé que l'écrit ou l'oral « préparé » auraient du mal à retrouver. Rien en effet n'est jamais spontané dans le langage, dans la mesure notamment où il s'agit toujours d'acquis culturels qui conditionnent une succession non aléatoire des mots. En toutes circonstances, le langage est le fruit d'un calcul, même si les conditions de ce calcul ne sont pas toujours les mêmes.

Dans la langue parlée, la contrainte du point disparaît, à la fois par nécessité et par convention. La langue parlée n'est pas une succession d'espaces syntaxiques clos et autonomes. C'est une succession d'espaces instables, d'errances sur l'axe paradigmatique, entrecoupés de *fillers* et de bribes, qui avance par nature en disfluant. Cela ne signifie pas pour autant que des tendances coalescentes, fondement des approches probabilistes, n'existent pas et qu'on ne peut pas définir des règles de succession. Cela ne signifie pas non plus qu'on ne peut pas bâtir des « modèles de langage » *ad hoc*, même si ce ne seront pas nécessairement les mêmes que pour la parole préparée. L'enjeu des SRAP pourrait alors être d'identifier automatiquement les degrés de spontanéité, afin d'appliquer des systèmes différenciés.

5.2.3 Le lexique

D'un point de vue lexical, la parole spontanée se caractérise principalement par l'emploi de deux morphèmes typiquement oraux, et souvent très utilisés : *euh* et *ben* (ainsi que ses formes dérivées), auxquels nous allons successivement nous intéresser.

5.2.3.1 *euh*

En premier lieu, le choix de la graphie *euh*, que nous avons adoptée tout au long de ces travaux de thèse, doit trouver ici sa justification. En effet, nous avons pris le parti, pour *euh* comme pour *ben*, de suivre l'orthographe des dictionnaires, soit celle qui est la plus usitée. Pour autant, quelques autres possibilités étaient envisageables :

- la variante graphique *heu*, également utilisée à l'écrit, mais tendant à connoter plutôt l'étonnement que l'hésitation.

- l'utilisation de l'Alphabet Phonétique International ou de l'alphabet SAMPA, auquel nous recourons à de nombreuses reprises dans le présent travail. Dans le cas de *euh*, que nous aurions alors représenté « [2] », ce choix a été écarté par souci de lisibilité.

D'un point de vue grammatical la forme *euh* est assimilée, au gré des dictionnaires et des études qui s'y consacrent, à différentes catégories : « interjection », « onomatopée », « mot invariable »... De leur côté, [Morel et Danon-Boileau, 1998] optent pour les terminologies « lexème-fantôme » ou »constituant-fantôme », d'une part pour signifier la brièveté de ce qu'ils appellent les « 'euh' d'hésitation » (entre 40 et 60 centisecondes en moyenne). D'autre part, lorsqu'un *euh* n'est pas suivi d'une pause marquée, il n'a généralement aucune influence sur le syntagme qu'il entrecoupe, et celui-ci se poursuit sans reprise du contexte gauche de *euh* :

71) « Mais je pense, monsieur le Ministre, que vous avez **euh** tout à fait raison »

Cela étant, quelle que soit l'appellation qu'on lui choisira, *euh* est à rapprocher étroitement de la notion d'hésitation, comme le font explicitement [Morel et Danon-Boileau, 1998]. Et, par extension, on le rapprochera donc de la parole spontanée. En effet, au même titre que les répétitions ou les faux départs par exemple, *euh* contribuent à provoquer ce que

[Blanche-Benveniste, 1997] appelle des « turbulences », c'est-à-dire des reprises ou des modifications d'énoncés typiques de la langue orale, où les disfluences abondent :

72) « Mais euh non (en)fin je voulais ce que je voulais pré() dire, c'est que... »

À l'inverse, on trouvera rarement le morphème *euh* dans la parole préparée : un présentateur de journal n'hésite que très peu sur les propos qu'il tient, tout simplement parce qu'il les « lit » sur un prompteur ou sur une fiche, et également parce que cette élocution journalistique très normée fait partie de ses compétences professionnelles. Ainsi, le moindre *euh* prononcé dans un tel contexte fait souvent figure d'événement ; à l'inverse, et sauf s'il est particulièrement marqué, il ne choque pas le moins du monde lors d'une conversation traditionnelle.

Toutefois, il arrive de temps en temps que ce *euh* n'ait pas réellement valeur d'hésitation, bien qu'il apparaisse très fréquemment : nous parlerons alors d'un tic de langage. Nous avons pu observer ce phénomène lors de l'expérience présentée en 5.3, qui porte sur la fréquence des disfluences dans un cadre spontanée. L'un des locuteurs participant à cette expérience, Jean Malaurie, qui s'exprime par ailleurs de façon tout à fait limpide, tend à parsemer ses propos de *euh*. Pourtant, bien souvent, aucune hésitation n'est perceptible au niveau auditif : pas de pause significative, pas d'émotions particulières... Plus que jamais, le qualificatif « fantôme » de Morel et Danon-Boileau semble approprié, puisque ni la syntaxe ni la sémantique ne sont modifiées par ces multiples *euh* dans les deux exemples ci-dessous. Il semble simplement que ce morphème agisse chez le locuteur comme une sorte de respiration, comme un palier indispensable à la bonne tenue de son énoncé :

73) « et à cet égard, je veux rendre **euh** hommage **euh** au film qui vient de passer sur Arte de Stéphane Breton sur les papous, qui est un film extraordinaire **euh** et qui est un film nouveau **euh** qui nous resitue dans les turbulences **euh euh** qui ne cesseront **euh**... »

74) « et dans cette collection **euh** nous avons voulu **euh** réagir contre une tendance je dirais presque 'Louis quatorzième' **euh euh** de s'exprimer **euh** comme si **euh** les faits **euh** sociaux étaient des choses **euh** »

5.2.3.2 ben

De la même façon que nous l'avons fait pour *euh*, nous avons choisi l'usage le plus

répandu pour la représentation graphique de *ben*. Une nouvelle fois, la représentation en alphabet SAMPA [be~] aurait posé, à la longue, des problèmes de lisibilité. Il existe bien une variante graphique, *bin*, assez répandue dans le langage SMS notamment (sans doute parce que le son [e~] est plus intuitivement lu sous la forme « -in », et aussi pour éviter la confusion avec le diminutif *Ben*, issu des prénoms Benoît ou Benjamin), mais cette forme est encore trop peu répandue dans la langue écrite pour supplanter l'orthographe *ben*.

Nous nous appuierons essentiellement, pour la suite de ce paragraphe, sur les travaux de [Luzzati, 1985], consacrés à l'analyse périodique du discours et traitant largement de la forme *ben*. L'auteur y considère *ben* comme un « appui du discours », classe sous laquelle il range également *quand*, *alors* ou *hein*. Ces appuis du discours sont définis comme étant des « mots ayant une nature grammaticale reconnue [...] ou n'en ayant pas [...], et qui se répartissent en deux catégories : les *articulateurs* et les *phatiques* ».

ben a pour sa part longtemps appartenu à la classe des mots n'ayant pas une « nature grammaticale reconnue », mais la tendance a évolué au cours de ces dernières années puisque certains dictionnaires ou grammaires françaises le classent parmi les adverbes, le considérant comme une forme oralisée de *bien*[47]. Gougenheim et ses collègues, dès 1964 [Gougenheim *et al.*, 1964], avaient ouvert la voie en démontrant dans *Le Français fondamental* qu'il était l'un des « mots » les plus utilisés de la langue parlée française

Toujours selon [Luzzati, 1985], *ben* se range du côté des articulateurs. C'est-à-dire qu'il est un « appui du discours ayant, sinon une fonction structurante dans le discours, du moins une place donnée dans la période : en tête des syntagmes appartenant à la *condition* ou à la *résolution* ». Enfin, par « période », l'auteur entend une « unité sémantique et formelle qui comporte trois éléments : la *tension*, la *condition* et la *résolution*. » Nous avons reproduit ci-dessous un schéma proposé par [Luzzati, 1985], qui représente l'ensemble des définitions données et le rôle des appuis du discours (dont *ben*) à l'intérieur d'un énoncé :

47 Gougenheim et ses collègues [Gougenheim *et al.*, 1964] avaient pourtant ouvert la voie il y a cinquante ans en démontrant dans le *Français fondamental* que *ben* était l'un des mots les plus utilisés de la langue française.

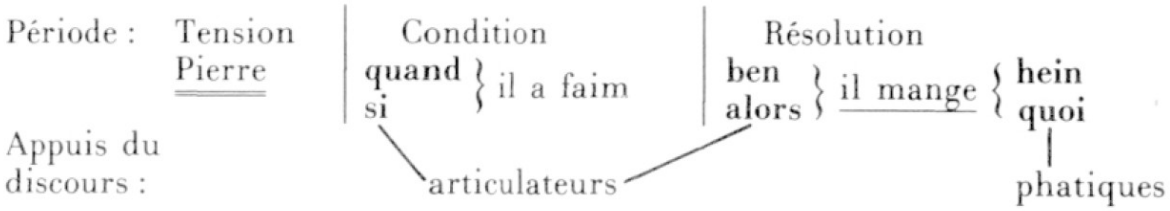


Figure 31 : Appuis du discours et périodes

Comme le laisse supposer la figure 31, il est intéressant d'opposer, au niveau de la construction énonciative et des articulateurs utilisés, parole préparée et parole spontanée. [Luzzati, 1985] prend en exemple le rapport cause / conséquence, généralement exprimé dans cet ordre à l'intérieur d'un propos spontanée :

75) « Pierre il a faim **ben** il mange »

À l'inverse, dans un discours préparé, outre la disparition du morphème *ben*, on constate que le rapport est inversé et que la conséquence précède cette fois la cause :

76) « Pierre il mange **parce qu**'il a faim »

Par rapport à la structure présentée dans la figure 31, on constate cette fois que la condition et la résolution se sont inversées. Daniel Luzzati y voit la volonté, lorsqu'on « surveille son langage, [...] d'énoncer tout de suite l'essentiel, le but du propos, pour évoquer ensuite circonstances et explications ». Alors que dans un discours oral spontané, l'énonciateur « crée un suspense en insérant la condition avant la résolution ».

Comme nous le soulignions précédemment, *ben* est considéré par certaines grammaires et dictionnaires comme une forme oralisée de *bien*. Ce glissement se retrouve notamment dans l'expression *eh bien*[48], que nous avons identifiée à plusieurs reprises dans le corpus EPAC sous la forme *eh ben* :

77) « eh ben ton installation est impeccable »

78) « eh ben moi en l'occurrence c'était un ami qui était écossais »

48 Voir [Sirdar-Iskandar, 1980] pour une analyse détaillée de la locution *eh bien.*

79) « eh ben des femmes me sourient »

Il est par ailleurs remarquable que dans le corpus EPAC, *ben* est souvent remplacé par *bah* dans cette locution particulièrement :

80) « eh bah c'est toutes ces choses-là qui m'ont fait créer ce personnage »
81) « eh bah je connais pas votre mari »

Il est certes difficile d'affirmer que *bah* est plus naturellement substituable à *ben* dans ces cas précis. Par exemple, la forme *Pierre il a faim **bah** il mange* ne choque pas outre mesure, et il ne faut pas non plus oublier que le corpus EPAC, aussi dense soit-il, n'a pas vocation à être représentatif de toutes les facettes de l'oral spontané. Peut-être cette observation est-elle tout simplement le fruit du hasard, mais elle permet au moins de s'interroger sur les liens qui existent entre *ben* et *bah*. Pour ce que nous avons vu, ce dernier semble pour l'instant bien soutenir le test de la substitution, et l'on trouve également dans le corpus EPAC des exemples tels que :

82) « **bah** c'est ce que m'avait dit Georges aussi »
83) « **bah** si maintenant depuis que j'ai lu le livre »
84) « oh **bah** elle lui faisait carrément un massage thaïlandais »
85) « ce qui est **bah** étrange puisque moi j'écris pour les femmes »

Dans ces quatre énoncés, *bah* est employé différemment : en 82, il se situe en tête d'énoncé, dans un rôle d' « appui du discours » [Luzzati, 1982], et on pourrait en conséquence lui substituer *ben*, voire *oui*. Dans les exemples 83 et 84, *bah* se trouve à l'intérieur des locutions *bah oui* et *oh bah*, que l'on rencontre également fréquemment avec la forme *ben*[49]. Enfin, l'exemple 85 est quelque peu à part, dans la mesure où *bah* y joue un rôle plus proche de *euh* que de *ben*. En effet, il indique explicitement une hésitation du locuteur, qui de plus poursuit son propos après s'être « appuyé » sur *bah*, sans rien modifier de la structure morpho-syntaxique initiale. Ce qui nous renvoie justement à la notion de « constituant-fantôme », que nous évoquions dans le paragraphe précédent consacré à *euh*.

49 Un célèbre humoriste français a d'ailleurs intitulé un de ses spectacles « Oh ben oui ! ».

Pour autant, *bah* a également une valeur onomatopéique que l'on ne retrouve pas avec *ben*. Ainsi, [Nodier et Jeandillou, 2008], dans leur *Dictionnaire raisonné des onomatopées*, consacrent une entrée à *bah* en le qualifiant de « mot factice ou artificiel qui échappe aux gens étonnés », qui imiterait « le bruit de la bouche d'un homme ébahi ». Par ailleurs, ils relatent une anecdote évoquant l'hypothèse selon laquelle *bah* serait à l'origine du verbe « ébahir », verbe indiquant lui aussi l'étonnement ou la surprise.

Ainsi, il faudrait sans doute bien plus que les quelques lignes ci-dessus pour parvenir à saisir toute la complexité du mot *ben*, et les quelques analyses et substitutions effectuées ici ne peuvent y suffire. Qui plus est, comme nous venons de le voir, *ben* se glisse régulièrement sous les traits de *bah*, mais parfois aussi *beh* ou *boh.* Cette diversité densifie d'autant plus le niveau d'analyse qu'il faudrait envisager pour mener une étude spécifique.

Cela dit, pour ce qui est de l'emploi de *ben* dans un cadre spontané, nous avons pu observer qu'il n'apparaissait que dans certaines situations précises : à l'inverse de *euh*, dont *a priori* aucun locuteur ne peut se passer lors d'une intervention improvisée, *ben* en est parfois totalement absent. Toujours en nous fondant sur les données de l'expérience en 5.3, nous avons ainsi recensé au moins deux locuteurs (Marie-Josée Mondzain et Odile Decq) qui, au cours d'un entretien d'environ un quart d'heure, ne prononcent à aucun moment le mot *ben* ou l'un de ses dérivés. Et bien que cela paraisse surprenant, Jean Malaurie suit de très près avec une seule occurrence, de même que Jacques-Alain Miller (deux occurrences), qui pour sa part utilise également très peu *euh*. Également, Stéphane Braunschweig, qui a les résultats les plus élevés en termes de disfluences, n'emploie qu'à quatre reprises reprises les mots *ben* ou *bah*. Ceux-ci ne semblent donc pas être particulièrement liés à une mauvaise qualité d'élocution ou a une construction du discours hésitante. Il semble plutôt que c'est du côté du type d'émission, de son ambiance générale qu'il faut chercher : un programme tel que « Sous les étoiles exactement », dont nous parlerons longuement par la suite, instaure une réelle complicité entre le présentateur et les invités, et les questions y sont aussi fréquentes que diversifiées et inattendues. Ainsi, les réponses commençant par *ben oui*, *bah ouais* ou *bah en fait* viennent plus facilement, d'une part parce que le présentateur lui-même n'hésite pas à employer ce genre d'expression, et d'autre part parce que les invités n'emploient sans doute pas le même vocabulaire d'une émission et d'une radio à l'autre. La liberté de ton peut en effet varier considérablement, pour toutes sortes de raison (« étiquette » à respecter, personnalité de

l'animateur, public visé, etc.). La fréquence des emplois de *ben*, *bah* et leurs dérivés semble donc motivée, au-delà d'une simple opposition parole préparée / parole spontanée, par une certaine corrélation entre l'élocution naturelle du locuteur, mais aussi le milieu dans lequel il se trouve et les personnes à qui il s'adresse.

Notons que ces morphèmes et les analyses qui s'y rattachent ne sont pas spécifiques à la langue française. L'anglais, par exemple, possède avec la forme *well* une expression 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.

5.2.4 La phonétique et la phonologie

La phonologie se distingue de la phonétique en ce sens qu'elle s'intéresse aux sons avec une valeur linguistique, c'est-à-dire à des *phonèmes* en relation avec un *signifié*. De ce point de vue, la parole spontanée se caractérise surtout par deux phénomènes importants : la disparition des schwas et les phénomènes d'assimilation qui en découlent.

5.2.4.1 Le schwa

Le mot « schwa », qui est la transcription d'un mot hébreu signifiant « vain », désigne dans le domaine linguistique le son [@][50]. On peut également le qualifier de « e » central, car telle serait sa position au sein du *triangle vocalique* de la langue française (figure 32). Cette appellation désigne une représentation graphique de l'appareil vocal humain basée sur deux éléments : la profondeur du point d'articulation et le degré d'aperture. Selon les langues, le triangle vocalique ne connaît que peu de variations car « on retrouve à peu près universellement les voyelles extrêmes (les plus fermées et les plus ouvertes) » [Riegel *et al.*, 2009].

50 Ou [ə] en Alphabet Phonétique International

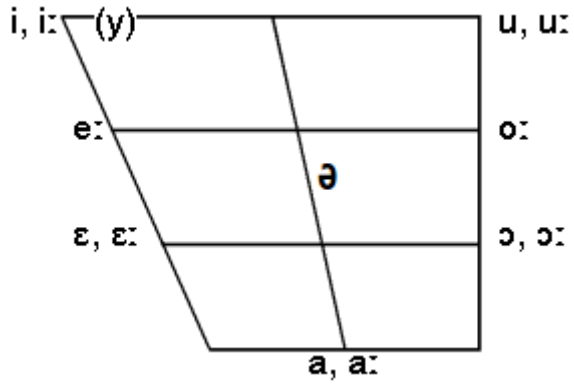


Figure 32 : Triangle vocalique français

Le schwa est également à envisager sous les appellations de « e » muet ou caduc, ou encore « e » instable. La différence entre ces deux premiers adjectifs est particulièrement ambiguë, et le troisième est paradoxalement celui qui résume le mieux le statut du schwa dans l'*usage*.

En effet, on s'accordera assez aisément sur le fait que de manière générale, un « e » situé à la fin d'un mot (« plag**e** », « écol**e** ») est muet (ou caduc), c'est-à-dire qu'il ne se prononce pas. Ainsi, les deux exemples ci-dessus comptent respectivement une et deux syllabes. On pourrait ensuite envisager de tenir le même raisonnement pour un « e » situé au milieu d'un mot («p**e**louse », « franch**e**ment »).

Cela étant, le terme « instable » prend tout son sens dès lors que l'on considère par exemple les différents accents régionaux français. Pour un locuteur avec l'accent du midi, l'adverbe « franchement » comptera traditionnellement trois syllabes, contre deux seulement pour un natif de Paris[51]. Plus encore, il prononcera souvent le mot *pneu* en y ajoutant un schwa pourtant inexistante dans la graphie originelle du mot, donnant la prononciation [p@n2].

L'instabilité évoquée ne s'arrête pas à des frontières géographiques, puisqu'elle défie également celles du temps. Dans la versification du théâtre et de la poésie classique, le schwa jouait ainsi le même rôle phonologique que n'importe quelle autre voyelle. Pour illustrer ces propos, observons les deux alexandrins suivants tirés de « La Légende des Siècles » :

51 Des études plus complètes telles que [Coquillon, 2007] ont été réalisées sur le sujet, et nous renvoyons le lecteur à celles-ci s'il souhaite davantage de précisions.

86) « Il **chanta**, calme et triste. Alors sur le Taygète »

87) « et Joss **chante** fort bien. - Oui, nous avons un maître »

Dans chacun des deux vers ci-dessus (86 et 87) se trouve une flexion du verbe « chanter » : « chanta » et « chante ». Et dans ce second cas, le « e » de « chante » doit se prononcer, au même titre que le « a » de « chanta ». Les règles de la poésie classique en décident ainsi, même si à l'oral la prononciation la plus naturelle serait aujourd'hui [eZOsSa~tfORbje~].

