Un grand corpus oral " disponible " : le corpus d'Orléans 1 1968-2012

Iris Eshkol-Taravella, Olivier Baude, Denis Maurel, Linda Hriba, Céline Dugua, Isabelle Tellier

Un grand corpus oral « disponible » : le corpus d'Orléans[1] 1968-2012

Iris Eshkol-Taravella* — Olivier Baude* — Denis Maurel — Linda Hriba* — Céline Dugua*— Isabelle Tellier*****

** Université d'Orléans - Laboratoire Ligérien de Linguistique - UMR 7270*
{Iris.Eshkol, Olivier.Baude, Linda.Hriba, Celine.Dugua}@univ-orleans.fr
***Université François Rabelais Tours – Laboratoire d'Informatique*
{Denis.Maurel@univ-tours.fr}
****Université Paris 3 - Sorbonne Nouvelle, Lattice[2]*
{isabelle.tellier@univ-paris3.fr}

RÉSUMÉ. *Cet article présente la constitution et la mise à disposition du corpus oral ESLO. Notre objectif est de montrer qu'il ne s'agit pas seulement de recueillir et rendre disponible des données langagières mais aussi de rendre explicite l'ensemble de la chaîne de traitement qui permet d'élaborer un tel corpus. Après avoir présenté le projet et le corpus nous préciserons les problèmes juridiques et méthodologiques qui ont conditionné les opérations de traitement du corpus et notamment les procédures d'anonymisation indispensables à la libre diffusion de cette ressource. Dans une seconde partie, nous présenterons les différentes annotations effectuées sur les données brutes avec quelques exemples de leurs exploitations. Nous expliquerons la méthodologie suivie qui est toujours guidée par la nature des données et l'objectif final visé : constituer un grand corpus oral variationniste du français. Nous aborderons enfin les questions de mise à disposition du corpus en ligne.*

ABSTRACT. *This article presents the building and putting online the oral corpus ESLO. Our purpose is to show that it is important not only to collect and make available language data and metadata but also to make explicit the whole chain of treatments. In the first part, we will present the project and the corpus, then we will specify the legal and methodological problems which determined all corpus treatments, in particular the anonymisation procedures which are required to freely make available this kind of resource. In the second part, we present different annotations made on the raw data with some examples of their use. We will explain the followed methodology which is always guided by the nature of the data and by the final objective: build a large sociolinguistic variationist oral corpus of French. Finally, we will discuss the issues of putting the corpus online.*

MOTS-CLÉS : *corpus oral, corpus variationniste, mise à disposition, anonymisation, transcription, annotation, variations.*

KEYWORDS: *oral corpus, variationniste corpus, anonymisation, transcription, corpus annotation, variations.*

1. http://eslo.in2p3.fr/
2. *Ce travail a été réalisé à l'université d'Orléans (au LIFO).*

1. Introduction

L'apparition d'Internet et le développement des outils informatiques ont permis la mise à disposition et la consultation des différents corpus. Des corpus nationaux représentatifs de leur langue comme le BNC[3], le *Russian National Corpus*[4] ou encore le *National Corpus of Polish*[5] apparaissent sur la Toile. Un corpus national représente la langue dans son développement en essayant de tenir compte de toutes les variétés de genres, de styles, d'utilisation, de variantes territoriales, sociales, etc. Ainsi, tous ces corpus contiennent les textes écrits mais aussi des transcriptions d'échanges oraux. Il s'agit souvent de la langue parlée par des locuteurs variés et dans les situations différentes : des conversations spontanées, des entretiens ou des émissions radio. D'autres corpus consacrés exclusivement à la langue parlée sont mis en ligne. Le *Santa Barbara Corpus of Spoken American English*[6] contient des enregistrements d'interactions orales des locuteurs de différentes origines régionales, ethniques, sociales. La majorité du corpus contient des entretiens en face-à-face, mais le corpus comprend aussi des conversations téléphoniques, des parties de jeux de cartes, des conférences, des narrations, des assemblées publiques, etc. Un autre exemple est le projet CORPAFROAS[7]. Il s'agit du premier corpus oral de langues afro-asiatiques (chamito-sémitiques) dont l'objectif consiste en une « *typological comparability among languages: prosodic analysis, and morphosyntactic glossing* » (Mettouchi, A. et Chanard, 2010, p. 258). Nous finirons par mentionner le livre-DVD (Cresti *et al.,* 2005) qui regroupe des corpus comparables en langues romanes (français, italien, portugais et espagnol) de discours spontanés avec des exemples de quelques études comparatives et avec l'accès simultané au son et à la transcription.