En effet, force est de constater que de nos jours, la réalisation des schwas tient plus du cas de figure spécifique que d'un usage largement répandu : acteurs interprétant une pièce de théâtre classique, déclamation d'un poème, registre de langue particulièrement soutenu, prononciation volontairement accentuée (mentionnons pour exemple le répertoire de Georges Brassens en général, et la chanson « Le Gorille » en particulier)... Même la parole journalistique, rentrant dans un cadre pourtant très artificiel, tend à s'affranchir de façon significative de la réalisation des schwas :

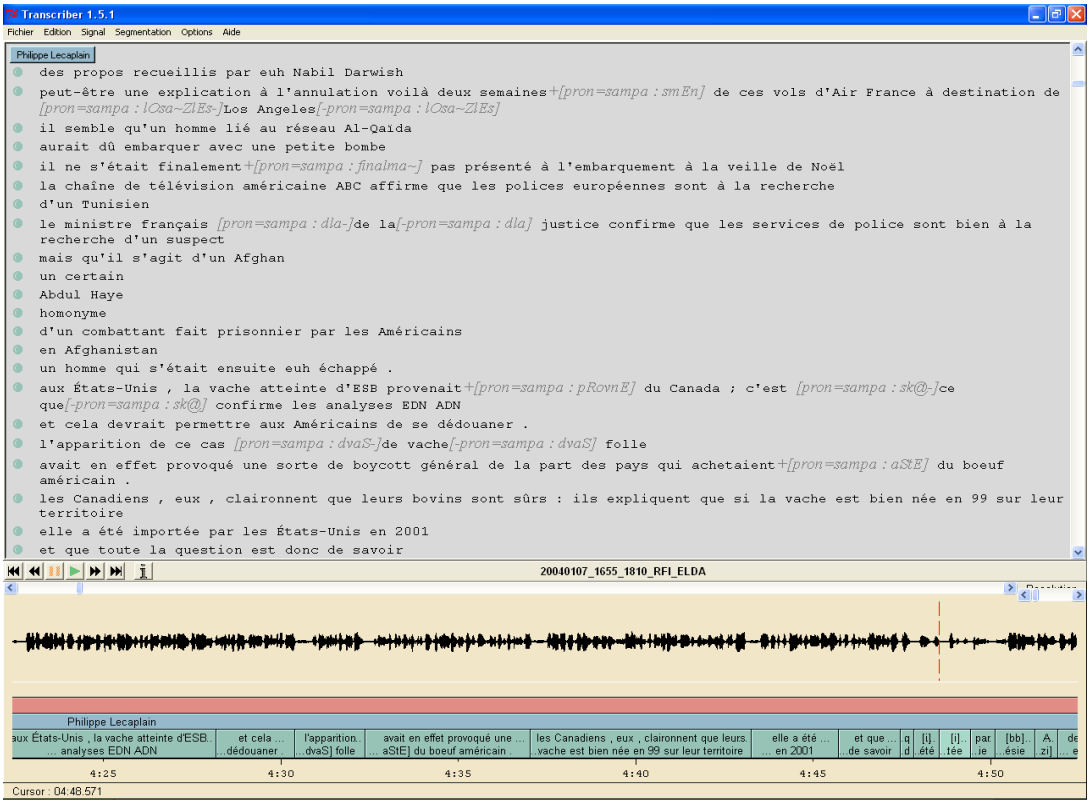


Figure 33 : Exemples de schwas élidés dans un journal d'informations

La figure 33 présente la transcription d'un extrait de journal d'informations radiophonique de la chaîne RFI (précisons que le journaliste est un français né en France, qui parle sans accent particulier). Outre le texte, les principales non-réalisations de schwas ont également été transcrites, à l'aide de balises. Comme on le voit, certaines d'entre elles sont presque surprenantes par rapport au très faible degré de spontanéité de la parole journalistique, et tendent vers l'oral (« ce cas d'vache folle », « le ministre français d'la justice »), voire l'incorrection (« Los Ang'les »).

Pourtant, si l'on se fie à la *Grammaire générale et raisonnée* d'Arnauld et Lancelot, citée par Christophe Rey, « […] en ancien et moyen français, les e dits 'caducs' étaient graphiquement marqués et prononcés comme un véritable *schwa* central et neutre ». Il semble donc acquis qu'il y ait eu une véritable évolution phonologique au fil du temps, tendant à restreindre de façon significative le nombre de schwas effectivement prononcés en français.

5.2.4.2 L'élision du schwa et les assimilations

De nombreux cas de non-réalisations du schwa donnent lieu à ce que l'on appelle une assimilation, c'est-à-dire un rapprochement de deux consonnes, l'une sourde (qui ne fait pas vibrer les cordes vocales) et l'autre sonore. Lorsque l'appareil phonatoire français se trouve confronté à la juxtaposition de ces deux types de phonèmes, il ne peut les retranscrire oralement de façon correcte et naturelle. Il va donc modifier la prononciation de la séquence consonantique, jusqu'à modifier parfois l'image des mots. En effet, dans des situations spontanées, il n'est pas rare que les élisions du schwa et les assimilations se succèdent et se combinent. Ainsi, une séquence comme *Il faut que je te dise* se réalisera soit par [fogZ@ddiz], soit par [fokSt@diz]. C'est-à-dire que sur les six phonèmes de la séquence *que je te,* il n'en restera plus que quatre et, dans le premier cas, deux seulement seront issus de la séquence originelle : [Z] et [@].

Les assimilations sont d'autant plus fréquentes dans la parole spontanée qu'elles sont très souvent provoquées par la disparition d'un schwa, caractéristique récurrente de ce type de discours. Ainsi, dans l'expression *je pense que oui,* le « e » central du pronom « je » aura facilement tendance à s'effacer à l'oral, pour donner la forme artificielle *j'pense que oui*. Or, dans cette structure, le « j » (consonne sonore) est en contact direct avec le « p » (consonne sourde), ce qui va entraîner un glissement de la consonne « j » vers « ch ». On obtiendra donc la prononciation [Spa~s]. Notons que dans ce cas précis, c'est la deuxième consonne qui

« influence » la première ; on parlera donc d'une assimilation régressive. Dans la situation inverse (par exemple, le mot *cheveux* prononcé [Sf2] : c'est ici le son « ch » qui influence le son « v »), l'assimilation est dite progressive.

Par ailleurs, les assimilations peuvent à la fois être progressives et régressives. On parlera alors d'assimilation double. Ces cas se produisent lors de l'influence de deux voyelles nasalisées sur une consonne qui se trouve entre elles. Ainsi, dans l'énoncé *maintenant c'est fini*, si l'on considère que le schwa de *maintenant* n'est pas prononcé (ce qui est, rappelons-le, très fréquent dans la parole spontanée), la consonne « t » va se trouver entre les voyelles « ain » et « ant ». Influencé par ces deux nasales, le « t » aura tendance à se transformer en « n », pour une prononciation que l'on pourrait représenter ainsi : [me~nna~sefini].

Parfois, ce glissement n'a même pas besoin de l'élision d'un schwa pour s'opérer : dans un contexte oral, l'expression *pendant les vacances* sera par exemple régulièrement prononcé [pa~na~levaka~s], toujours suite à l'influence des voyelles nasales.

Dans le corpus EPAC, la combinaison « *je* + consonne sourde » (qui devient donc [S]) est l'une des assimilations les plus observées : ainsi, des formes telles que « j'pense » ou « j'crois » deviennent respectivement [Spa~s] et [SkRwa]. Les mêmes arguments s'appliquent également à « de », et dans des proportions presque identiques. Des expressions comme « pas d'problème » ou « pas d'chance » reviennent également à de nombreuses reprises dans la langue parlée et, toujours par effet d'assimilation, sont prononcées [patproblEm] et [patSa~s].

À un degré moindre, on retrouve également l'élision du schwa avec les pronoms « te », « se » ou «que » suivis d'une consonne sonore :

- *on se donne* : [o~zdOn]
- *tu te demandes* : [tydd@ma~d]
- *que vous* : [gvu]

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.

5.2.5 La prosodie

Ce paragraphe s'appuiera largement sur [Morel et Danon-Boileau, 1998], en explicitant ce que leurs auteurs appellent les quatre « indices suprasegmentaux » de l'intonation : fondamental de la voix (ou F0), intensité, durée et pause. Nous définirons tout d'abord chacun de ces quatre éléments, puis nous verrons en quoi ils caractérisent de façon pertinente la parole spontanée par rapport à la parole préparée.

5.2.5.1 La fréquence fondamentale de la voix (F0)

L'abréviation F0, que nous emploierons par commodité dans la suite de ce paragraphe, désigne la fréquence fondamentale de la voix. Elle peut varier de 80Hz à 850Hz, mais dans une situation de parole « normale » elle se situe entre 110 et 300Hz.

C'est précisément à cette variation de hauteur que nous nous attacherons ici. [Morel et Danon-Boileau, 1998] utilisent le terme *montée mélodique* pour caractériser une F0 ascendante, terme qu'ils explicitent en l'assimilant à « une deixis vocale qui permet d'attirer l'attention sur un fragment du discours ». Ainsi, lorsque la F0 monte, elle a trait à un propos que l'énonciateur juge « déformable, négociable, argumentable dans son échange avec l'autre ».

Inversement, lorsque la F0 chute, elle indique une certaine distance entre le locuteur et son propos. Celui-ci ne cherche pas à susciter de réaction chez son interlocuteur, pas plus qu'il ne veut directement capter son attention.

Comme le résument [Morel et Danon-Boileau, 1998], une F0 ascendante renvoie à la *coénonciation*, avec « une pensée qui s'élabore dans le dialogue et la négociation ». Par opposition, une F0 qui chute est synonyme de *colocution*, avec un locuteur qui « se voit comme un informateur, sans plus ».

Cette opposition renvoie de façon plus ou moins directe à celle que nous étudions depuis le début de cette étude : parole préparée *vs* parole spontanée. En effet, le locuteur « informateur » de Morel et Danon-Boileau évoque les journaux d'informations dont nous avons déjà largement parlé. Le présentateur n'y est pas en situation d' « échange avec l'autre », pas plus que ses propos ne sont faits pour être « déformés, négociés, argumentés », bien au contraire. On observe donc dans ces situations plus de chutes que de montées de F0.

À l'opposé, lors d'un débat, l'essence même de ce mode d'expression fait que chaque

intervention de chaque locuteur est déformable, négociable, argumentable. L'intérêt est précisément la notion d'« échange avec l'autre », même si cet échange se transforme parfois en monologues isolés et disjoints. C'est sans doute pour cela que des remontées de F0 sont souvent observables dans un tel contexte, en fin de segment : c'est, selon [Morel et Danon-Boileau, 1998], « une façon comme une autre de forcer l'attention de l'autre en lui manifestant que l'on n'a pas fini de s'exprimer ».

5.2.5.2 L'intensité

L'intensité, selon [Morel et Danon-Boileau, 1998], caractérise « la façon dont le locuteur gère le canal interactif de l'échange, ainsi que son tour de parole ». Suivant l'intention du locuteur dans une situation d'énonciation (prendre la parole, conserver la parole ou abandonner la parole), l'intensité sera respectivement montante, stable ou descendante. Nous avons régulièrement au l'occasion, lors de la constitution du corpus EPAC, de mesurer l'importance de l'intensité dans les situations de débats politiques par exemple. Lorsqu'il s'agit de conserver la parole ou de la reprendre après avoir été interrompu, les politiciens jouent beaucoup sur l'intensité pour indiquer à leur(s) interlocuteur(s) qu'ils n'ont pas terminé leur intervention. Bien souvent, cela contribue à créer des situations de parole superposée très confuses, où chaque participant (et notamment le médiateur) recourt à l'intensité montante pour justement tenter de « prendre l'ascendant » au sein du débat. Parallèlement, on observe souvent la répétition simple ou multiple d'une même expression, avec justement une intensité de plus en plus montante pour imposer « sa » prise de parole.

Locuteur 1 : mais je vais vous dire c'est son boulot c'est un peu technique
Locuteur 2 : **pardonnez-moi pardonnez-moi pardonnez-moi** de vous interrompre

On observe parfois les mêmes phénomènes, mais en raison d'un problème technique par exemple : un interlocuteur s'exprime au téléphone, mais n'entend pas ce que lui dit le journaliste présent en studio. Ainsi, il poursuit son discours pendant que ce dernier tente de l'interrompre en répétant son prénom avec une intensité de plus en plus montante :

Locuteur 1 : comprends pas, donc je voudrais vous poser cette question
Locuteur 2 : **Frédéric, Frédéric, Frédéric, Frédéric, Frédéric**

À l'inverse, ces pics d'intensité n'apparaissent qu'exceptionnellement dans un cadre de parole préparée. Lors d'un journal d'informations, le présentateur n'a pas à marquer sa volonté de prise de parole, puisque celle-ci lui naturellement donnée par sa fonction, et que son discours ne sera *a priori* pas interrompu de façon impromptue. L'intensité stable domine donc très largement ce genre de situations.

5.2.5.3 La durée

La durée d'une syllabe varie selon « la représentation que le locuteur se fait de l'état de la formulation des idées qu'il s'apprête à exposer sitôt qu'il aura dit ce qu'il est en train de dire » [Morel et Danon-Boileau, 1998]. On comprend dès lors qu'ici encore, la distinction parole préparée / parole spontanée prend tout son sens. Un propos journalistique, écrit et même lu à l'avance, ne pose aucun problème de « formulation des idées » à celui qui en est l'énonciateur. Il peut se contenter de suivre le canevas que lui tisse son prompteur, et n'aura qu'accidentellement le besoin de s'en écarter. En conséquence, chaque syllabe prononcée tendra à conserver la durée des précédentes. Morel et Danon-Boileau appellent cela une « isochronie des syllabes au sein d'un continuum ».

Mais dans une situation d'énonciation spontanée, cette isochronie est beaucoup moins souvent observée. Les hésitations du locuteur allongent par exemple la durée des syllabes, tandis que sa précipitation les raccourcissent.

Dans le corpus EPAC, nous avons noté que les allongements syllabiques se trouvaient surtout en fin de mot. Ils se rapprochent en cela de la définition que fait [Caelen-Haumont, 2002] du mélisme.

Ces allongements, dans leur grande majorité, portent sur le monosyllabe *euh*, que nous avons abordé en 5.2.3.1. Nous avons déjà montré que *euh* avait valeur d'hésitation au niveau sémantique ; celle-ci se voit donc corroborée sur le plan prosodique par un allongement vocalique, qui tend à indiquer l'aspect particulièrement soutenu de l'hésitation.

Ce sont d'ailleurs essentiellement les monosyllabes qui, dans le corpus EPAC, sont sujets au mélisme[52] :

52 Par convention et pour éviter toute confusion, nous avons choisi de représenter l'allongement syllabique par la séquence « :: ».

- « me » : *tout ce parcours du combattant me:: me fascine*
- « un » : *et ça c'est aussi un:: un plaisir*
- « les » : *il a un stylo pour pouvoir signer les:: les polices d'assurance*
- « à » : *je vais pas passer à:: à l'exemple*
- « c'est » : *l'histoire c'est:: on va dire un avion*

On notera d'ailleurs qu'à l'exception du dernier cas, le mot dont la durée se prolonge est systématiquement répété par la suite. L'allongement syllabique semble donc considéré comme une « erreur » par celui qui le réalise, erreur qu'il s'empresse de corriger.

Enfin, nous aurons l'occasion de montrer en 5.2.6. que cet allongement syllabique, et notamment le mélisme en fin de mot, se révèle particulièrement efficace pour détecter automatiquement des zones de parole spontanée dans de gros corpus audio [Jousse *et al.*, 2008].

5.2.5.4 La pause

Les pauses, et plus précisément leur durée et leur fréquence, sont une autre caractéristique permettant de distinguer la parole spontanée. En effet, si l'on observe des données préparées (et notamment journalistiques), on s'aperçoit que les pauses dans le flux de parole y sont généralement peu nombreuses, relativement brèves, et bien souvent contraintes par la respiration et/ou la déglutition (d'après [Morel et Danon-Boileau, 1998], on respire environ 20 fois par minute). Elles n'ont donc pas de « valeur iconique définie ».

À l'inverse, les pauses dans un cadre spontané (interviews par exemple) sont en général beaucoup plus longues et nombreuses : d'une part, les locuteurs ne bénéficient pas d'un canevas (prompteur, notes) pour tisser des propos qu'ils conçoivent donc au fur et à mesure ; d'autre part, les intervenants, lors d'interviews ou de témoignages, ne sont pas toujours familiarisés avec ce type de prise de parole. Ainsi, il en résulte souvent de longs « blancs », marqueurs d'hésitations ou même de ruptures totales de construction.

5.2.6 La reconnaissance automatique de la parole

Enfin, au niveau de la reconnaissance automatique de la parole, la parole spontanée pose de réels problèmes aux SRAP, notamment à cause ses particularités morpho-syntaxiques et phonologiques que nous venons d'évoquer. Concrètement, ces difficultés de traitement se

caractérisent par des taux d'erreurs mots (WER) bien plus importants lorsqu'il s'agit de traiter un débat ou une interview plutôt qu'un journal d'informations. Ces écarts ne sont d'ailleurs pas seulement observables en français : [Kawahara *et al.*, 2003] évoquent les mêmes problèmes dans la langue japonaise, notamment à cause des disfluences.

De façon plus générale, et quelle que soit la langue, l'agrammaticalité (parfois récurrente dans un cadre spontané) pose problème [Dufour, 2010]. Nous l'avons souligné à plusieurs reprises : la parole préparée est normée, formatée et ne contient pas (ou peu) de « fautes » de français. Elle est donc très bien traitée par un SRAP comme LIUM RT, dont l'apprentissage a précisément été réalisé sur de la parole préparée journalistique. LIUM RT obtient ainsi sur ces données des taux d'erreur mot souvent inférieurs à 10%. À l'inverse, comme nous l'annoncions au début de ce paragraphe, la parole spontanée échappe souvent aux structures morpho-syntaxiques classiques. Les SRAP se trouvent ainsi confrontés à des situations qu'ils ne savent pas traiter efficacement, parce que leurs modèles de langage ne sont pas adaptés à ce genre de données.

A ce titre, le dictionnaire de prononciations joue également un rôle essentiel, puisqu'il « permet de faire le lien entre le mot et sa représentation phonémique » [Dufour, 2010]. C'est lui que le SRAP interrogera une fois qu'il aura identifié une séquence acoustique, afin de tenter d'établir une correspondance entre ladite séquence et un mot de son vocabulaire. Comme pour la phase d'apprentissage, le dictionnaire de prononciations peut donc être plus ou moins adapté suivant le type de données auxquelles on le confronte. Pour ce qui est de la parole spontanée, nous avons démontré qu'elle correspondait généralement assez peu à la grammaire « classique » de notre langue, celle que [Luzzati, 2010] nomme grammaire bêta VERIFIER CARACTERE SPECIAUX. Les variantes de prononciations peuvent par exemple être considérables, selon le niveau de spontanéité face auquel on se trouve. Ainsi la séquence *c'est-à-dire que je pense pas que ce soit lui* est-elle susceptible d'être prononcée, au fil des élisions et assimilations :

- [sEtadiRk@Z@p~aspak@s@swalHi]
- [stadiRk@Z@p~aspak@s@swalHi]
- [stadiRgZ@p~aspak@s@swalHi]
- [stadiRgZ@p~aspaks@swalHi]
- [stadiRgZ@p~aspakswalHi]

- [stadiRkSp~aspakswalHi]
- etc.

On le voit, les variantes possibles sont très nombreuses. Il est donc difficile pour un SRAP de les prendre toutes en considération et de parvenir à les associer systématiquement à la prononciation « initiale », notamment parce que la plupart de ces variantes n'ont qu'un seul phonème qui les différencie les unes des autres. La création de dictionnaire de prononciations adaptés, que nous avons détaillée en 2.2.3, apparaît donc comme étant une étape indispensable pour que la parole spontanée soit reconnue efficacement par les SRAP.

Les paramètres acoustiques influencent également de manière considérable les performance des SRAP. Puisque nous venons d'évoquer les variantes de prononciations, il convient désormais d'y associer un facteur acoustique important : le débit. Celui-ci influence considérablement les variantes de prononciation utilisées (ou non) par un locuteur [Fosler-Lussier et Morgan, 1999]. Encore une fois, la parole spontanée est particulièrement concernée puisque, étant « conçue et perçue dans le fil de son énonciation », elle est sujette à de nombreux arrêts, pauses, ralentissements et surtout accélérations selon le cheminement discursif de l'énonciateur. L'émotion du locuteur (angoisse, excitation, énervement...) peut également provoquer un débit excessivement rapide, et donc des prononciations particulièrement altérées [Polzin et Waibel, 1998].

En conséquence, plus le débit d'un locuteur sera rapide ou instable, plus les performances des SRAP diminueront puisque le lexique est en fait directement concerné par cette variable.

Par opposition, la parole préparée est pour sa part beaucoup plus régulière dans son débit puisque le propos n'a pas à être conçu, mais simplement lu ou répété, sans modifications nécessaires. Par ailleurs, ce genre de parole étant souvent manipulé par des professionnels (journalistes, hommes politiques...), la vitesse d'élocution est maîtrisée, de sorte que le discours reste toujours compréhensible pour l'auditorat et également, en quelque sorte, pour les SRAP. Des chercheurs japonais qui se sont focalisés sur l'acoustique ont d'ailleurs démontré que l'espace spectral[53], plus faible avec des données spontanées, était le phénomène acoustique majeur expliquant la baisse de performance des SRAP [Nakamura *et al.*, 2008].

53 Façon dont est perçue l'amplitude et l'évolution en fréquences d'un son.

Concernant la parole spontanée, le corpus EPAC a permis de bénéficier d'une base solide (une centaine d'heures richement annotées) pour tenter d'optimiser les performances des SRAP sur ce genre de données. Le système LIUM RT a ainsi récemment atteint le chiffre de 17,25% de taux d'erreur mots lors de la campagne ESTER 2, notamment grâce aux apports fournis par la transcription et l'annotation des cent heures de parole essentiellement spontanée.

5.3 Les disfluences : un critère réellement discriminant ?

Comme indiqué dans la première partie de ce chapitre, la parole spontanée comprend sur le plan morpho-syntaxique des phénomènes tels que la troncation, la répétition ou encore la rupture de syntaxe. Ceux-ci appartiennent à la catégories des « disfluences », terme que nous avons défini en 5.1. Ces disfluences recouvrent également, d'un point de vue lexical, des *fillers* comme « ben » et surtout « euh », qui est l'un des mots les plus utilisés dans le corpus EPAC.

Il faudrait avoir pour un segment de corpus, pouvant inclure par sp et par prép un lexique qui compte les occurrences, de façon à pouvoir dire la proportion des euh, même si ce n'est pas exactement un mot, et qu'on peut l'assimiler au monosyllabe vocalique, le plus neutre qui soit. Si c'est trop long, un petit comptage à la main est concevable...

Toutes ces disfluences ont été spécifiquement annotées lors de la transcription des données EPAC, en vue d'analyses linguistiques futures. La première que nous avons souhaité réaliser, une fois la transcription achevée, avait précisément pour but de quantifier ces disfluences au sein de différents extraits de parole spontanée, très similaires dans leur structure et leur contenu.

C'est cette expérience que nous allons désormais présenter, afin d'éprouver la pertinence du lien linguistique entre parole spontanée et disfluences.

5.3.1 Protocole

Les données que nous avons utilisées sont les suivantes : 12 entretiens issues de l'émission « Les Matins de France Culture », durant tous à peu près un quart d'heure. Le cadre de ces entretiens est toujours le même : l'animateur, Nicolas Demorand, reçoit une personne (dont la langue maternelle est le français) issue du milieu politique, culturel ou artistique pour parler d'un événement précis : sortie d'un livre, création d'une pièce de théâtre... Le temps de parole alloué à chaque invité est globalement identique, et l'environnement est le même lors de tous les entretiens, à une petite exception près : deux invités s'expriment par voie téléphonique, et ne sont donc pas présents physiquement face à Nicolas Demorand. En dehors de cette particularité, les locuteurs sont donc tous soumis aux mêmes règles.

À partir des annotations réalisées lors de la constitution du corpus EPAC, nous avons

donc pu relever, pour chacun des douze locuteurs, un certain nombre de disfluences :

- les répétitions
- les troncations
- les faux départs
- les ruptures de syntaxe
- les *fillers* « ben » et « euh »
- les mots prononcés de façon incorrecte

Outre le fait que ces données étaient facilement repérables grâce au travail d'annotation effectué en amont, il nous a également semblé que ces six indices étaient représentatifs de l'appellation « disfluence » telle que nous l'avons définie en 5.1.

Ces annotations ont été réalisées grâce aux balises personnalisables dont dispose le logiciel TRANSCRIBER. La figure 34 donnera au lecteur un aperçu du processus d'annotation.

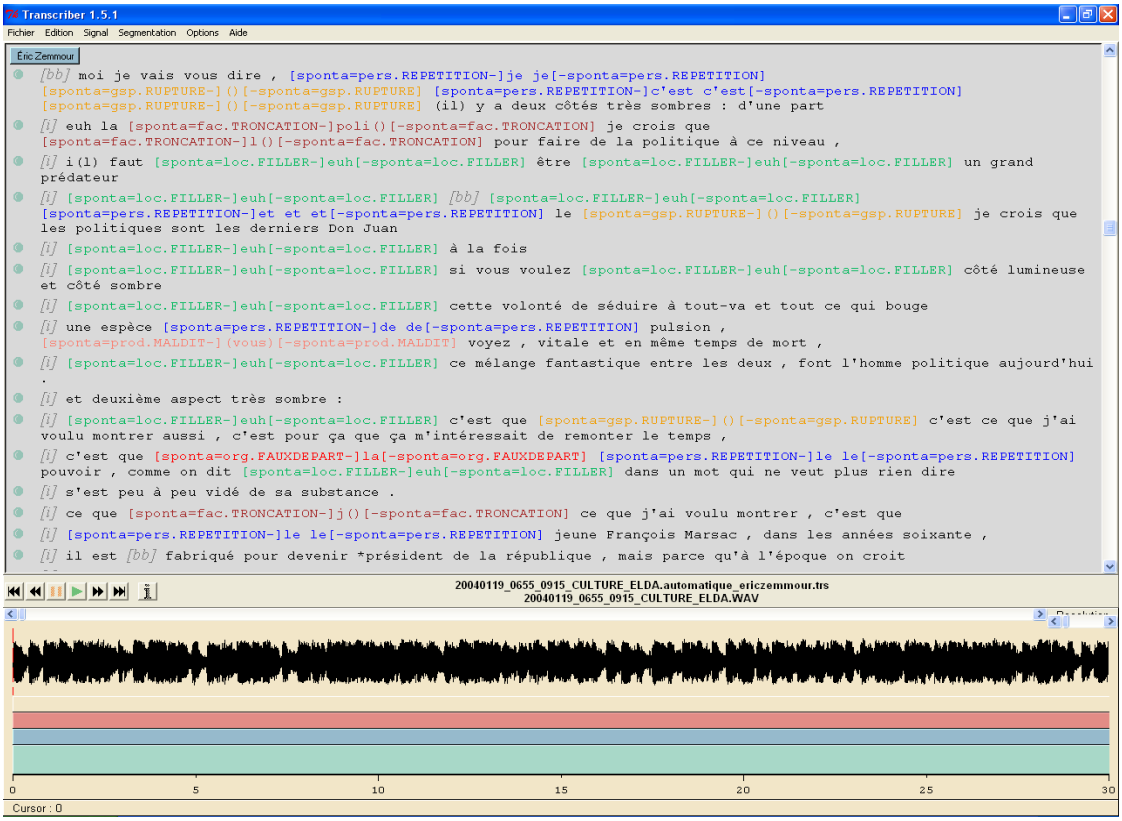


Figure 34 : Exemple d'annotation des disfluences sous TRANSCRIBER

S'il en était encore besoin, cet exemple démontre, sur un plan purement visuel, que les disfluences peuvent être omniprésentes dans un cadre de parole spontanée. Cependant, il faut nuancer cette impression : comme on peut le voir, chaque balise « encercle » le phénomène qu'elle représente, et apparaît donc sous une forme double.

Nous avons choisi d'assigner une couleur à chaque type de disfluence, afin de mettre plus en avant encore l'aspect visuel. Il est ainsi possible, au premier coup d'œil, d'avoir une idée de la fréquence de chaque disfluence.

Le processus d'annotation a été facilité par le travail réalisé en amont lors de la transcription du corpus EPAC. En effet, les troncations, faux départs, ruptures de syntaxe et mots mal prononcés avaient déjà fait l'objet d'annotations spécifiques, ce qui nous a permis de les repérer plus facilement. Nous avons cependant tenu à rajouter des balises d'annotations propres à cette étude, afin que toutes les disfluences apparaissent sous la même forme. Pour leur part, les *fillers* et les répétitions n'ayant pas fait l'objet d'annotations antérieures, ils ont été relevés et annotés spécifiquement pour le présent travail.

Chaque catégorie de disfluence a ensuite fait l'objet d'un double comptage manuel, pour s'assurer que les résultats obtenus étaient corrects. Ceux-ci ont ensuite été divisés par le temps de parole de chaque invité, afin d'obtenir une moyenne de disfluences par minute. La section suivante présente des résultats généraux puis détaillés de cette étude.

5.3.2 Résultats généraux

Locuteurs	Disfluences / minute	Disfluences les plus fréquentes	Plus long segment sans disfluence
Stéphane Braunschweig	28	Fillers (10,7) ; Répétitions (9)	8 s
Odile Decq	18,6	Troncations (6,6) ; Faux départs (3,6)	12 s
Éric Zemmour	17,7	Fillers (7,5) ; Répétitions (4,6)	22 s
Alain Ducasse	16,7	Fillers (7,9) ; Répétitions (4,2)	16 s
Jean Malaurie	16,5	Fillers (11,4) ; Répétitions (2,1)	15 s
André Engel	16,2	Fillers (7,3) ; Répétitions (4,4)	12 s
Bertrand Lavier	14,5	Répétitions (4,6) ; Fillers (4,1)	31 s
Claude Mollard	13,4	Fillers (8,1) ; Répétitions (2)	32 s
René Burri	13,3	Fillers (5,2) ; Répétitions (4,6)	14 s
Bernard Thibault	8,5	Fillers (6,1) ; Répétitions (2,1)	46 s
Marie-Josée Mondzain	7,6	Fillers (2,6) ; Troncations (2)	30 s
Jacques-Alain Miller	4,9	Fillers (3,2) ; Répétitions / Troncations (0,6)	32 s

Tableau 11 : Quelques résultats relatifs aux disfluences

Les chiffres ci-dessus permettent de dégager quelques tendances très intéressantes, et la colonne « disfluences/minute » nous donne un enseignement majeur : si les disfluences sont une caractéristique de la parole spontanée, alors il existe *plusieurs* paroles spontanées, et les disfluences ne valent que pour certaines d'entre elles.

En effet, les écarts entre Stéphane Braunschweig et Jacques-Alain Miller sont tels (six fois plus de disfluences pour le premier cité alors que les conditions de l'entretien sont, rappelons-le, strictement identiques) qu'il est impossible de considérer les disfluences comme un facteur *toujours* pertinent pour caractériser la parole spontanée. Sauf à imaginer, donc, qu'il puisse exister différentes sortes de parole spontanée. Et que parmi celles-ci, certaines soient, sur les plans morpho-syntaxique et lexical, très proches de la parole préparée ; nous développerons ultérieurement cette hypothèse.

Cela étant, « très proches » ne signifient pas « identiques », et la colonne intitulée « disfluences les plus fréquentes » explicite à elle seule cette nuance. En effet, il en ressort

explicitement que les *fillers*, à une très large majorité, demeure le type de disfluence le plus fréquent chez les locuteurs étudiés. Et pour être plus précis, le morphème « euh », déjà défini en 5.2.3.1, est omniprésent dans leurs discours, bien plus encore que « ben ». Ainsi, chez trois locuteurs (Stéphane Braunschweig, Christian Thorel et Jean Malaurie), il apparaît aux alentours de dix fois par minute, soit une occurrence de « euh » toutes les six secondes.

Certes, ces chiffres peuvent être remis en cause, tant la frontière entre ce filler et un allongement vocalique du schwa est mince. Il est ainsi difficile, lorsqu'on entend une séquence comme [Z@:], de trancher entre les séquences écrites « je » ou bien « je euh ». Si [Adda-Decker, 2007] considère que « l'allongement du schwa évoque plutôt des phénomènes de pauses remplies (de type *euh*) », Maria Candea distingue pour sa part les *euh* des « allongements dits “d'hésitation” », deux phénomènes qui « auraient en français oral non lu une distribution complémentaire et seraient pour ainsi dire des variantes combinatoires d'une seule et même marque » [Candea, 2000a].

Pour la présente étude, notre position a essentiellement dépendu des choix ayant été faits au moment de l'annotation du corpus EPAC, puisque c'est sur ces transcriptions que nous nous sommes fondés.

À l'époque où la transcription d'EPAC a été réalisée, plusieurs choix avaient été effectués. Tout d'abord, la représentation graphique du filler *euh*. Choisir de l'orthographier avec la séquence E+U+H était en soi une prise de position, et d'autres formes telles que *heu* ou une représentation phonétique auraient été tout à fait envisageables. Cela étant, dans un souci de lisibilité et d'inter-opérabilité des données, c'est la graphie la plus usitée qui avait été retenue, à savoir *euh*. Ensuite, il restait à déterminer dans quelles conditions la graphie « euh » serait utilisée. Bien que forcément arbitraire, la décision se résumait à trois cas de figures :

- utiliser la graphie « euh » lorsque le son [2] semblait nettement détaché du phonème, morphème ou mot précédent, et n'était donc pas assimilable à un allongement vocalique mais bien, sinon à un « mot » en tant que tel, du moins à un morphème, à référence hésitatoire.

- ne pas annoter ou représenter d'une quelconque façon les allongements vocaliques jugés « peu significatifs ».

- utiliser une balise spécifique pour les allongements vocaliques jugés « significatifs » par l'annotateur.

En conséquence, d'aucuns pourront peut-être s'étonner des chiffres que nous avons relevés dans notre corpus. Bien que, nous l'aons souligné, de tels choix d'annotation soient forcément arbitraires, ils ont été cohérents au sein même du corpus EPAC. Ainsi, les douze interviews de cette étude ont toutes été annotées par la même personne, et selon les mêmes règles.

Dans un cadre de parole préparée (journal d'informations par exemple), ce genre de problème est résolu beaucoup plus rapidement : les fillers comme « euh » n'y apparaissent qu'en de rares occasions puisque le message du locuteur est lu, et a donc *déjà* été construit. Sauf exception, il ne justifie donc plus la moindre hésitation ni le moindre connecteur[54] supplémentaire. A cela s'ajoute que l'on a affaire à des locuteurs professionnels, susceptibles de prévenir l'apparition de morphèmes qui risqueraient d'être perçus comme parasites dans leurs discours. En tout état de cause, on y constate presque systématiquement l'absence de morphèmes comme « euh » ou « ben ».

Dans le cas de notre étude, le tableau l'indique clairement : même dans les scores les plus bas, les disfluences occupent une part non négligeable (2,6 pour Marie-Josée Mondzain et 3,2 pour Jacques-Alain Miller par exemple).

Les mêmes remarques pourraient presque s'appliquer à la répétition, disfluence qui apparaît également à de très nombreuses reprises dans notre tableau. Cependant, il faut signaler que des locuteurs comme Jacques-Alain Miller, Bernard Thibault ou Marie-Josée Mondzain parviennent presque à éviter toute répétition dans leurs discours (avec des scores respectifs de 0,6, 1,2 et 1,8 répétitions par minute).

Ce sont d'ailleurs des discours comme les leurs qui donnent à réfléchir sur la pertinence des critères morpho-syntaxiques et lexicaux pour caractériser la parole spontanée. Nous l'avons dit tout à l'heure, les scores obtenus par ces locuteurs sont très proches de ceux obtenus par des journalistes – autrement dit très proches de zéro. Ceux de Jacques-Alain Miller sont assurément les plus remarquables :

- 0,6 répétitions /minute
- 0,6 troncations / minute
- 0,3 semi faux-départs / minute
- 0 faux-départs / minute

54 Au sens de « conjonction ou adverbe introducteur » [Candea, 2000a].

- 3,2 *fillers* / minute
- 0,2 mots mal prononcés / minute

Certes, les *fillers* à eux seuls suffiraient sans doute à différencier son discours d'un texte lu sur un prompteur par un professionnel, mais sinon ? L'absence quasi totale de faux départs et de ruptures de syntaxe montre bien à quel point le propos de ce locuteur est contrôlé, afin d'apparaître comme structuré et cohérent. Il ne fait aucun doute que ce propos est « conçu et perçu dans le fil de son énonciation », mais il est aussi et surtout « *bien* conçu *donc bien* perçu dans le fil de son énonciation ».

Il est sans doute plus tentant que pertinent de chercher les causes d'une telle aisance élocutoire, tant elles peuvent être multiples en même temps que contradictoires. L'une des premières serait sans doute l'habitude qu'ont les locuteurs face à l'exercice de l'interview. Si, dans l'absolu, il est tout à fait légitime de tenir compte de cet aspect, il nous semble que ça l'est beaucoup moins dans le cas présent. En effet, les 12 locuteurs concernés sont tous rompus à l'« épreuve » de l'interview, et en ont tous donné plusieurs dizaines avant celles que nous avons analysées. Qui plus est, les corps de métiers qu'ils représentent (homme politique, artiste, universitaire) sont amplement rompus à la communication orale et médiatique, que ce soit pour un discours, un débat, un cours magistral ou un rôle au théâtre. Certes, certaines de ces situations sont extrêmement instrumentées ou préparées, mais il n'empêche qu'elles confèrent à leurs auteurs une expérience illocutoire appréciable. Ainsi, ces locuteurs donnent une image d'eux-mêmes tantôt empreinte de naturel et de spontanéité, tantôt surtout maîtrisée.

Par ailleurs, elles rendent également caduc l'aspect sociologique qui pourrait être pris en considération dans une telle étude. En effet, il ne nous semble pas ici pertinent de tenir compte des origines, des racines ou de l'éducation de chacun des invités : chacun fréquente le paysage audio-visuel français depuis de nombreuses années, et a donc assimilé les codes d'expression de celui-ci.

Enfin, les chiffres de la dernière colonne représentent la durée de la plus longue séquence de parole sans disfluence de chaque locuteur. On pourrait penser, presque logiquement, qu'ils devraient être inversement proportionnels au nombre de disfluences par minutes. Or, dans notre étude, ces résultats ne sont pas aussi limpides. À vrai dire, le raisonnement précité ne fonctionne réellement que dans deux cas : les deux premiers, à savoir Stéphane Braunschweig et Odile Decq. Les rapports entre leurs résultats dans les deuxième et

quatrième colonne sont même d'une précision quasi mathématique : 50% de disfluences en plus pour Stéphane Braunschweig, et un segment sans disfluences 50% plus long pour Odile Decq.

Éventuellement, nous pourrions également considérer que les trois dernières entrées du tableau (Bernard Thibault, Marie-Josée Mondzain et Jacques-Alain Miller) le corroborent également, dans la mesure où les scores de ces locuteurs sont, certes avec des écarts importants, les trois plus élevés de notre étude.

Pour le reste, les résultats ne peuvent être analysés de manière conventionnelle. Par exemple, Jacques-Alain Miller et Bertrand Lavier parviennent à s'exprimer sans disfluence pendant une trentaine de secondes, alors que le second « commet » en moyenne trois fois plus de disfluences que le premier au cours de son interview. Cette ambivalence tend à montrer que certains locuteurs sont en mesure, dans certains cas, de tenir un discours plus fluide et plus limpide qu'à l'accoutumée. Est-ce lorsqu'ils développent un thème qu'ils maîtrisent particulièrement bien, ou qu'ils expriment une idée qu'ils tiennent absolument exprimer de façon aussi peu parlée que possible ? Il est difficile, voire impossible, de répondre à cette question, puisque cela impliquerait de connaître les sentiments, émotions ou opinions de chaque locuteur. Cela étant, ces chiffres permettent au moins d'affirmer que les disfluences ne sont pas un phénomène régulier au sein d'un discours. Il est ainsi des situations, sujets ou moments où elles apparaissent très fréquemment, et d'autres où elles s'effacent presque.