La situation semble être différente en France. Force est de constater que l'oral a été longtemps marginalisé dans le champ de la linguistique française (Blanche-Benveniste et Jeanjean, 1983) comme dans celui de la linguistique de corpus. Faisant l'inventaire des corpus oraux en français, Cappeau et Gadet (2007) notent qu' « il n'y a pas eu en France de volonté institutionnelle qui aurait conduit à la constitution d'un grand corpus oral. C'est, en contraste, ce qui a été fait pour l'écrit ». Cependant les travaux sur « le français parlé » puis l'apport des nouvelles technologies ont permis un engouement récent pour ce domaine. Parmi les initiatives actuelles, nous pouvons citer la base CLAPI[8] constituée pour étudier les interactions orales, le corpus PFC[9] pour analyser certains phénomènes

3. British National Corpus, http://www.natcorp.ox.ac.uk/
4. http://www.ruscorpora.ru/en/index.html
5. http://nkjp.pl/index.php?page=0&lang=1
6. http://www.linguistics.ucsb.edu/research/sbcorpus_obtaining.html
7. http://corpafroas.tge-adonis.fr/index_fr.html
8. Corpus de langues parlées en interaction, http://clapi.univ-lyon2.fr/
9. Phonologie du français contemporain, http://www.projet-pfc.net/?accueil:intro

phonologiques, le corpus CRFP[10] pour la morphosyntaxe ou le cas du corpus de français spontané EPAC[11] composé des interviews et des débats d'émissions de télé.

Des initiatives institutionnelles (Centre de ressources numériques du CNRS, ANR Corpus, Programme corpus de la parole de la DGLFLF en partenariat avec les fédérations de recherche en linguistique du CNRS, la création du TGE-ADONIS et du TGIR CORPUS) n'ont pas encore permis une mise à disposition d'envergure des corpus oraux. En 2011, la création du consortium Corpus oraux et multimodaux au sein de l'IRCORPUS qui doit répondre à l'objectif de « fédérer les équipes, laboratoires, chercheurs et enseignants-chercheurs engagés dans la constitution de corpus oraux et multimodaux, afin de faire converger les pratiques et de les rendre conformes aux standards internationaux[12] » confirme l'intérêt important pour ce champ d'études.

À la différence de l'écrit qui n'utilise qu'un seul support, l'oral associe le plus souvent la parole enregistrée à une représentation écrite et/ou codée (transcriptions, traductions, annotations). Cette donnée « secondaire » permet son exploitation par les outils de la linguistique de corpus. Toutefois les spécificités de l'oral nécessitent une adaptation des outils. La superposition des voix, ou « chevauchement de parole », ainsi que les phénomènes de disfluences – hésitations, répétitions, reprises, fausses amorces – n'existent pas à l'écrit et rendent le traitement de l'oral compliqué.

Cet article présente un exemple de constitution et de mise à disposition d'un grand corpus oral de français. Il s'agit de porter un regard sur le travail effectué au sein du programme de recherche ESLO (*Enquête Sociolinguistique à Orléans*) et par-delà, sur les méthodes actuelles d'exploitation des corpus et des données sociolinguistiques[13].

ESLO est un corpus de référence du français parlé hier et aujourd'hui à Orléans. Sa caractéristique première est de permettre des analyses sur la variation dans le français. La seconde caractéristique réside dans la structure même de ce corpus, composé de deux enquêtes : ESLO1 en 1968-1971 (un corpus clos/stable) et ESLO2 en 2008-2012 (un corpus ouvert/évolutif).

Cette expérience est donc l'occasion d'aborder quelques problèmes liés aux opérations de conservation et de diffusion d'un tel corpus. Comment tenir compte de l'hétérogénéité des données ? Quelles sont les données à constituer et à traiter ? Comment les coder ? Quels outils utiliser ? Toutes ces questions sont indissociables

10. Corpus de référence du français parlé, http://www.up.univ-mrs.fr/delic/crfp
11. Exploration de masse de documents audio pour l'extraction et le traitement de la parole conversationnelle, http://projet-epac.univ-lemans.fr/doku.php?id=accueil.
12. http://www.corpus-ir.fr/index.php?page=ircom
13. Ce travail a été réalisé grâce au soutien de l'ANR (projet Variling) et du Feder Région Centre (projet Entités).

et doivent être posées : de la collecte jusqu'à la diffusion. Mettre à disposition ce corpus implique de maîtriser l'ensemble de la chaîne de traitement tout en veillant à une cohérence au sein des contraintes imposées par la nature des données et par la nécessité de conserver les moyens d'une comparaison entre les deux corpus.

Nous développerons plus précisément les travaux d'annotation et de structuration du corpus rendus nécessaires par l'anticipation des problèmes juridiques de diffusion d'un corpus oral.