En conséquence, il semble tout simplement que certains locuteurs ont une plus grande « aisance élocutoire » que d'autres. Une comparaison intéressante pourrait ici être établie avec le travail d'un écrivain comme Marcel Proust. Lorsqu'on l'interrogeait sur la longueur et la complexité des phrases de ses romans, l'écrivain utilisait l'image suivante : si un quidam marche dans les rues d'une grande métropole, il se retrouvera perdu parmi des milliers d'autres passants, et ne parviendra pas à percevoir le fonctionnement de ce flux humain. Or, il suffit qu'il monte au sommet d'un haut monument pour qu'instantanément, sa perception des déplacements de la foule soit différente. Parce qu'il aura pris de la hauteur, il pourra plus facilement distinguer les mouvements d'individus qu'il verra désormais comme des groupes, et non plus comme des entités uniques au milieu desquelles il serait perdu. Ainsi, Proust procédait exactement de la même façon lorsqu'il écrivait : il arrivait toujours à prendre assez de « hauteur » pour que la cohérence et la construction de ses phrases lui apparaissent de façon limpide, même lorsqu'elles étaient d'une complexité extrême.

Il nous semble qu'un locuteur comme Jacques-Alain Miller est en mesure d'être assimilé à ce genre de raisonnement. Vraisemblablement, au fur et à mesure qu'il parle (et peut-être même avant), il parvient à prendre ce recul qui lui permet d'entrevoir clairement les grandes lignes et articulations de son propos. Il sait où celui-ci va l'emmener, et par quels détours il passera. Et son mérite par rapport à l'image de Proust n'en est que plus grand puisque Jacques-Alain Miller n'est pour sa part pas dans le cadre de l'écrit : ses propos ne sont pas matérialisés devant lui au fur et à mesure qu'il parle, et il doit donc structurer et anticiper son discours « au fil de son énonciation », sans possibilité de le corriger autrement que par une disfluence. Or, notre étude l'a montré : ces corrections sont extrêmement rares.

5.3.3 Résultats détaillés

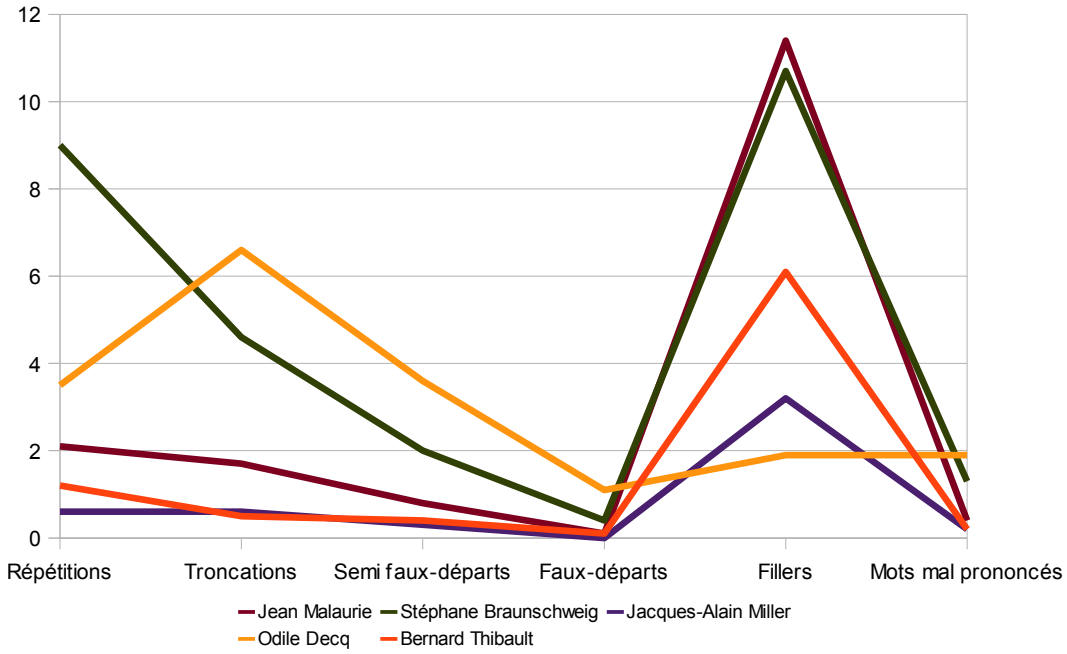


Figure 35 : Représentation graphique des disfluences

Nous avons choisi, dans la figure 35, de représenter les résultats détaillés de certains locuteurs en particulier. En effet, ces derniers sont tout à fait représentatifs des différents « types de parole spontanée » que nous venons d'évoquer, et vont nous permettre de les analyser avec précision.

En premier lieu, ce graphique nous permet de visualiser de façon patente l'importance des *fillers,* que nous mentionnions précédemment. Ceux-ci font apparaître des pics de

disfluences très nets pour l'ensemble des locuteurs, d'autant plus visibles chez des locuteurs comme Jean Malaurie. En effet, bien que cet intervenant utilise très fréquemment le filler « euh », son élocution est par ailleurs claire et bien structurée, sans aucune autre disfluence significative. C'est en cela que sa courbe est remarquable ici : elle ne s'élève jamais au dessus du chiffre 2, sauf dans le cas des *fillers* où elle se situe entre 11 et 12.

Nous pourrions presque ici parler de tic de langage. Il s'agit manifestement d'une démarche inconsciente du locuteur, sans justification discursive précise, et qui ne témoigne que rarement d'une véritable hésitation au sein de son propos :

88) « et à cet égard je veux rendre **euh** hommage **euh** au film qui vient de passer sur Arte de Stéphane Breton sur les papous, qui est un film extraordinaire **euh** et qui est un film nouveau **euh** qui nous resitue dans les turbulences **euh euh** qui ne cesseront **euh**... »

89) « et dans cette collection **euh** nous avons voulu **euh** réagir contre une tendance je dirais presque 'Louis quatorzième' **euh euh** de s'exprimer **euh** comme si **euh** les faits **euh** sociaux étaient des choses **euh** »

À ce titre, il est intéressant d'observer que même Stéphane Braunschweig, qui obtient de loin les résultats globaux les plus élevés, utilise moins de *fillers* que Jean Malaurie. La répartition et la fréquence des disfluences sont donc loin d'être des sciences exactes, et confirment encore un peu plus l'existence de plusieurs types de parole spontanée.

Il est d'ailleurs un autre cas, concernant les *fillers*, qui mérite que l'on s'y intéresse : celui d'Odile Decq. Tout d'abord parce que, à l'inverse de Jean Malaurie, elle les emploie très peu : à peine deux occurrences par minute, soit presque six fois moins que Jean Malaurie. Et, de façon plus large, Odile Decq est le seul locuteur chez qui les *fillers* ne figurent pas parmi les disfluences les plus utilisées (voir tableau 11).

Ensuite, alors qu'une fréquence si peu élevée de *fillers* laisserait à penser que les scores d'Odile Decq vont être globalement très bas, il n'en est rien. Par exemple, ses chiffres concernant les troncations (6,6 / minute) et les faux départs (3,6 / minute) sont de loin les plus élevés de notre étude. En conséquence, alors que le seul nombre de *fillers* suppose un discours plus fluide que celui de ses confrères, les nombreuses reprises et ruptures qui jalonnent le discours d'Odile Decq en font au contraire l'un des plus heurtés.

Sur la figure 35, la courbe de ce locuteur est d'ailleurs visuellement remarquable : son

tracé vient presque s'opposer à ceux des quatre autres locuteurs. En effet, de façon générale, les tracés sont descendants au départ, connaissent un brutal « pic de disfluences » au niveau des *fillers* et enfin chutent fortement. La courbe d'Odile Decq, quant à elle, commence par être franchement ascendante, puis chute radicalement pour ensuite se stabiliser autour de la valeur 2. Elle ne connaît donc pas de « pic » puisque, nous l'avons vu, son utilisation des *fillers* est beaucoup moins importantes que chez les autres locuteurs.

Cette relative similitude que l'on observe entre les courbes de Stéphane Braunschweig, Jacques-Alain Miller, Bernard Thibault et Jean Malaurie permet, d'une certaine façon, d'envisager une hiérarchie des disfluences. Ainsi, les *fillers* en sont clairement la manifestation la plus évidente, celle qu'aucun locuteur ne parvient réellement à éviter, si tant est qu'il s'y emploie. Ensuite, les répétitions, bien que quantitativement très variables, arrivent presque toujours en deuxième position, parfois de façon très significative (Stéphane Braunschweig). La remarque vaut également pour les troncations, troisième disfluence significative, même si là encore les écarts quantitatifs sont importants. Enfin, le tryptique composé des faux-départs, semi faux-départs et mots mal prononcés constitue un ensemble plutôt faiblement représenté chez les quatre locuteurs retenus.

Cette dernière remarque amène inévitablement à se poser plusieurs questions : les faux-départs, semi faux-départs et mots mal prononcés sont-ils les « écueils » les plus simples à éviter dans un discours ? Ou bien sont-ils si rédhibitoires pour la compréhension de ce dernier, qu'un locuteur a naturellement tendance à s'en écarter au maximum ? Et si tel est le cas, cela veut-il dire que le propos d'Odile Decq, par exemple, est nettement moins intelligible que les autres ?

Il faudrait assurément plus de données, d'analyses et de temps pour être en mesure de répondre précisément à ces interrogations. Cependant, nous espérons avoir ouvert la voie à de futurs travaux sur la place et l'influence des disfluences dans un discours spontané, qui ne fonctionnent pas nécessairement de façon binaire comme on serait parfois tenté de le croire.

Nous allons à présent nous intéresser, pour terminer cet ultime chapitre, à un autre élément prépondérant dans un cadre de parole spontanée : l'interaction.

5.4 La parole spontanée et l'interaction

Une des particularités remarquables de la parole spontanée est la notion d'interaction. En effet, s'il semble de prime abord pertinent de considérer comme spontané tout « énoncé conçu et perçu dans le fil de son énonciation », cette définition ne prend nullement en compte le caractère interactif dudit énoncé, à l'intérieur d'un cadre dialogique. Pour le dire autrement, cela signifierait qu'un témoignage long de plusieurs minutes, jamais interrompu par l'intervention d'un autre locuteur que le « témoignant », serait tout autant spontané qu'un débat rugueux où de multiples participants se coupent sans cesse la parole. Or, il nous apparaît que la dénomination « parole spontanée » sous-entend un minimum d'interaction entre au moins deux locuteurs pour être attestée. Ensuite, à l'intérieur de ce cadre, il conviendra de dégager des « degrés d'interaction » suivant différents critères : nombre de locuteurs, nombre de tours de parole, durée des tours de parole, importance de la parole simultanée...

En somme, c'est à ce niveau interactionnel que le terme « conversationnelle » nous paraît pertinent, car il sous-entend de fait un échange, une *conversation* entre différents locuteurs. La parole spontanée n'est donc pas systématiquement, au sens où nous l'entendons, conversationnelle.

Afin d'illustrer ce propos, nous avons établi un corpus de différentes émissions radiophoniques et télévisées (voir en 1.5.1.), sur lesquelles nous avons mesuré les critères précités. Toutes s'apparentaient à une catégorie que l'on pourrait intituler « débat / interview ». Les résultats sont visibles dans le tableau ci-dessous :

Nom de l'émission	Nombre d'émissions	Durée moyenne (minutes)	Nombre moyen de locuteurs	Tours de parole / minute	Durée moyenne d'un tour de parole (secondes)	Nombre de segments superposés / minute
Ripostes	1	51	6	12,5	5	4,5
Sous les étoiles exactement	9	34	3	9,7	8	1,9
C dans l'air	1	51	5	5,5	11	1,3
Culture Vive	13	27	2	4,8	13	0,9
Le téléphone sonne	33	37	11	3,8	17	0,7
Les Matins	15	16	2	2,8	23	0,5
Quartiers d'été	10	37	2	2,4	28	0,3

Tableau 12 : Quelques statistiques sur les émissions transcrites

5.4.1 Analyse générale des résultats

La première observation qui se dégage du tableau 12 est que les deux principaux critères retenus (nombre de tour de parole et nombre de segments simultanés par minute) sont étroitement liés. Cela peut sembler somme toute logique dans la mesure où, si l'on observe un grand nombre de tours de parole au cours d'un programme, c'est souvent parce que ses différents protagonistes s'interrompent et se coupent la parole à de nombreuses reprises. L'émission *Ripostes* (France 5), dont nous n'avons certes transcrit qu'un numéro, illustre parfaitement cette dualité : avec une moyenne de 12,5 tours de parole / minute (ce qui signifie que chacun dure à peu près 5 secondes), elle obtient d'assez loin le score le plus élevé. Parallèlement, son nombre de segments simultanés est très nettement supérieur aux autres. On peut donc dire, en s'appuyant sur ces deux données, que *Ripostes* possède un degré d'interaction très élevé, et surtout très largement supérieur à une émission comme *Quartiers d'été*.

Dans une moindre mesure, il est tout à fait possible de mener le même type de raisonnement sur les autres émissions sélectionnées : les deux chiffres précités sont toujours corrélés, bien que ce soit parfois dans des proportions différentes (pour *Sous les étoiles exactement* par exemple, le nombre de segments simultanés est relativement faible par rapport au nombre de tours de parole).

Cela étant, il devient nécessaire, à la vue de ces résultats, de se poser la question suivante : des programmes tels que *Les Matins* ou *Quartiers d'été* sont-ils réellement plus

interactifs qu'un journal d'informations, par exemple ? Avec moins de trois tours de parole par minute et un nombre de segments superposés proche de zéro, la comparaison mérite au moins d'être évoquée, voire vérifiée. Le corpus EPAC nous ayant permis de transcrire quelques journaux d'informations, nous avons en effet pu nous apercevoir que ces derniers étaient parfois bien loin de l'image rigide et formatée qu'on leur prête bien souvent : co-présentateurs, nombreux chroniqueurs intervenant en direct, invités téléphoniques... Il devenait alors très intéressant de mesurer le réel écart d'interactivité constatée entre un journal d'informations « vivant » et quelques émission perçues comme davantage spontanées, mais finalement moins interactives qu'il n'y paraît, selon les critères objectivables que nous avons retenus.

Programme	**Durée (min.)**	**Nombre de locuteurs**	**Tours de parole / minute**	**Durée moyenne d'un tour de parole (sec.)**	**Nombre de segments superposés / minute**
Journal Inter (20/09/04)	11	7	2,1	28,6	0
Les Matins (22/12/03)	18	2	1,5	40	0,28
Quartiers d'Été (16/08/04)	38	2	1,84	32,6	0,26

Tableau 13 : Comparaison entre des situations dites "spontanées" et "préparées"

Comme l'illustre le tableau 13, les différences sont loin d'être flagrantes entre les deux types de parole représentés ici. Pour donner du relief à l'expérience, nous avons certes choisi un journal relativement interactif et deux entretiens qui, eu égard à leur forme, le sont assez peu. Précisons cependant que ces derniers ont été réalisés dans des conditions d'interaction optimales, puisque l'intervieweur et l'interviewé se trouvaient dans le même studio, ce qui n'est pas toujours le cas. Il n'en reste pas moins que les résultats sont éloquents. S'appuyant sur un nombre de locuteurs beaucoup plus élevé (qui s'explique par la co-présentation et les nombreuses interventions de chroniqueurs, sur le plateau ou en duplex), le journal d'informations de France Inter en devient beaucoup moins stéréotypé qu'à l'accoutumée. Les tours de parole y sont plus nombreux que dans certains entretiens spontanés, ce qui est tout de même très surprenant. Certes, aussi interactif qu'il soit, ce journal ne contient aucun segment de parole superposée, mais les deux entretiens auxquels nous le comparons sont peu prolixes en la matière (moins de 0,3 par minute).

Il est assez surprenant de constater que le nombre de locuteurs n'influe pas

spécifiquement sur le nombre et la durée des tours de parole. *Le Téléphone Sonne* en est un exemple patent. Mêlant appels d'auditeurs et avis d'experts en studio, on dénombre dans cette émission une moyenne d'environ onze intervenants. Pour autant, ne se coupant que très peu la parole, chacun d'entre eux s'exprime longuement (dix-sept secondes en moyenne pour un tour de parole), ce qui confère au programme un caractère finalement assez peu interactif. A l'inverse, *Sous les étoiles exactement*, émission où ne sont reçus en général qu'un ou deux invités, présente des résultats tout à fait opposés. La parole superposée y est relativement fréquente, et les tours de parole plutôt brefs (huit secondes environ).

On opposera avec les mêmes constats des programmes tels que *Culture Vive* d'un côté, et *Les Matins* ou *Quartiers d'Été* de l'autre. Ils fonctionnent tous trois sur le mode « un intervieweur / un interviewé », et pourtant leurs résultats varient du simple au double, voire au triple pour ce qui est de la parole superposée.

Quel que soit le nombre de locuteurs présents, le cadre d'interactivité d'une émission est donc relativement formaté, et ne varie qu'en de rares occasions. Le type de radio, la personnalité de l'animateur et de ses invités, et (du moins peut-on le supposer) quelques directives éditoriales sont certainement des critères beaucoup plus déterminants, mais beaucoup moins quantifiables et objectivables. Il est clair enfin qu'il s'agit d'une parole spontanée et/ou interactive toute relative. Il n'est ici question que de parole radiophonique et en aucun cas de conversation quotidienne.

De même, on notera que la durée d'une émission n'influence pas non plus son degré d'interaction. Si nous avions pu supposer dans un premier temps qu'un entretien ou une débat relativement longs étaient susceptibles de contenir un plus fort niveau d'interaction (pour de simples raisons statistiques), il n'en est finalement rien. Plus l'entretien sera long, plus il portera l'empreinte du programme lui-même et plus son degré d'interaction sera confirmé, au lieu d'évoluer.

En résumé, la notion d'interaction nous amène à établir une distinction fondamentale entre « parole spontanée » et « parole conversationnelle ». La parole spontanée se veut naturelle, sans pour autant supposer un échange verbal entre plusieurs locuteurs. Certaines interviews ou témoignages illustrent parfaitement cette idée, et l'émission *Quartiers d'été* se pose dans notre étude en modèle du genre. À l'inverse, des débats comme ceux proposés par *Ripostes* sont spontanés mais aussi conversationnels, parce qu'ayant un fort degré d'interaction (nombreux tours de parole/minute et parole simultanée très présente).

Ainsi, il est à l'inverse tout à fait envisageable de se trouver confronté à de la parole préparée qui soit également conversationnelle. C'est notamment le cas des pièces de théâtre ou des duos comiques par exemple. Dans ces cadres précis, le degré d'interaction est souvent très important, mais à aucun moment (ou alors lors d'une improvisation exceptionnelle) le propos n'est naturel. Au contraire, il devient de plus en plus scripté au fil des représentation et/ou des répétitions, bien plus encore que dans un journal télévisé où la teneur du discours change quotidiennement.

5.4.2 Radio / télévision : une seule parole spontanée ? L'exemple du débat

Comme mentionné précédemment, nous avons, au cours de notre travail de constitution du corpus EPAC, fait une petite entorse aux données radiophoniques pour nous intéresser à celles issues de la télévision. Notre étude s'est focalisée sur cinq émissions de la chaîne France 5 : *Arrêt sur Images*, *C dans l'air*, *C'est notre Affaire*, *Madame Monsieur* et *Ripostes*. Ou bien celles-ci étaient des débats à part entière, ou bien elles en contenaient au moins un. Deux d'entre elles se sont avérées particulièrement riches d'enseignements : *C dans l'Air* et *Ripostes*, car elles se résumaient toutes deux à un unique débat avec plusieurs intervenants, ce qui rentrait tout à fait dans le cadre de notre thématique sur la parole spontanée. L'interaction y est en effet très présente, et la nature même des sujets abordés (volontairement polémiques) provoque inévitablement des prises de parole brutales, des interruptions, des bafouillages, etc. Les trois autres programmes étaient soit trop régulièrement entrecoupés de reportages, soit composés d'une suite de plusieurs débats avec plusieurs thématiques et plusieurs invités. Par rapport au corpus radiophonique dont nous disposions, une telle disparité n'avait pas sa place et n'autorisait pas de comparaison quantifiable. Pour autant, les deux émissions retenues permettent d'établir quelques parallèles intéressants entre parole spontanée télévisuelle et radiophonique.

Les premiers éléments à relever sont ceux concernant les deux émissions de télévision elles-mêmes. Bien qu'ayant la même forme (un débat), la même durée et peu ou prou le même nombre d'intervenants, les données qui en sont issues sont diamétralement opposés : 13 tours de parole par minute en moyenne pour *Ripostes*, contre 5 seulement pour *C dans l'air*. 4,5 segments superposés contre 1,3. On notera cependant que les thématiques sont elles aussi opposées, puisque là où *Ripostes* s'aventure sur le chemin sinueux de la prolifération des armes nucléaires, *C dans l'air* traite d'un sujet sans doute moins périlleux : les perturbateurs

de l'élection présidentielle en 2004. Les prises de position et velléités de chacun des invités sont plus vives et plus houleuses dans le premier cas, ce qui peut être une première explication concernant le nombre de tours de parole et la fréquence des segments superposés.

Ensuite, et c'est là sans doute une des principales explications, les personnalités des deux animateurs sont totalement divergentes. Serge Moati est démonstratif, contestataire et parfois même théâtral, tandis que Thierry Guerrier affiche une personnalité beaucoup plus sobre et discrète. Cela s'exprime à travers leurs interventions respectives lorsqu'ils tentent d'interrompre un invité : Thierry Guerrier se montre alors beaucoup moins téméraire que Serge Moati, se rétractant volontiers après une première tentative avortée. A l'inverse, ce dernier n'hésite pas à couper franchement un de ses interlocuteurs pour être entendu.

Balladur il était franchement pas bon
[i] Jospin il était franchement pas bon

Thierry Guerrier

mais

Roland Cayrol

ils l'ont payé
[i] donc i(l) se passe des choses et tout le monde sait ça : ne faisons pas comme si
[i] soixante pour cent des Français étaient d'ores et déjà jusqu'au jusqu'au bout
[i] décidés à voter pour ces deux-là

Thierry Guerrier

mais aujourd'hui quand vous posez puisque vous évoquez le
[i] la réaction du public euh qui trouve que on en fait trop

Figure 36 : Extrait de la transcription d'un débat animé par Thierry Guerrier

Pascal Boniface

et donc ils ont
[i] à la fois l'Iran fait peur au monde et eux ils ont peur du reste du monde donc c'est une situation
[i] un petit peu compliquée*+[conv]*

Serge Moati

attendez attendez

Serge Moati + Any Bourrier

1: je voudrais vraiment je voudrais vraiment
2: je voulais revenir un tout petit peu sur le

Serge Moati

que v() attendez
non on va y venir *[bb]*

Figure 37 : Extrait de la transcription d'un débat animé par Serge Moati

D'une manière générale, Serge Moati prend une part beaucoup plus active au débat qu'il anime, se muant presque en intervenant, au même titre que ses invités. Cette donnée est essentielle eu égard à l'interactivité d'une émission, et explique en partie les écarts que nous avons obtenus. Lorsque l'animateur se « contente » de donner la parole ou d'acquiescer, les

échanges sont par nature beaucoup plus policés, et deviennent parfois même mécaniques, jusqu'à en être capable d'anticiper les interventions du journaliste-présentateur. Ce processus est tout à fait impossible à mettre en place avec des animateurs comme Serge Moati, puisqu'il va parfois jusqu'à se mettre, non pas à la place de l'invité mais à celle du téléspectateur, lorsqu'il fait l'hypothèse qu'une notion ou le sens d'un mot aurait pu lui échapper :

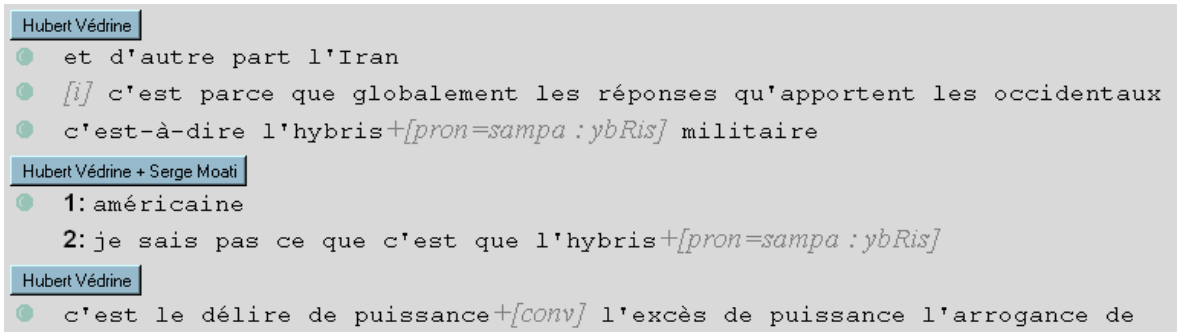


Figure 38 : Exemple d'intervention de Serge Moati

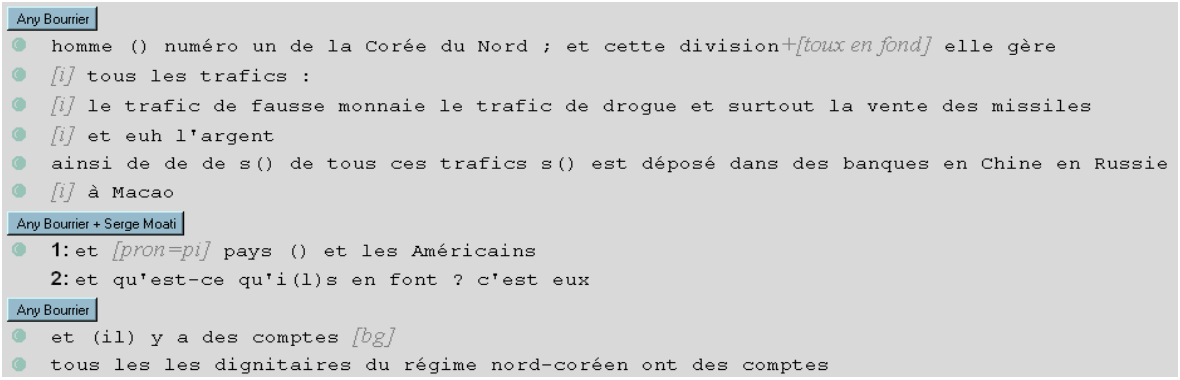


Figure 39 : Exemple d'intervention de Serge Moati (2)

Serge Moati ne se soucie pas du tout du moment où il pose sa question, quitte à interrompre à nouveau son invité. Ce dernier, d'ailleurs, ne relève pas forcément la remarque, comme c'est le cas sur la figure 39.

Dans le cas de *Ripostes,* les chiffres élevés que nous avons relevés (12,5 tours de parole/minute et 4,5 segments superposés/minute) s'expliquent en grande partie par la personnalité de son animateur qui, en l'occurrence, mérite bien cette dénomination.

En ce qui concerne l'émission *C dans l'air*, Thierry Guerrier évolue, comme nous l'avons

précisé, dans un tout autre registre. Il se montre beaucoup plus effacé par rapport à ses invités, qu'il laisse volontiers partir dans de longues interventions sans jamais les interrompre. Tel ses *alter ego* radiophoniques, il agit en maître de cérémonie et prend grand soin de distribuer les tours de parole de façon régulière, en attendant qu'un de ses interlocuteurs ait fini de s'exprimer pour s'adresser alors au suivant. Les chiffres concernant les tours de parole et les segments superposés placent d'ailleurs *C dans l'air* à peine au-dessus de la plupart des émissions radiophoniques de notre corpus. Sans doute eût-il été intéressant de pouvoir inverser les rôles et de voir comment se serait comporté Thierry Guerrier face aux invités plutôt incisifs de Serge Moati, sur un sujet aussi grave que celui de l'armement nucléaire. Il est en effet à peu près certains que les invités, au même titre que la thématique abordée, jouent un rôle dans la notion d'interaction abordée ici. Eux-mêmes d'ailleurs, peut-être, se comportent volontairement différemment suivant l'émission à laquelle ils participent. Toutefois, il s'agit de données difficiles à prendre en compte ici, dans la mesure où nous ne disposons que d'une émission de *C dans l'air* et d' une émission de *Ripostes*.

En dehors d'aspects tels que la personnalité ou les traits de caractère, il existe une différence interactionnelle essentielle entre les animateurs télévisuels et radiophoniques : la nécessité de mentionner le nom du locuteur en train de s'exprimer. A la radio, cette précision est indispensable car, ne bénéficiant pas du support de l'image, avec possibilité de sous-titrage, il est parfois bien difficile pour l'auditeur de savoir exactement qui il est en train d'écouter, si l'animateur ne le précise pas. Naturellement, dans de la parole préparée - un journal d'informations par exemple -, ces précisions sont peu fréquentes car les interventions des différents protagonistes (chroniqueurs, envoyés spéciaux...) sont planifiées et leur chronologie ne connaît que rarement un bouleversement. Inversement, lors d'une situation spontanée, et *a fortiori* au cours d'un débat très animé, il est très souvent impossible de déterminer à l'avance quel locuteur va parler, et surtout à quel moment. Le rôle de l'animateur se modifie alors, puisqu'il doit, à chaque prise de parole qu'il ne sollicite pas lui-même, préciser pour son auditorat l'identité de l'intervenant. De telles interventions peuvent ainsi donner une couleur et un rythme très particuliers aux émissions radiophoniques, puisque la spontanéité des interventions y est contrebalancée par une sorte d'impératif journalistique presque automatisé. On en viendrait presque, en tant qu'auditeur, à être capable de déterminer le moment précis de ces petites apostrophes. Au final, cela donne lieu à des segments de

parole superposée, où l'intervenant peine parfois à poursuivre son discours, perturbé par la brève intervention de l'animateur.

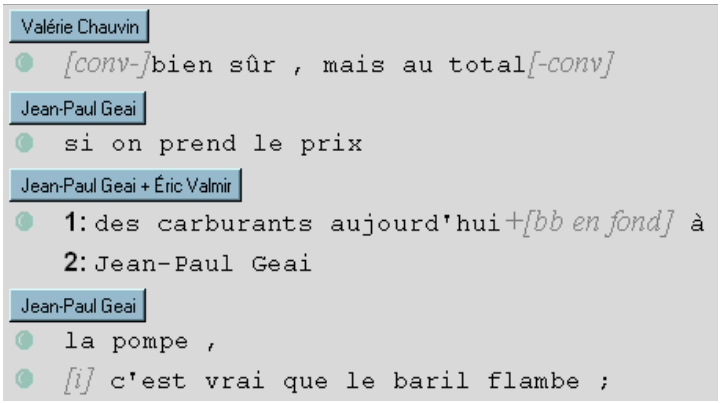


Figure 40 : Exemple d'apostrophe de l'animateur

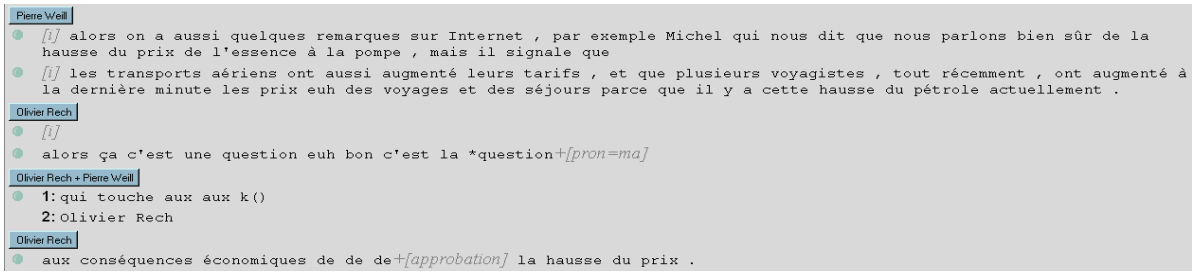


Figure 41 : Exemple d'apostrophe de l'animateur (2)

La différence entre ces exemples et une émission télévisée comme *Ripostes* est patente : au cours de cette dernière, il arrive fréquemment que trois ou quatre intervenants (dont l'animateur lui-même parfois) s'interrompent mutuellement, sans que Serge Moati ne précise qui prend la parole à chaque fois. Il n'en a en effet pas besoin : à la télévision, les caméras assurent parfaitement ce rôle en alternant gros plans, plans larges, etc. pour indiquer aux téléspectateurs la provenance du flux de parole. Le cas échéant, des sous-titrages peuvent également assurer cette fonction.

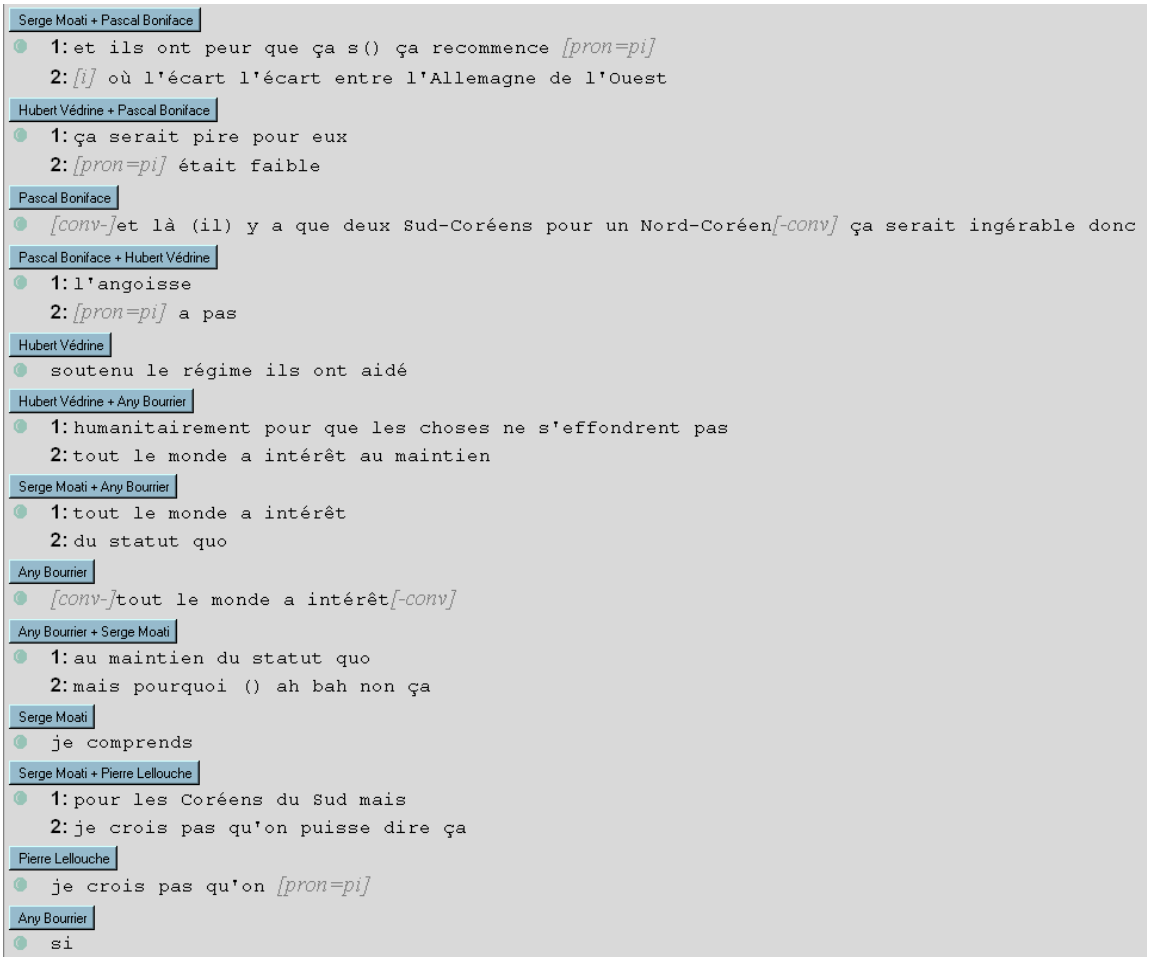


Figure 42 : Transcription d'une situation d'interaction fortement marquée

Une situation aussi interactive que celle présentée ci-dessus serait difficilement envisageable à la radio, sous peine de devenir rapidement incompréhensible. Car si l'appui de l'image n'a pas forcément vocation à expliciter plus clairement les propos de chaque intervenant, il permet au moins de « visualiser » le discours global, ainsi que ceux qui le produisent.

Il est d'ailleurs intéressant d'observer les conséquences d'une telle situation sur le travail de transcription lui-même : pour les données radiophoniques, bien aidé par l'animateur comme nous l'avons vu, l'annotateur n'éprouve que rarement des difficultés à identifier la source de chaque tour de parole. Dans le cas de la télévision, si l'on ne se fie qu'à la partie sonore, cette identification pose de gros problèmes. Ne bénéficiant plus de la mention régulière du nom des intervenants, l'annotateur perd énormément de temps à tenter de discerner les différents locuteurs (notamment dans un passage comme celui présenté ci-

dessus), et doit au final s'appuyer nécessairement sur l'image pour y parvenir.

Qui plus est, en dépit de la présence conjointe de l'image et du son, les deux débats télévisés que nous évoquons ici ont donné lieu à de nombreux passages non transcrits. En effet, lors de ces séquences, de nombreux intervenants parlaient en même temps, ce qui les rendaient inaudibles. A nouveau la différence avec la radio est réelle, puisque lors des débats du *Téléphone Sonne*, tout excès oratoire est rapidement réprimandé par l'animateur.

Puisque nous évoquons la présentation du *Téléphone Sonne*, nous allons nous attarder sur une particularité qui lui est propre : cette émission n'est pas toujours animée par la même personne. Au sein de notre corpus, nous avions ainsi soit affaire à Alain Bédouet (le présentateur principal à l'époque), soit à Eric Valmir (son remplaçant estival), soit à Pierre Weill (qui assura quelques prestations ponctuelles). Il nous a du coup semblé pertinent de dissocier les résultats du tableau 12, de façon à les faire apparaître spécifiquement pour chacun des trois présentateurs cités. Voici les résultats que nous avons obtenus :

Animateur	**Nombre d'émissions**	**Nombre moyen de locuteurs (invités + auditeurs)**	**Tours de parole / minute**	**Durée moyenne d'un tour de parole (sec.)**	**Nombre de segments superposés / minute**
Alain Bédouet	18	12	3,39	18,88	0,64
Eric Valmir	13	11	4,39	14,11	0,91
Pierre Weill	2	9,5	3,33	18	0,36

Tableau 14 : Statistiques comparées d'Alain Bédouet, Eric Valmir et Pierre Weill

S'il est sans doute délicat de vouloir tirer beaucoup d'enseignements des résultats obtenus par Pierre Weill par rapport à ceux de ses confrères, eu égard au faible nombre d'émissions qu'il a animées, nous pouvons en revanche légitimement opposer les chiffres d'Alain Bédouet et Eric Valmir.

En effet, tous deux ont animé un nombre conséquent d'émissions du *Téléphone Sonne*, et qui plus est avec (presque) le même nombre de locuteurs respectifs. Cela étant, on note tout de même une vraie différence, en fonction de nos critères d'interaction tout au moins. Tout d'abord, concernant les tours de parole, on constate que les émissions animées par Eric Valmir en compte un de plus par minute. Cette différence ne paraît forcément considérable, mais elle

prend une autre ampleur si on la ramène à la durée moyenne des tours de parole. Chacun d'entre eux fait ainsi presque cinq secondes de moins que dans une émission dirigée par Alain Bédouet. Étendu à la durée totale du programme (environ 37 minutes), ce chiffre n'est pas anodin.

Il l'est d'ailleurs d'autant moins qu'on peut le corréler avec le nombre de segments superposés par minute. Si, d'une manière générale, *Le Téléphone Sonne* laisse peu de place à ce genre de disfluences, il faut pourtant observer que sous la direction d'Eric Valmir, elles sont environ 1,5 fois plus fréquentes. Il est délicat d'avancer une explication précise pour justifier ces écarts, tant les invités reçus et les thématiques abordées sont variés au cours de la trentaine d'émissions concernées ici. Pourtant, il nous a semblé que les invités présents sur le plateau du *Téléphone Sonne* étaient plus à l'écoute de son présentateur emblématique que de son remplaçant. Ou alors, si l'on prend le problème à l'envers, Éric Valmir peinait parfois à s'imposer comme le « patron » du débat qu'il animait (figures 43 et 44), à la différence d'Alain Bédouet qui s'accommodait très bien de ce rôle, sans jamais hausser le ton, mais en faisant montre d'une autorité sinon plus naturelle, du moins davantage reconnue (figure 45).