2. Corpus

2.1. *Historique*

ESLO1, la première enquête sociolinguistique à Orléans, a été réalisée en 1968-1971 par des professeurs de français de l'University of Essex, Language Centre, Colchester (Royaume-Uni), en collaboration avec des membres du B.E.L.C. (Bureau pour l'étude de l'enseignement de la langue et de la civilisation françaises de Paris). L'objectif en était double. D'une part rendre disponible l'ensemble du corpus – « Des listes de transcriptions et enregistrements sont disponibles à ceux qui s'adressent à nous. » (Lonergan *et al.*, 1974) – et d'autre part, constituer un corpus « sociolinguistique » autour du concept de « portrait sonore d'une ville » afin de croiser représentativité et variations au sein d'une communauté d'auditeurs dans un espace géographique et socioéconomique clairement défini (Bergounioux *et al.*, 1992). Dans les années 1980-1990, une partie du corpus a été transcrite et étiquetée puis mise à disposition sur la Toile dans le cadre du projet ELILAP/LANCOM[14]. Entre 1993 et 2001, une partie du corpus a été reprise par des chercheurs de l'Université de Louvain (Debrock *et al.*, 2000) dans le cadre du projet ELICOP[15].

Quarante ans après le projet initial, le Laboratoire Ligérien de Linguistique (LLL) de l'université d'Orléans a entrepris un double projet : diffuser largement le corpus ESLO1 et réaliser un nouveau corpus représentatif du français parlé à Orléans dans les années 2010 (ESLO2), en prenant en compte l'expérience d'ESLO1 et l'évolution des cadres théoriques et méthodologiques de la constitution et de l'exploitation de grands corpus oraux à visée variationniste.

14. ELILAP 1980-83 puis LANCOM 1993-2001, voir Mertens (2002).
15. http://bach.arts.kuleuven.be/elicop/

2.2. *ESLO en chiffres*

Actuellement, ESLO1 est composé de 470 enregistrements, d'une durée totale de 317 heures et évalué à 4 500 000 mots. Plusieurs genres y sont représentés : la grande partie des enregistrements consiste en des entretiens en face-à-face (157 enregistrements comportant autant de profils sociologiques différents), mais ESLO1 compte également des enregistrements libres dans des situations privées ou professionnelles faites en l'absence des chercheurs (16 enregistrements), des interviews des personnalités de la ville (45 enregistrements), des reprises de contact complètement informelles (55 enregistrements), des communications téléphoniques (50 enregistrements), des conférences-débats ou même des discussions (26 enregistrements), des interviews au centre médico-psychopédagogique entre des parents d'élèves et des assistantes sociales (37 enregistrements), ainsi que des enregistrements divers dans les lieux publiques : magasins, marchés, visites d'ateliers, etc. (84 enregistrements).

ESLO2, débuté en 2008, est un corpus en cours. À terme, il comprendra plus de 350 heures d'enregistrement afin de former avec ESLO1 un corpus de plus de 700 heures et atteignant les dix millions de mots. Il s'agira alors d'un grand corpus oral réalisé selon des bonnes pratiques de constitution garantissant l'interopérabilité des données avec d'autres projets semblables.

La mise à disposition d'un tel corpus se heurte en premier lieu aux aspects juridiques. Nous verrons que l'anticipation de ceux-ci a eu une incidence sur l'ensemble de la chaîne de traitement.

3. Mettre à disposition : les aspects juridiques

À l'époque de la conservation pérenne et de la diffusion des archives numériques il est important de prendre en compte l'ensemble des aspects juridiques dès la conception du projet. Nous avons bénéficié des réflexions et recommandations émanant du groupe de travail du ministère de la Culture et du CNRS[16] : *Corpus oraux. Guide des bonnes pratiques 2006*.

Les problèmes rencontrés se concentrent sur deux grands domaines juridiques : le respect de la vie privée et la protection de la propriété intellectuelle.

16. Groupe de travail composé de linguistes, juristes, informaticiens et conservateurs.

3.1. *Respect de la vie privée*

3.1.1. *Complexité de la notion*

Selon le *Guide des bonnes pratiques 2006* pour gérer les droits liés au respect de la vie privée, il convient de suivre scrupuleusement le cadre légal de gestion des données personnelles, de s'assurer que les locuteurs ont exprimé leur consentement « éclairé » ou, à défaut de celui-ci, de procéder à l'anonymisation des données. Juridiquement, l'anonymisation consiste à rendre impossible l'identification d'une personne. Cette notion est assez complexe à mettre en place dans le cadre de recherches linguistiques sur l'oral (il n'est pas question de « brouiller » totalement la voix des locuteurs). On restreindra l'objectif au masquage ou à la suppression des éléments permettant une identification par un large public qui utiliserait des moyens classiques de requêtes. Il faut donc repérer et traiter les indices permettant d'identifier directement ou indirectement la personne, ainsi que les éléments qui peuvent lui porter préjudice. Il peut s'agir des formes nominatives, des professions, statuts, ou titres, des activités sociales, des liens de parenté, des réseaux, des références à des lieux et/ou des références à des caractéristiques de la personne ou encore des propos légalement répréhensibles.