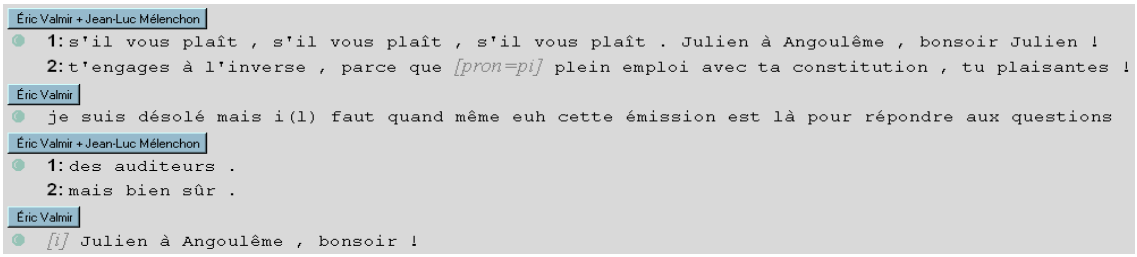


Figure 43 : Exemple d'intervention d'Eric Valmir

Figure 44 : Exemple d'intervention d'Eric Valmir (2)

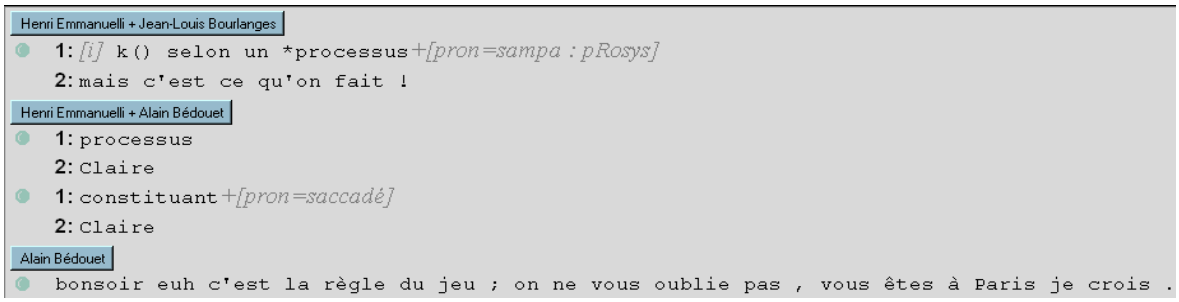


Figure 45 : Exemple d'intervention d'Alain Bédouet

Enfin, nous terminerons par une brève lecture des résultats de Pierre Weill qui, comme nous l'avons dit en préambule, doivent être regardés avec beaucoup de précautions. Toutefois, on notera que pour un nombre de tours de parole finalement très proche de celui des émissions d'Alain Bédouet, le nombre de segments superposés est quant à lui quasiment nul. Nous proposerons deux explications à cela : d'une part, pour les deux thèmes abordés, il n'y a que trois invités à chaque fois, alors que ce chiffre est régulièrement d'au moins quatre. Ensuite, sur ces trois invités, l'un est systématiquement en duplex téléphonique, ce qui limite les interactions puisque seulement deux interlocuteurs sont réellement « présents » sur le plateau de l'émission.

En conclusion, on peut légitimement se poser la question suivante : parle-t-on de la même parole spontanée dans le cas de la télévision et de la radio, au vu des importantes différences d'interaction que nous avons évoquées ? Toute réponse précise est impossible à apporter, car notre corpus de données télévisuelles est trop restreint pour en tirer des enseignements définitifs. Il suffit de voir les importantes différences existant déjà entre deux débats télévisés ayant pourtant beaucoup de points commun (durée, nombre de locuteurs, chaîne...) pour s'en convaincre. Toutefois, il est certain que *Ripostes* laisse apparaître un degré d'interaction qui (et les chiffres du tableau 12 l'attestent) va bien au-delà de ce que nous avons pu voir avec *Le Téléphone Sonne* ou *C dans L'air*. Il serait intéressant, dans l'optique de futurs travaux, d'essayer de rassembler et surtout de transcrire un corpus télévisuel plus conséquent. Davantage de numéros d'une même émission, plus d'émissions différentes, et pourquoi pas des cas particuliers où un animateur interviendrait à la télévision *et* à la radio. Cela permettrait peut-être d'infirmer ou de confirmer certaines hypothèses que nous n'avons pu qu'ébaucher ici,

faute de données en nombre suffisant.

5.4.3 Sous les étoiles exactement : un bon exemple de variation interactionnelle

N°	Durée (min.)	Nombre moyen tours de parole / minute	Durée moyenne d'un tour de parole (sec.)	Nombre moyen de segments superposés / minute	Locuteur 1	Locuteur 2	Locuteur 3	Loc. 4
1	34	10	6	2,1	Serge Levaillant	François Boucq	Karim Belkrouf	
2	32	3	19	0,7	Serge Levaillant	Pierre Santini		
3	34	14	4	1,5	Serge Levaillant	Cali		
4	41	9	6	2,2	Serge Levaillant	Aziz Chouaki	Ariane Gardel	
5	29	17	3	3,8	Serge Levaillant	Belle du Berry	François Jeannin	Mano
6	36	14	4	3,8	Serge Levaillant	Virginie Lemoine	Bernard Mabille	
7	32	9	7	1,5	Serge Levaillant	Monique Ayoun	Georges Wolinski	
8	28	5	11	0,7	Serge Levaillant	Hans Ruedi Giger	Bijan Aalam	
9	36	5	12	1,1	Serge Levaillant	Geoffrey Oryema		

Tableau 15 : Statistiques détaillées de l'émission "Sous les étoiles exactement"

L'émission *Sous les étoiles exactement* offre un panorama représentatif des différents degrés inhérents au concept d'« interaction ». Parmi les neuf émissions que nous avons transcrites, la n°2 (dont l'invité est Pierre Santini) prend le contre-pied complet des moyennes observées généralement sur cette émission. Avec seulement trois tours de parole et moins d'un cas de parole superposée par minute, il obtient des résultats similaires à ceux d'une parole plutôt... préparée. La raison de ce décalage est en fait une corrélation de deux phénomènes. Tout d'abord, une fois n'est pas coutume, Pierre Santini est le seul invité de l'émission. Avec seulement deux locuteurs, l'interactivité, par définition, a déjà de fortes chances d'être moindre. Ensuite, Pierre Santini se révèle dans le cas présent très bavard, et offre aux auditeurs de nombreuses anecdotes sur ses jeunes années. Intimidé, respectueux ou tout simplement passionné, Serge Levaillant n'essaie ni de l'interrompre, ni même de s'engouffrer dans le moindre silence pour reprendre le fil de son émission. Au final, l'acteur parle parfois

pendant plus de quatre minutes sans être interrompu, ce qui est assez rare dans ce type d'émission.

L'émission suivante de *Sous les étoiles exactement* propose un contraste intéressant par rapport à ce que nous venons de développer. A nouveau, il n'y a qu'un invité (le chanteur Cali), mais on dénombre cette fois plus de quatorze tours de parole par minute. La timidité du chanteur et son manque d'aisance face aux médias lui font produire des énoncés très brefs, et souvent hésitants. Serge Levaillant doit régulièrement relancer le débat pour maintenir l'émission à un certain rythme de croisière. Pour autant, le nombre de segments superposés n'évolue pas dans les mêmes proportions : 1,5 par minute, soit un chiffre en dessous de la moyenne globale de l'émission. Les remarques faites précédemment ne sont sans doute pas étrangères à ce chiffre. D'un naturel réservé, Cali confère à l'émission (peut-être de façon tout à fait involontaire) une atmosphère très paisible, sans nuances particulières au niveau des prises de parole, des intonations, etc. En conséquence, le jeu interviewer / interviewé demeure très policé, et pour ainsi dire sans véritable relief. Nul ne semble souhaiter interrompre l'autre, le corriger ou le reprendre. En somme, Serge Levaillant et Cali s'écoutent plus qu'ils ne se parlent.

L'émission n°6 nous offre un point de comparaison très intéressant, puisque le nombre de tours de parole par minute y est presque rigoureusement identique (un peu plus de quatorze). Malgré cela, une différence fondamentale la distingue de la précédente. Il y a cette fois deux invités (Virginie Lemoine et Bernard Mabille), et non un seul. Cette dualité se ressent immédiatement dans le nombre de segments superposés : près de 4 par minute, alors que nous en comptions seulement 1,5 avec Cali. De façon pragmatique, ces chiffres signifient que les interventions des différents locuteurs ne sont en moyenne pas plus brèves (d'après nos relevés, 4 secondes semble être le seuil nécessaire à la bonne tenue d'une discussion), mais que nombre d'entre elles se font de façon simultanée. Cela s'explique assez simplement : en règle générale, plus il y a d'intervenants dans une émission, plus il est difficile pour eux d'obtenir du temps de parole.

Toutefois, ce principe n'est pas toujours vérifié. Il n'est qu'à considérer l'émission du 10 septembre, où Serge Levaillant reçoit Monique Ayoun et Georges Wolinski, soit le même nombre d'invités que dans notre précédent exemple. On note pourtant une nette diminution du nombre de tours de parole (8,5 par minute) et de segments superposés (1,5 par minute). Comment expliquer de tels écarts ? Sans doute en considérant la personnalité des invités.

Bernard Mabille et Virginie Lemoine se connaissent très bien et depuis longtemps, et sévissent tous deux dans le domaine de l'humour. En conséquence, on perçoit une très grande aisance lors de leurs prestations médiatiques. Ils n'hésitent pas à s'interrompre, à parler en même temps, à plaisanter... et Serge Levaillant lui-même se prend au jeu, d'où les chiffres particulièrement élevés que nous avons mentionnés ci-dessus.

Monique Ayoun et Georges Wolinski, quant à eux, ne sauraient être rangés dans la même catégorie. Ils ne sont certes pas mal à l'aise face à l'exercice de l'interview, bien au contraire, mais ils ne s'interrompent mutuellement que très peu. Pendant une bonne partie de l'émission, chacun écoute attentivement ce que dit l'autre, sans jamais intervenir. Sans doute est-ce dû au fait – information que les invités nous donnent eux-mêmes – qu'ils ne se connaissaient guère avant de collaborer sur le livre qu'ils viennent promouvoir ensemble. L'atmosphère, bien que très détendue, n'est ainsi pas comparable avec celle de l'émission précédente.

Enfin, pour terminer cette analyse détaillée des émissions de *Sous les étoiles exactement*, nous évoquerons conjointement les émissions n°8 et 9. Elles ont un point commun : le ou les invités n'ont pas le français pour langue maternelle. En conséquence, leur expression n'est pas aussi naturelle et spontanée que celle des autres invités. De son côté, Serge Levaillant, soucieux de laisser ses invités s'exprimer, intervient beaucoup moins qu'à l'accoutumée et prend grand soin de ne pas les interrompre. Au niveau des chiffres, cela donne des résultats jusqu'à deux fois inférieurs à la moyenne générale de l'émission.

5.5 Conclusion

La notion de *disfluence* a été le fil conducteur de cet ultime chapitre, et assurément l'un de ceux de nos travaux en général. L'expérience menée en 5.1 prouve par les chiffres qu'elle est très étroitement liée au taux d'erreur mot des SRAP. Mais ses rapports avec la parole spontanée sont-ils aussi patents ? Rien n'est moins sûr. Si l'on se rappelle les trois questions que nous posions à l'issue de cette l'expérience mentionnée ci-dessus, il semble désormais plus simple d'y apporter quelques éléments de réponse :

- *Peut-on envisager la parole spontanée sous d'autres angles que celui des « disfluences » mentionnées dans cette étude ?*

Il nous semble à présent que oui. Les disfluences, bien sûr, occupent sur les plans morpho-syntaxique, lexical et phonologique une place considérable. Par ailleurs, ce sont ces aspects qui posent le plus de problèmes aux SRAP, problèmes qui pour certains sont difficilement résolvables sans une aide humaine : un SRAP pourra-t-il jamais faire la différence entre « c'est pas ce que je veux manger » et « c'est parce que je veux manger », lorsqu'il sera confronté à la séquence acoustique [sepask@Zv2m~aZe] ?

Cela étant, l'énonciation et la prosodie sont par exemple deux autres angles d'approches qui permettent d'envisager la parole spontanée de façon pertinente, sans placer les disfluences au centre de l'analyse. [Morel et Danon-Boileau, 1998] ont notamment démontré que la fréquence fondamentale ou la durée sont deux paramètres prosodiques mettant en exergue la parole spontanée, en ce sens que leurs valeurs sont radicalement différentes lorsqu'elles s'appliquent à de la parole préparée. On observe ainsi plus de chutes que de montées de F0 dans un cadre spontané, et inversement pour un contexte préparée. De même, on observe dans la parole préparée une certaine « isochronie » de la durée des syllabes, alors qu'elle fluctue au gré des hésitations et des émotions d'un locuteur qui tient un propos spontané.

- *Les disfluences sont-elles un critère de spontanéité toujours pertinent ?*

La distinction parole préparée / parole spontanée que nous venons d'évoquer prend ici tout son sens. Notre expérience en 5.3 l'a montré : si l'on ne retient que les disfluences comme critère discriminatoire, certains contextes énonciatifs peuvent prêter à confusion. S'il est

parfois difficile d'expliquer pourquoi certains locuteurs sont plus aptes que d'autres à tenir un discours fluide, nous avons tenté d'émettre quelques hypothèses pour expliquer cet état de fait : émotions, faculté à concevoir efficacement et rapidement une réponse... Toujours est-il que les résultats que nous avons obtenus placent sur un plan presque identique un propos journalistique et certains entretiens essentiellement spontanés. S'il est donc réducteur d'affirmer que les disfluences sont systématiquement synonymes d'un contexte spontané, il ne faut pas pour autant oublier que les cas comme celui de Jacques-Alain Miller demeurent plutôt marginaux.

- Un propos spontané ne doit-il pas aussi être vu via son degré d'interaction avec le public auquel il s'adresse ?

C'est à cette dernière question que nous nous sommes efforcés de répondre dans la quatrième partie du présent chapitre. À la lecture des résultats obtenus, il est au moins possible d'affirmer qu'il existe des degrés d'interaction radicalement différents, dans des situations pourtant toutes considérées comme spontanées (en ce sens qu'elles sont soit des interviews, soit des entretiens, soit des débats).

Dans un premier temps, un peu comme nous l'avons fait avec les disfluences, nous avons démontré qu'un contexte préparé était parfois très semblable à un contexte spontané, mais cette fois du point de vue de l'interaction : peu ou pas de parole superposée, des tours de parole très longs, etc. La conclusion à en tirer est d'ailleurs également similaire à ce que nous énoncions ci-dessus : parole préparée et parole spontanée peuvent se confondre sous certains angles d'approche, mais dans des cas qui restent assez peu fréquents.

Ensuite, nous avons mis à profit le fait de disposer d'un petit corpus de transcriptions d'émissions télévisées pour voir s'il existait « plusieurs » paroles spontanées, selon le média auxquels les locuteurs étaient confrontés. Nous avons ainsi pu montrer, principalement à travers l'exemple de l'émission *Ripostes*, que la télévision générait une interactivité beaucoup plus importante, notamment grâce au support de l'image. Celui-ci autorise des prises de parole beaucoup plus soudaines et des propos simultanés très fréquents, puisqu'assurant toujours au (télé)spectateur des repères visuels qui compensent l'éventuelle confusion auditive. Par ailleurs, le tempérament des intervenants n'est pas non plus étranger aux écarts obtenus dans nos résultats. Et à ce titre, nous avons analysé en détails l'émission *Sous les étoiles exactement*, dont nous disposions de nombreuses transcriptions. Cette étude nous a en effet

permis de constater que non seulement le nombre, mais aussi et surtout la personnalité des intervenants avaient une importance considérable dans le degré d'interactivité d'une situation énonciative.

Il nous semble donc, au vu des écarts importants observés tout au long de cette étude, que l'interactivité, notion sous-entendue lorsque l'on parle de parole *conversationnelle*, ne doit pas être prise à la légère lorsqu'on s'intéresse à la catégorisation de la parole spontanée. Au même titre que les disfluences, le nombre de segments superposés ou de tours de parole permettent d'établir une hiérarchie laissant à penser qu'il existe bel et bien non pas une, mais plusieurs paroles spontanées.

Conclusion générale et perspectives

Tout au long de cette étude, nous avons tenté de circonscrire un objet dont le qualificatif même est sujet à débat : la parole *spontanée*. Le projet EPAC, dans sa description scientifique, lui préférait par exemple l'adjectif *conversationnelle*. Nous avons justifié à plusieurs reprises notre position lexicale sur ce point mais qui pourrait, après avoir lu ce travail, affirmer que l'une est plus « vraie » que l'autre ? Que l'on ne s'y trompe pas : la parole spontanée, la parole conversationnelle, l'oral, la *langue de tous les jours*... Aucun de ces qualificatifs n'est réellement plus approprié qu'un autre tant qu'il n'est pas associé à un objet dont les contours doivent être clairement définis.

Les disfluences sont une couche de ce contour, peut-être la plus évidente si on ne se confronte pas directement aux données. Car c'est lorsque l'on se trouve face à des locuteurs comme Stéphane Braunschweig et Jacques-Alain Miller que la pertinence des répétitions, *fillers* et faux départs vole en éclat. Mais les disfluences ne sont pas les seuls éléments permettant de cataloguer un propos comme étant *spontané*. Nous avons eu l'occasion de montrer que des aspects prosodiques, énonciatifs, phonétiques ou phonologiques étaient également autant d'angles à envisager pour distinguer, de façon plus ou moins patente, ce qui pouvait se rapprocher ou non de notre définition de la spontanéité.

Celle-ci englobait également un paramètre qui nous semble fondamental : l'interaction. Qu'est-ce en effet que la spontanéité d'un propos si elle ne trouve pas un, voire plusieurs échos ? De notre point de vue, cet aspect est essentiel, au même titre que les disfluences par exemple. Il permet de distinguer explicitement, comme nous l'avons montré dans la dernière partie de cette thèse, les débats animés des interviews tournant au monologue. C'est là que le terme *conversationnel* est riche de sens : pour que de la parole puisse être conversationnelle, il faut par essence qu'il y ait une conversation, et donc un véritable échange d'opinions, de regards, de gestes et pourquoi pas même de disfluences.

Avoir travaillé dans et avec des corpus – et plus encore en avoir constitué un – nous

permet aujourd'hui de dire que l'on ne peut résumer la parole spontanée à une parole disfluente. L'expérience détaillée dès le début du présent manuscrit va dans ce sens en insistant plutôt sur la notion de « qualité d'élocution », pierre angulaire de notre raisonnement. Si cette qualité est bonne, le travail du transcripteur sera facilité : moins d'annotations concernant les disfluences, moins d'écoutes répétées d'un même passage pour en comprendre le sens, moins de corrections à apporter si le travail est assisté... En un mot, le transcripteur sera plus efficace. Et cette efficacité n'est pas le moindre des avantages lorsque l'on regarde le deuxième chapitre : quelle que soit la qualité d'élocution, il y a toujours besoin d'un transcripteur, même lorsque le travail manuel est précédé par le déblayage d'un SRAP performant. La transcription parfaite n'existe pas (encore), pour peu que l'on souhaite un résultat de qualité.

Pour ce faire, des outils sont certes disponibles, eux qui ont été développés dans le but de rendre moins contraignantes et surtout plus pérennes la transcription et l'annotation de gros corpus. Tous ne sont pas dédiés aux mêmes types de données, et leur efficacité s'en ressent suivant ce pour quoi on les sollicite. Cependant, tous ont leur communauté d'utilisateurs plus ou moins attitrés, et on ne peut que louer le fait que la plupart d'entre eux appartiennent au monde des logiciels libres. Il est assurément plus regrettable que diversité ne rime pas ici avec interopérabilité, puisque chacun de ces outils a un format de sortie spécifique, plus ou moins facilement échangeable et diffusable. A l'heure où les projets de transcription à grande échelle – notamment francophones – se multiplient (EPAC, mais aussi et surtout les ESLO, VARILING, etc.), le moment semble venu de s'accorder sur un format de sortie commun, dont la TEI semble être le meilleur représentant à l'heure actuelle. Si certains groupes ou laboratoires n'ont pas attendu cette étude pour s'y intéresser (CLAPI, CATCOD), ils sont encore trop marginaux pour peser suffisamment dans la balance.

Il ne faudrait pourtant pas négliger une telle possibilité, ouvrant des perspectives de conservation et d'archivage des données sans précédents. Il y a moins d'un siècle, le magnétophone n'existait pas et personne n'était donc en mesure de capturer le moindre flux de voix ; quant à la transcription, elle se résumait à des bribes de conversations recueillis sur quelques feuillets épars, avec plus ou moins de rigueur et un très faible espoir de pouvoir conserver ces précieuses données dans un état acceptable au fil du temps.

Aujourd'hui, tout a changé : on peut enregistrer des heures de conversation en qualité numérique avec un appareil mesurant quelques centimètres seulement. Ces enregistrements

peuvent être copiés, gravés, hébergés et même retravaillés à l'envi. Une seule chose, au fond, n'a pas réellement évolué : il y a toujours besoin d' « enquêteurs » pour aller recueillir les enregistrements sur le terrain et, comme nous le disions plus haut, de transcripteurs pour traiter ces données.

Mais il convient aussi et surtout de s'interroger sur l'exploitation desdites données : enregistrer, transcrire, annoter... oui, mais pour quoi faire ? La question peut paraître dérisoire ou ironique, mais elle se veut surtout le reflet de situations – réelles – pour le moins étonnantes : il existe aujourd'hui, enfermés dans des placards, des corpus auxquels (presque) personne n'a accès. Lorsque l'on sait le coût que représente de tels entreprises, on peut être surpris que le commun des chercheurs ne puisse en bénéficier.

D'autant que ces données sont aujourd'hui très rares, du moins en France. Les corpus journalistiques existent en quantité raisonnable, mais leurs équivalents spontanés demeurent beaucoup moins nombreux. Les raisons en sont peut-être quantifiables (disponibilité des données audio, qualité de celles-ci, temps nécessaires à la transcription et à l'annotation...), mais elles ne doivent pas décourager les chercheurs désirant s'intéresser à un objet qui, jour après jour, ne cesse d'évoluer au gré des modes, des générations et des endroits où l'on se trouve. Autant d'opportunités et de raisons, en somme, de le saisir au vol pour en regarder les mécanismes. Faute de quoi, ce sont des pans entiers de son évolution qui risquent de nous échapper.

Bibliographie personnelle

Revues d'audience internationale avec comité de sélection

- T. Bazillon, V. Jousse, F. Béchet, Y. Estève, G. Linarès, D. Luzzati (2008). « La parole spontanée : transcription et traitement », in *Traitement automatique des langues*, vol. 49, n°3.

Conférences d'audience internationale avec comité de sélection

- P. Clément, T. Bazillon, C. Fredouille (2011). « Speaker diarization of heterogeneous web radio files: A preliminary study », ICASSP 2011, 22-27 mai 2011, Prague (République Tchèque).

- L. M. Rojas, T. Bazillon, M. Quignard, F. Lefèvre (2011). « Using MMIL for the High Level Semantic Annotation of the French MEDIA Dialogue Corpus », IWCS 2011, 12-14 janvier 2011, Oxford (Royaume-Uni).

- Y. Estève, T. Bazillon, J.-Y. Antoine, F. Béchet, J. Farinas (2010). « The EPAC corpus: manual and automatic annotation of conversational speech in French broadcast news », LREC 2010, 17-23 mai 2010, Malte.

- T. Bazillon, Y. Estève, D. Luzzati (2008). « Le codage des corpus oraux », CATCOD 2008, 4-5 décembre 2008, Orléans (France).

- T. Bazillon, Y. Estève, D. Luzzati (2008). « Manual *vs* assisted transcription of prepared and spontaneous speech », LREC 2008, 28-30 mai 2008, Marrakech (Maroc).

Conférences d'audience nationale avec comité de sélection

- T. Bazillon, B. Maza, M. Rouvier, F. Béchet, A. Nasr (2011) « Qui êtes-vous ? Catégoriser les questions pour déterminer le rôle des locuteurs dans des conversations orals », TALN 2011, 27 juin-1er juillet 2011, Montpellier (France).

- T. Bazillon, Y. Estève, D. Luzzati (2008). « Transcription manuelle *vs* assistée de la parole préparée et spontanée », JEP 2008, 9-13 juin 2008, Avignon (France).

- V. Jousse, Y. Estève, F. Béchet, T. Bazillon, G. Linarès (2008). « Caractérisation et détection de parole spontanée dans de larges collections de documents audio », JEP 2008, 9-13 juin 2008, Avignon (France).

- T. Bazillon (2007). « Le codage de la parole spontanée pour la reconnaissance automatique de la parole », RJCP 2007, 5-6 juillet 2007, Paris (France).

Bibliographie

[Abouda et Baude, 2005] Lotfi ABOUDA et Olivier BAUDE (2005). « Du Français Fondamental aux ESLO », colloque *Français fondamental et corpus oraux*, 8-10 décembre 2005, Lyon (France).

[Abouda et Baude, 2006] Lotfi ABOUDA et Olivier BAUDE (2006). « Constituer et exploiter un grand corpus oral : choix et enjeux théoriques. Le cas des ESLO », colloque *Corpus en lettres et sciences sociales – des documents numériques à l'interprétation*, 10-14 juillet 2006, Albi (France).

[Adda *et al.*, 2005] Gilles ADDA, Martine ADDA-DECKER, Claude BARRAS, Frédérique BÉNARD, Philippe BOULA DE MAREÜIL, Benoît HABERT et Patrick PAROUBEK (2005). « Disfluences et traitement automatique : l'Heure de vérité », journée d'étude de l'ATALA : *Hésitations, disfluences, répétitions, faux départs : quel ordre dans le désordre ?*, 2 avril 2005, Paris (France).

[Adda-Decker, 2006] Martine ADDA-DECKER (2006). « De la reconnaissance automatique de la parole à l'analyse linguistique de corpus oraux », JEP 2006, 12-16 juin 2006, Dinard (France).

[Adda-Decker, 2007] Martine ADDA-DECKER (2007). « Problèmes posés par le schwa en reconnaissance et en alignement automatique de la parole », JEL 2007, 27-28 juin 2007, Nantes (France).

[Adda-Decker *et al.*, 2004] Martine ADDA-DECKER, Benoît HABERT, Claude BARRAS, Gilles ADDA, Philippe BOULA DE MAREÜIL et Patrick PAROUBEK (2004). « 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)

[Afitska, 2009] Oksana AFITSKA (2009). « Transana 2.30 », *Language Documentation & Conservation* vol. 3, n° 2, pp. 226-235.

[Albrespit, 2007] Jean ALBRESPIT (2007). « Y a-t-il une grammaire de l'oral ? », 6ème Journée des langues vivantes, 14 novembre 2007, CDDP de la Gironde (France).

[Anderson, 2008] Wendy ANDERSON (2008). « Corpus Linguistics in the UK: Ressources for Sociolinguistic Research », in *Language and Linguistics Compass* 2/2, pp. 352-371.

[André et Goossens, 1995] Jacques ANDRÉ et Michel Goossens (1995). « Codage des caractères et multi-linguisme : de l'ASCII à UNICODE et ISO/IEC-10646 », in *Cahiers GUTenberg* n°20, mai 1995.

[Antoine, 2002] Jean-Yves ANTOINE (2002). « Corpus OTG : présentation générale », document téléchargeable à l'adresse : http://www.info.univ-tours.fr/~antoine/parole_publique/OTG/Pres_OTG.pdf

[Antoine *et al.*, 2002] Jean-Yves ANTOINE, Sabine LETELLIER-ZARSHENAS, Pascale NICOLAS, Igor SCHADLE et Jean CAELEN (2002). « Corpus OTG et ECOLE_MASSY : vers la constitution d'une collection de corpus francophones de dialogue oral diffusés librement », TALN 2002, 24-27 juin 2002, Nancy (France).

[Antoine, 2008] Jean-Yves ANTOINE (2008). « Outils d'aide à la transcription », séminaire à Paris X Nanterre, juin 2008.

[ASILA, 2003] Emmanuel SCHANG, Olivier BAUDE, Maria Caterina MANES GALLO, Lukas BALTHASAR, Lorenza MONDADA, Anne LACHERET, Jean-Jacques GIRARDOT, Serge HEIDEN, Philippe MARTIN, Michel JACOBSON, Jean-Yves ANTOINE, Daniel LUZZATI, Matthieu QUIGNARD, Christian PLANTIN, Christian BRASSAC, Laurent ROMARY, Anne NICOLLE et Laurence DEVILLERS (2003). Workshop ASILA, 23-25 juin 2003, La Bresse (France).

[Bahou *et al.*, 2010] Younès BAHOU, Abir MASMOUDI et Lamia Hadrich BELGUITH (2010). « Traitement des disfluences dans le cadre de la compréhension automatique de l'oral arabe spontané », TALN 2010, 19-23 juillet 2010, Montréal (Canada).

[Bally, 1929] Charles BALLY (1929). *Traité de stylistique française*, Klincksieck, Paris.

[Barras *et al.*, 1998] Claude BARRAS, Édouard GEOFFROIS, Zhibiao WU et Mark LIBERMAN (1998). « Transcriber: a free tool for segmenting, labeling and transcribing speech », LREC 98, 28-30 mai 1998, Grenade (Espagne).

[Barras *et al.*, 2000] Claude BARRAS, Édouard GEOFFROIS, Zhibiao WU et Mark LIBERMAN (2000). « Transcriber: development and use of a tool for assisting speech corpora production », *Speech Communication special issue on Speech Annotation and Corpus Tools*, vol. 33, n°1-2.

[Baude, 2006] Olivier BAUDE (2006). « Corpus oraux : les bonnes pratiques d'une communauté scientifique », colloque *Corpus en lettres et sciences sociales – des documents numériques à l'interprétation*, 10-14 juillet 2006, Albi (France).

[Baude, 2008] Olivier BAUDE (2008). « Remarques sur quelques corpus disponibles », Journée Rhapsodie *Corpus Design*, 4 avril 2008, Paris (France).

[Bazillon *et al.*, 2008a] Thierry BAZILLON, Vincent JOUSSE, Frédéric BÉCHET, Yannick ESTÈVE, Georges LINARÈS et Daniel LUZZATI (2008). « La parole spontanée : transcription et traitement », in *Traitement automatique des langues*, vol. 49, n°3.

[Bazillon *et al.*, 2008b] Thierry BAZILLON, Yannick ESTÈVE et Daniel LUZZATI (2008). « Manual *vs* assisted transcription of prepared and spontaneous speech », LREC 2008, 28-30 mai 2008, Marrakech (Maroc).

[Bazillon *et al.*, 2008c] Thierry BAZILLON, Yannick ESTÈVE et Daniel LUZZATI (2008). « Le codage des corpus oraux », CATCOD 2008, 4-5 décembre 2008, Orléans (France).

[Berez, 2007] Andrea L. BEREZ (2007). « EUDICO Linguistic Annotator (ELAN) », *Language Documentation & Conservation* vol. 1, n° 2, pp. 283-289.

[Benzitoun et Véronis, 2005] Christophe BENZITOUN et Jean VÉRONIS (2005). « Problèmes d'annotation d'un corpus oral dans le cadre de la campagne EASY », TALN 2005, 6-10 juin 2005, Dourdan (France).

[Biber, 2006] Douglas BIBER (2006). *University language: A corpus-based study of spoken and written registers*, Amsterdam: John Benjamins.

[Bilger et Cappeau, 2004] Mireille BILGER et Paul CAPPEAU (2004). « Ce que les corpus nous apprennent sur la langue », in C. Vargas (éd.), *Langue et étude de langue. Approches linguistiques et didactiques*, Presses Universitaires d'Aix-en-Provence, pp. 59-68.

[Binnenpoorte *et al.*, 2003] Diana BINNENPOORTE, Simo GODDIJN et Catia CUCCHIARINI (2003). « How to Improve Human and Machine Transcriptions of Spontaneous Speech », SSPR 2003, 13-16 avril 2003, Tokyo (Japon).

[Blanche-Benveniste, 2010] Claire BLANCHE-BENVENISTE (2010). *Approches de la langue parlée en français*, Paris : Ophrys.

[Blanche-Benveniste et Bilger, 1999] Claire BLANCHE-BENVENISTE et Mireille BILGER (1999). « Français parlé – oral spontané. Quelques réflexions. », in *Revue française de linguistique appliquée* n°IV-2, pp. 21-30.

[Blanche-Benveniste et Jeanjean, 1987] Claire BLANCHE-BENVENISTE et Colette JEANJEAN (1987). *Le français parlé. Transcription et édition*, Paris : Didier Érudition.

[Blanche-Benveniste *et al.*, 1990] Claire BLANCHE-BENVENISTE, Mireille BILGER, Christine ROUGET et Karel VAN DEN EYDE (1990). *Le français parlé : études grammaticales*, CNRS Éditions, Paris.

[Boersma, 2001] Paul BOERSMA (2001). « Praat, a system for doing phonetics by computers », *Glot International* 5:9/10, pp. 341-345.

[Bonneau-Maynard *et al.*, 2005] Hélène BONNEAU-MAYNARD, Sophie ROSSET, Anne KUHN et Djamel MOSTEFA (2005). « Semantic annotation of the French Media dialog corpus », Interspeech 2005, 4-8 septembre 2005, Lisbonne (Portugal).

[Boula de Mareüil *et al.*, 2005] Philippe BOULA DE MAREÜIL, Benoît HABERT, Frédérique BÉNARD, Martine ADDA-DECKER, Claude BARRAS, Gilles ADDA et Patrick PAROUBEK (2005). « A quantitative study of disfluencies in French broadcast interviews », DiSS 2005, 10-12 septembre 2005, Aix-en-Provence (France).

[Boves et Oostdijk, 2003] Lou BOVES et Nelleke OOSTDIJK (2003). « Spontaneous Speech in the Spoken Dutch Corpus », SSPR 2003, 13-16 avril 2003, Tokyo (Japon).

[Brinckmann, 2008] Caren BRINCKMANN (2008). « Transcription bottleneck of speech corpus exploitation », Lesser Used Languages and Computer Linguistics (LULCL) II, 13-14 novembre 2008, Bolzano (Italie).

[Caelen-Haumont, 2002] Geneviève CAELEN-HAUMONT (2002). « Prosodie et dialogue spontané : Valeurs et fonctions perlocutoires du mélisme », in *TIPA* n°21, pp. 13-24.

[Candea, 2000a] Maria CANDEA (2000). « Les euh et les allongements dits “d'hésitation” : deux phénomènes soumis à certaines contraintes en français oral non lu », JEP 2000, 19-23 juin 2000, Aussois (France).

[Candea, 2000b] Maria CANDEA (2000). *Contribution à l'étude des pauses silencieuses et des phénomènes dits d'hésitation en français oral spontané. Étude sur un corpus de récits en classe de français*, Thèse de doctorat, Université de Paris III.

[Cappeau, 2007] Paul CAPPEAU (2007). « Que peuvent nous apprendre en syntaxe des corpus oraux anciens ? », colloque *Études de syntaxe : français parlé, français hors de France, créoles*, 19 octobre 2007, Paris (France).

[Cappeau et Seijido, 2005] Paul CAPPEAU et Magali SEIJIDO (2005). « Les corpus oraux en français », document téléchargeable à l'adresse : http://www.dglf.culture.gouv.fr/recherche/corpus_parole/Presentation_Inventaire.pdf

[Carvajal, 2008] Isabel Colon de CARVAJAL (2008). « Tableau Comparatif – Logiciels de Transcription », document téléchargeable à l'adresse : http://icar.univ-lyon2.fr/projets/corinte/documents/CORINTE_Tableau_ComparatifPA.pdf

[Cerf-Danon *et al.*, 1991] Hélène CERF-DANON, Marc EL-BÈZE et Bernard MERIALDO (1991). « Reconnaissance automatique de la parole », in *Informatique et Santé* vol. 4 : *Nouvelles Technologies et Traitement de l'Information en Médecine*, Springer-Verlag France, Paris (France).

[Chevalier, 2000] Gisèle CHEVALIER (2000). « 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.

[Clopper et Pisoni, 2006] Cynthia G. CLOPPER et David B. PISONI (2006). « The Nationwide Speech Project: A new corpus of American English dialects », in *Speech Communication* n°48, pp. 633-644.

[Cohen, 1960] Jacob COHEN (1960). « A coefficient of agreement for nominal scales », in *Educational and Psychological Measurement* n°20, pp. 37-46.

[Coquillon, 2007] Annelise COQUILLON (2007). « Le français parlé à Marseille : exemple d'un locuteur PFC », in *Bulletin PFC* n°7, pp. 145-156.

[Corley et Stewart, 2008] Martin CORLEY et Olivier W. STEWART (2008). « Hesitation disfluencies in spontaneous speech: The meaning of *um* », in *Language and Linguistics Compass* vol. 2, n°4, pp. 589-602.

[Coursil, 2000] Jacques COURSIL (2000). *La fonction muette du langage*, Ibis Rouge Éditions, Matoury (Guyane française).

[Damourette et Pichon, 1911-1940] Jacques DAMOURETTE et Édouard PICHON (1911-1940). *Des mots à la pensée, essai de grammaire de la langue française*, D'Artrey, Paris.

[Delais-Roussarie *et al.*, 2006] Élisabeth DELAIS-ROUSSARIE, Geneviève CAELEN-HAUMONT, Daniel HIRST, Philippe MARTIN et Piet MERTENS (2006). « Outils d'aide à l'annotation prosodique de corpus », in *Bulletin PFC* n°6, pp. 7-26.

[Deléglise *et al.*, 2005] Paul DELÉGLISE, Yannick ESTÈVE, Sylvain MEIGNIER et Teva MERLIN (2005). « The LIUM Speech Transcription System: A CMU Sphinx III-Based System for French Broadcast News », Interspeech 2005, 4-8 septembre 2005, Lisbonne (Portugal).

[Deulofeu, 1977] Henri-José DEULOFEU (1977). « Recherches sur l'ordre des syntagmes en français : le cas des groupes prépositionnels », Thèse de 3ème cycle, Université de Paris III.

[Deulofeu, 2001] Henri-José DEULOFEU (2001). « L'innovation linguistique en français contemporain : mythes tenaces et réalité complexe », in *Le français dans le monde* n°21, “Oral : variabilité et apprentissages”.

[Dewaele, 2001] Jean-Marc DEWAELE (2001). « Une distinction mesurable : corpus oraux et écrits sur le continuum de la deixis », in *Journal of French Language Studies* n°11, Presses Universitaires de Cambridge, pp. 179-199.

[Dister et Simon, 2008] Anne DISTER et Anne-Catherine SIMON (2008). « La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé », in *Arena Romanistica* 1/1, pp. 54-79.

[Draxler, 2005] Christoph DRAXLER (2005). « WebTranscribe – An Extensible Web-Based Speech Annotation Framework », TSD 2005, 12-15 septembre 2005, Karlovy Vary (République tchèque).

[Dufour, 2010] Richard DUFOUR (2010). *Transcription automatique de la parole spontanée*, Thèse de doctorat, Université du Maine, Le Mans.

[Dufour et Estève, 2008] Richard DUFOUR et Yannick ESTÈVE (2008). « Correcting ASR outputs: specific solutions to specific errors in French », IEEE Workshop on Spoken Language Technology (SLT 2008), 15-18 décembre 2008, Goa (Inde).

[Dufour *et al.*, 2009] Richard DUFOUR, Yannick ESTÈVE, Paul DELÉGLISE et Frédéric BÉCHET (2009). « Local and global models for spontaneous speech segment detection and characterization », ASRU 2009, 13-17 décembre 2009, Merano (Italie).

[Durand et Tarrier, 2006] Jacques DURAND et Jean-Michel TARRIER (2006). « PFC, corpus et systèmes de transcription », in *Cahiers de Grammaire* n° 30, pp. 139-158.

[Durand *et al.*, 2002] Jacques DURAND, Bernard LAKS et Chantal LYCHE (2002). « Synopsis du projet PFC, La phonologie du Français contemporain : Usages, Variétés et Structure », in *Bulletin PFC* n°1, pp. 5-7.

[Durand *et al.*, 2005] Jacques DURAND, Bernard LAKS et Chantal LYCHE (2005). « Un corpus numérisé pour la phonologie du français », in G. Williams (ed.) *La linguistique de corpus,* Presses Universitaires de Rennes, pp. 205-217.

[Estève *et al.*, 2004] Yannick ESTÈVE, Paul DELÉGLISE et Bruno JACOB (2004). « Systèmes de transcription automatique de la parole et logiciels libres », in *Traitement Automatique des Langues* vol. 45, n° 2.

[Estève *et al.*, 2010] Yannick ESTÈVE, Thierry BAZILLON, Jean-Yves ANTOINE, Frédéric BÉCHET et Jérôme Farinas (2010). « The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news », LREC 2010, 19-21 mai 2010, Malte.

[Fillmore *et al.*, 1998] Charles J. FILLMORE, Nancy IDE, Catherine MACLEOD et Daniel JURAFSKY (1998). « An American National Corpus: a proposal », LREC 1998, 28-30 mai 1998, Grenade (Espagne).

[Fosler-Lussier et Morgan, 1999] Éric FOSLER-LUSSIER et Nelson MORGAN (1999). « Effects on speaking rate and word frequency on pronunciations in conversational speech », in *Speech Communication* vol. 29, pp. 137-158.

[Frei, 1929] Henri FREI (1929). *La grammaire des fautes*, Slatkine, Genève.

[Gadet, 1996] Françoise GADET (1996). *Le français ordinaire*, deuxième édition revue et augmentée, Paris : Armand Colin.

[Gadet, 2007] Françoise GADET (2007). *La variation sociale en français*, nouvelle édition revue et augmentée, Paris : Ophrys.

[Galliano *et al.*, 2005] Sylvain GALLIANO, Édouard GEOFFROIS, Djamel MOSTEFA, Khalid CHOUKRI, Jean-François BONASTRE et Guillaume GRAVIER (2005). « The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast New », Interspeech 2005, 4-8 septembre 2005, Lisbonne (Portugal).

[Garcia-Fernandez et Lailler, 2008] Anne GARCIA-FERNANDEZ et Carole LAILLER (2008). « Morphosyntaxe de l'interrogation pour le système question-réponse RITEL », RECITAL 2008, 9-13 juin 2008, Avignon (France).

[Garric et Léglise, 2007] Nathalie GARRIC et Isabelle LÉGLISE (2007). « Aspects syntaxiques et discursifs d'un français parlé des médias : le discours d'information télévisé », in Mathias Broth, Mats Forsgren, Coco Norén et François Sullet-Nylander (éds.), *Le français parlé des médias*, pp. 243-258.

[Gauvain *et al.*, 2004] Jean-Luc GAUVAIN, Gilles ADDA, Lori LAMEL, Fabrice LEFÈVRE, Holger SCHWENK (2004). « Transcription de la parole conversationnelle », JEP 2004, 19-22 avril 2004, Fès (Maroc).

[Gendner et Adda-Decker, 2002] Véronique GENDNER et Martine ADDA-DECKER (2002). « Analyse comparative de corpus oraux et écrits français : mots, lemmes et classes morpho-syntaxiques », JEP 2002, 24-27 juin 2002, Nancy (France).

[Gendrot, 2005] Cédric GENDROT et Frédérique BÉNARD (2005). « Propositions de Normalisation pour une Base de Corpus Multimédia à l'ED268 », RJCED268 2005, 21 mai 2005, Paris (France).

[Geoffrois *et al.*, 2000] Édouard GEOFFROIS, Claude BARRAS, Steven BIRD et Zhibiao WU (2000). « Transcribing with Annotation Graphs », LREC 2000, 31 mai-2 juin 2000, Athènes (Grèce).

[Geoffrois, 2004] Édouard GEOFFROIS (2004). « Le traitement automatique de la parole pour la recherche d'informations dans les documents audio », in *Revue Scientifique et Technique de la Défense* n°64.

[Goldman, 2006] Jean-Philippe GOLDMAN (2006). « Tutoriel PRAAT », document téléchargeable à l'adresse :
http://latlcui.unige.ch/phonetique/easyalign/tutorielpraat.pdf

[Gougenheim *et al.*, 1964] Georges GOUGENHEIM, René MICHEA, Paul RIVENC et Aurélien SAUVAGEOT (1964). *L'élaboration du français fondamental : étude sur l'établissement d'un vocabulaire et d'une grammaire de base*, Didier, Paris.

[Guénot, 2005] Marie-Laure GUÉNOT (2005). « Parsing de l'oral : traiter les disfluences », TALN 2005, 6-10 juin 2005, Dourdan (France).

[Hellwig *et al.*, 2009] Birgit HELLWIG, Diter van UYTVANCK et Micha HULSBOSCH (2009). « ELAN – Linguisitc Annotator version 3.8 », document téléchargeable à l'adresse :
http://www.mpi.nl/corpus/manuals/manual-elan.pdf

[Henry, 2002] Sandrine HENRY (2002). « Étude des répétitions en français parlé spontané pour les technologies de la parole », RECITAL 2002, 24-27 juin 2002, Nancy (France).

[Jacobson, 2002] Michel JACOBSON (2002). « Les outils modernes pour la transcription de la parole », in *Revue Parole* n°22/23/24, pp. 213-229.

[Jelinek, 1988] Frederick JELINEK (1988). « Applying Information Theoritic Methods: Evaluation of Grammar Quality », Workshop Evaluation of NLP Systems, Wayne PA, décembre 1988.

[Jelinek, 2004] Frederick JELINEK (2004). « Some of my best friends are linguists », LREC 2004, 26-28 mai 2004, Lisbonne (Portugal).

[Johannessen *et al.*, 2007] Janne Bondi JOHANNESSEN, Kristin HAGEN, Joel PRIESTLEY et Lars NYGAARD (2007). « An Advanced Speech Corpus for Norwegian », NODALIDA 2007, 25-26 mai 2007, Tartu (Estonie).

[Jousse *et al.*, 2008] Vincent JOUSSE, Yannick ESTÈVE, Frédéric BÉCHET, Thierry BAZILLON et Georges LINARÈS (2008). « Caractérisation et détection de parole spontanée dans de larges collections de documents audio », JEP 2008, 9-13 juin 2008, Avignon (France).

[Jousse *et al.*, 2009a] Vincent JOUSSE, Sylvain MEIGNIER, Christine JACQUIN, Simon PETITRENAUD, Yannick ESTÈVE, Béatrice DAILLE (2009), « Analyse conjointe du signal sonore et de sa transcription pour l'identification nommée de locuteur », in *Traitement automatique des langues* vol. 50, n°1.

[Jousse *et al.*, 2009b] Vincent JOUSSE, Simon PETITRENAUD, Sylvain MEIGNIER, Yannick ESTÈVE, Christine JACQUIN (2009), « Automatic named identification of speakers using diarization and asr systems », ICASSP 2009, 19-24 avril 2009, Taïpei (Taïwan).

[Kawahara *et al.*, 2003] Tatsuya KAWAHARA, Hiroaki NANJO, Takahiro SHINOZAKI et Sadaoki FURUI (2003). « Benchmark Test for Speech Recognition Using the Corpus of Spontaneous Japanese », ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition, 13-16 avril 2003, Tokyo (Japon).

[Kelley, 1984] John F. KELLEY (1985). « An iterative design methodology for user-friendly natural language office information applications », in *ACM Transactions on Office Information Systems* 2:1, mars 1984, pp. 26-41.

[Kipp, 2008] Michael KIPP (2008). « Spatiotemporal Coding in ANVIL », LREC 2008, 28-30 mai 2008, Marrakech (Maroc).

[Kipp, 2010] Michael KIPP (2010). « Multimedia Annotation, Querying and Analysis in ANVIL », in M. Maybury (éd.), *Multimedia Information Extraction*, chapitre 19, MIT Press.

[Kobus *et al.*, 2008] Catherine KOBUS, François YVON et Géraldine DAMNATI (2008). « Transcrire les SMS comme on reconnaît la parole », TALN 2008, 9-13 juin 2008, Avignon (France).

[Kurdi, 2003] Mohamed-Zakaria KURDI (2003). « Contribution à l'analyse du langage oral spontané », thèse de doctorat, Grenoble (France).

[Lacheret, 2007] Anne LACHERET (2007). « Prosodie-discours : une interface à multiples facettes », in *Nouveaux cahiers de linguistique française* n°28, pp. 7-40.

[Lamel *et al.*, 1991] Lori F. LAMEL, Jean-Luc GAUVAIN et Maxine ESKÉNAZI (1991). « BREF, a Large Vocabulary Spoken Corpus for French », EUROSPEECH 91, 24-26 septembre 1991, Genève (Suisse).

[Lee et Strassel, 2007] Haejoong LEE et Stéphanie STRASSEL (2007). « Using Xtrans for Broadcast Transcription – A User Manual version 3.0 », document téléchargeable à l'adresse :
http://www.ldc.upenn.edu/tools/XTrans/docs/XTransManualV3.pdf

[Luzzati, 1982] Daniel LUZZATI (1982). « "Ben", appui du discours », in *Le Français Moderne* vol. 50, n°3, pp. 193-207.

[Luzzati, 1985] Daniel LUZZATI (1985). « Analyse périodique du discours », in *Le Français Moderne* vol. 65, n°1, pp. 62-73.

[Luzzati, 2004] Daniel LUZZATI (2004). « Le fenêtrage syntaxique : une méthode d'analyse et d'évaluation de l'oral spontané », MIDL 2004, 29-30 novembre 2004, Paris.

[Luzzati, 2007] Daniel LUZZATI (2007). « Le dialogue oral spontané : quels objets pour quels corpora ? », RIHM, vol. 2, pp. 23-27.

[Luzzati, 2009] Daniel LUZZATI (2009). « Corpus d'hier et d'aujourd'hui : progrès quantitatifs et progrès qualitatifs », in *Cahiers de linguistique* vol. 32, n°2, pp. 97-112.

[Luzzati, 2010] Daniel LUZZATI (2010). *Le Français et son orthographe*, Didier, Paris.

[MacCowan *et al.*, 2004] Iain MACCOWAN, Darren MOORE, John DINES, Daniel GATICA-PEREZ, Mike FLYNN, Pierre WELLNER et Hervé BOULARD (2004). « On the Use of Information Retrieval Measures for Speech Recognition Evaluation », Rapport technique, IDIAP-RR-04-73, LIDIAP, Martigny (Suisse).

[MacWhinney, 2000] Brian MACWHINNEY (2000). « The CHILDES Project: Tools for Analyzing Talk », 3ème édition, Mahwah, NJ: Lawrence Erlbaum Associates.

[Marlet, 2008] Renaud MARLET (2008). « Annotation des disfluences », Journées Rhapsodie, 1er juillet 2008, Aix-en-Provence (France).

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

[Martin, 2004] Philippe MARTIN (2004). « WinPitchPro – A Tool for Text to Speech Alignment and Prosodic Analysis », Speech Prosody 2004, 23-26 mars 2004, Nara (Japon).

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

[Mauclair, 2006] Julie MAUCLAIR (2006). *Mesures de confiance en traitement automatique de la parole et applications*, Thèse de doctorat, Université du Maine (Le Mans).

[Mertens, 2002] Piet MERTENS (2002). « Les corpus de français parlé Elicop : consultation et exploitation », in J. Binon, P. Desmet, J. Elen, P. Mertens et L. Sercu (éds.), *Tableaux Vivants. Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock*, Presses Universitaires de Louvain (Belgique).

[Meurs *et al.*, 2008] Marie-Jean MEURS, Frédéric DUVERT, Frédéric BÉCHET, Fabrice LEFÈVRE et Renato DE MORI (2008). « Annotation en frames sémantiques du corpus de dialogue MEDIA », TALN 2008, 9-13 juin 2008, Avignon (France).

[Moeschler et Auchlin, 1999] Jacques MOESCHLER et Antoine AUCHLIN (1999). *Introduction à la linguistique contemporaine*, deuxième édition, Paris : Armand Colin.

[Moeschler *et al.*, 1994] Jacques MOESCHLER, Anne REBOUL, Jean-Marc LUSCHER et Jacques JAYEZ (1994). *Langage et pertinence*, Presses Universitaires de Nancy.

[Mondada, 2005] Lorenza MONDADA (2005). « Constitution de corpus de parole-en-interaction et respect de la vie privée des enquêtés : une démarche réflexive », rapport du projet *Pour une archive des langues parlées en interaction. Statuts juridiques, formats et standards, représentativité*.

[Morel et Danon-Boileau, 1998] Mary-Annick MOREL et Laurent DANON-BOILEAU (1998). *Grammaire de l'intonation, l'exemple du français*, Paris : Ophrys.

[Nakamura *et al.*, 2008] Masanobu NAKAMURA, Koji IWANO et Sadaoki FURUI (2008). « Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance », in *Computer Speech and Language* vol. 22, pp. 171-184.

[Nasr et Béchet, 2009] Alexis NASR et Frédéric BÉCHET (2009). « Analyse syntaxique en dépendances de l'oral spontané », TALN 2009, 24-26 juin 2009, Senlis (France).

[Nodier et Jeandillou, 2008] Charles NODIER et Jean-François JEANDILLOU (2008). *Dictionnaire raisonné des onomatopées françaises*, Genève, Droz.

[Ogden, 1932] Charles Kay OGDEN (1932). *Basic English: A General Introduction with Rules and Grammar,* Kegan Paul, Londres.

[Oostdijk *et al.*, 2002] Nelleke OOSTDIJK, Wim GOEDERTIER, Frank VAN EYNDE, Louis BOVES, Jean-Pierre MARTENS, Michael MOORTGAT et Harald BAAYEN (2002). « Experiences from the Spoken Dutch Corpus Project », LREC 2002, 29-31 mai 2002, Las Palmas (Espagne).

[Pallaud, 2004] Berthille PALLAUD (2004). « 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, 2006] Berthille PALLAUD (2006). « Troncations de mots, reprises et interruption syntaxique en français parlé spontané », JADT 2006, 19-21 avril 2006, Besançon (France).

[Petitrenaud *et al.*, 2010] Simon PETITRENAUD, Vincent JOUSSE, Sylvain MEIGNIER, Yannick ESTÈVE (2010), « Reconnaissance Automatique de Locuteurs à l'aide de Fonctions de Croyance », Actes de RFIA'10, 20-22 janvier 2010, Caen (France).

[Piu et Bove, 2007] Marie PIU et Rémi BOVE (2007). « Annotation des disfluences dans les corpus oraux », RÉCITAL 2007, 5-8 juin 2007, Toulouse (France).

[Polzin et Waibel, 1998] Thomas S. POLZIN et Alexander WAIBEL (1998). « Pronunciation variations in emotional speech », ESCA Workshop Modeling Pronunciation Variation for Automatic Speech Recognition, 3-7 mai 1998, Kerkrade (Pays-Bas).

[Raymond *et al.*, 2002] William D. RAYMOND, Mark PITT, Keith JOHNSON, Elizabeth HUME, Matthew MAKASHAY, Robin DAUTRICOURT et Craig HILTS (2002). « An analysis of transcription consistency in spontaneous speech from the buckeye corpus », ICSLP 2002, 16-20 septembre 2002, Denver (États-Unis).

[Raymond *et al.*, 2007] Christian RAYMOND, Giuseppe RICCARDI, Kepa Joseba RODRIGUEZ et Joanna WISNIEWSKA (2007). « The LUNA Corpus: an Annotation Scheme for a Multi-domain Multi-lingual Dialogue Corpus », Workshop DECALOG'07, Rovereto (Italie).

[Riegel *et al.*, 2009] Martin RIEGEL, Jean-Christophe PELLAT et René RIOUL (2009). *Grammaire méthodique du français*, 7ème édition revue et augmentée, Presses Universitaires de France, Paris.

[Rivenc, 2005] Paul RIVENC (2005). « Espoirs et limites de l'étude quantitative des corpus (dialogues et vocabulaires dits « disponibles ») », colloque *Français fondamental, corpus oraux. 50 ans de travail et d'enjeux*, 8-10 décembre 2005, Lyon (France).

[Roulet *et al.*, 1985] Eddy ROULET, Antoine AUCHLIN, Jacques MOESCHLER, Christian RUBATTEL et Marianne SCHELLING (1985). *L'articulation du discours en français contemporain*, Berne ; Francfort-s.-Main ; New-York : P. Lang.

[Sauvageot, 1962] Aurélien SAUVAGEOT (1962). *Français écrit, français parlé. Les procédés expressifs du français contemporain*, Larousse, Paris.

[Schiffrin, 2001] Deborah SCHIFFRIN (2001). « Discourse markers : Language, Meaning, and Context », in *The Handbook of Discourse Analysis*, Blackwell Publisher, pp. 54-75.

[Schmidt, 2009] Thomas SCHMIDT (2009). « Creating and Working with Spoken Language Corpora in EXMARaLDA », in Lyding, V. (éd.) : *LULCL II: Lesser Used Languages & Computer Linguistics II*, pp. 151-164.

[Schmidt, 2010] Thomas SCHMIDT (2010). « EXMARaLDA : un système pour la constitution et l'exploitation de corpus oraux », Actes du colloque international *Pour une épistémologie de la sociolinguistique*, 10-12 décembre 2009, Montpellier (France).

[Schuurman *et al.*, 2004] Ineke SCHUURMAN, Wim GOEDERTIER, Heleen HOEKSTRA, Nelleke OOSTDIJK, Richard PIEPENBROCK et Machteld SCHOUPPE (2004). « Linguistic annotation of the Spoken Dutch Corpus : If we had to do it all over again... », LREC 2004, 26-28 mai 2004, Lisbonne (Portugal).

[Searle, 1972] John Rogers SEARLE (1972). *Les actes de langage. Essai de philosophie du langage*, Hermann.

[Shriberg, 1994] Elizabeth SHRIBERG (1994). *Detecting disfluency in spontaneous speech*, Thèse de doctorat, Université de Berkeley (États-Unis).

[Shriberg, 2005] Elizabeth SHRIBERG (2005). « Spontaneous Speech : How People Really Talk and Why Engineers Should Care », Interspeech 2005, 4-8 septembre 2005, Lisbonne (Portugal).

[Silver et Lewins, 2009] Christina SILVER et Ann LEWINS (2009). « Transana 2.40 – Distinguishing features and functions », *QUIC Working Papers* n°6, University of Guildford.

[Sirdar-Iksandar, 1980] Christine SIRDAR-IKSANDAR (1980), « *Eh bien !* Le russe lui a donné cent francs », in O. Ducrot (éd.), *Les mots du discours*, Paris : Éditions de Minuit.

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

[Véronis, 2000] Jean VÉRONIS (2000). « Annotation automatique de corpus : panorama et état de la technique », in J.-M. Pierrel (éd.)*, Ingénierie des langues*, pp. 111-129, Hermès, Paris.

[Véronis et Guimier de Neef, 2006] Jean VÉRONIS et Émilie GUIMIER DE NEEF (2006). « Le traitement des nouvelles formes de communication écrite », in Sabah, G. (éd.), *Compréhension automatique des langues et interaction*, Hermès Science, Paris.

Résumé

Le projet EPAC, visant à proposer des méthodes d'extraction d'information et de structuration de documents audio, est le point de départ de nos travaux de recherche. Il nous a permis d'envisager un traitement de la parole conversationnelle en contexte, puisque cette thèse s'est greffée sur un corpus conçu parallèlement à son déroulement. C'est notamment ce que nous exposons dans notre premier chapitre, en relatant notre parcours de recherche, toujours étroitement lié aux tâches de transcription et d'annotation du corpus EPAC.

Dans un deuxième temps, nous nous intéressons plus spécifiquement à la tâche de transcription de la parole. Nous présentons ici quelques « jalons historiques », notamment avec les travaux de Gougenheim ou de Damourette et Pichon. Ensuite, devant la masse émergente de corpus « parole », en France comme à l'étranger, il nous a semblé essentiel de faire le point sur les données disponibles aujourd'hui. Cet état des lieux détaillé regroupe et synthétise un nombre important d'informations, en privilégiant l'accessibilité des données mentionnées. Enfin, nous comparons deux méthodes de transcription, l'une manuelle et l'autre semi-assistée avec l'aide d'un SRAP. Cette dernière offre des résultats plus probants, malgré des taux d'erreur mot parfois élevés. En cela, quelques pistes sont proposées pour tenter d'améliorer ces performances.

Par la suite, nous réalisons une étude comparative de huit logiciels d'aide à la transcription. Cela afin de démontrer que, suivant les besoins du transcripteur et la granularité recherchée, certains sont plus indiqués que d'autres. Des informations objectives telles que les formats et systèmes d'exploitation supportés, les modalités d'acquisition ou la langue de l'interface sont également détaillées pour chaque outil.

Le codage des données est ensuite au centre de notre quatrième chapitre. Nous l'avons vu, il existe de nombreux logiciels permettant de transcrire de l'oral, mais peut-on facilement échanger les transcriptions de l'un à l'autre ? Nous démontrons que l'interopérabilité est un domaine où beaucoup de travail reste à faire, même si des initiatives comme la TEI sont très prometteuses en matière d'uniformisation des données.

Enfin, nous terminons par une analyse détaillée de ce que nous appelons au cours de cette étude la parole *spontanée*. Par différents angles, définitions et expériences, nous tentons de circonscrire ce que cette appellation recouvre. Nous nous intéressons tout particulièrement à ses aspects morpho-syntaxiques et lexicaux, et plus précisément aux *disfluences*, pierre angulaire de notre approche.

Mots-clés : parole spontanée, transcription, corpus, système de reconnaissance automatique de la parole, taux d'erreur mot.