Les données traitées concernent les fichiers audio, les données primaires textuelles ainsi que les données secondaires comme les transcriptions, les métadonnées ou les analyses effectuées sur ces données.

La diffusion des données anonymisées présuppose aussi la préservation et la conservation des données originales non anonymisées ainsi que l'accès restreint à ces données.

3.1.2. *Respect de la vie privée dans ESLO : anonymisation*

Dans le cas du corpus ESLO1, le recueil de consentement pose deux problèmes. Premièrement, il n'existe aucun document rempli par les locuteurs qui permettrait d'exprimer ce consentement ; deuxièmement, il serait illusoire de penser que les locuteurs de la fin des années soixante imaginaient le type d'exploitation et notamment la diffusion instantanée par Internet. Toutefois le contenu des enregistrements et certains documents annexes prouvent que les locuteurs étaient conscients d'être enregistrés à des fins d'exploitation scientifique et didactique. Pour ESLO2, le recueil d'un consentement éclairé est systématique et permet une diffusion de l'ensemble des données brutes.

Le choix de l'équipe a néanmoins été d'anonymiser l'ensemble des données d'ESLO1 et d'ESLO2. L'idée est de repérer dans les enregistrements des éléments sensibles pouvant donner une information personnelle sur le locuteur.

Nous nous sommes tout d'abord intéressés aux noms propres. Il faut considérer que tous les noms propres ne sont pas à anonymiser : la *Loire* et *Jeanne d'Arc* ne sont pas à inclure dans l'effacement, ainsi que les toponymes se trouvant dans la

réponse à la question « *Où parle-t-on bien le français ?* » ou encore le nom des animateurs célèbres de l'époque, dans les réponses sur les questions concernant les émissions télévisées ou radiophoniques. En revanche, dans la phrase « *Je travaille au collège de Saint-Jean-de-Braye* », l'entité *collège de Saint-Jean-de-Braye* n'est plus seulement un établissement scolaire, mais également un lieu de travail du locuteur. Les noms communs désignant les métiers, par exemple, peuvent aussi à leur tour donner une information personnelle. En observant le corpus, nous avons constaté que c'est souvent le regroupement de plusieurs indices qui peut renvoyer vers l'identité du locuteur. Être un *professeur* ne permet pas d'identification, mais il n'en va pas de même s'il est précisé par ailleurs que c'est un *professeur d'université spécialisé en électronique* et, ailleurs encore, que c'est une *femme*, auquel cas on peut arriver à un singleton. À cela s'ajoutent les exemples comme « *mon père a fondé le plus grand cabinet d'ophtalmologiste de la ville* » qui sont rarement présents dans le corpus et permettent, en revanche, l'identification directe du locuteur.

Si au début, nous avons pensé automatiser complètement le processus d'anonymisation en nous fondant sur des couches d'annotations automatiques de ce type d'information, nous avons vite été confrontés à l'impossibilité d'effectuer cette tâche, d'où le travail manuel de la validation selon le contexte des entités repérées automatiquement[17]. Celles qui identifient le plus le locuteur ont été remplacées par leur hyperonyme et le son a été masqué. Pour cette raison, il a été décidé de simplifier le processus d'anonymisation pour ESLO2 : la phase de remplacement par un hyperonyme s'y fait dès la transcription.

3.2. *Propriété intellectuelle*

À la différence du corpus constitué en 1968-1971, en 2012, la question de la propriété intellectuelle de corpus contenant des paroles de locuteurs enregistrés, des enrichissements de ces paroles (transcriptions, annotations), constitués en base de données dans le cadre de projets financés par les institutions de l'État et réalisés par des enseignants-chercheurs en fonction est éminemment complexe.

L'objectif annoncé du projet ESLO de mise à disposition de l'ensemble du corpus a permis à l'équipe de se positionner dans une démarche de ressources libres et accessibles. Ainsi ESLO est un corpus mis à disposition librement sous licence *Creative Commons*[18]. L'usage de ce type de licences permet de gérer les droits d'exploitation en spécifiant la paternité et les conditions d'utilisation. Nous verrons

17. Dans la section 4.4., nous présenterons une couche d'annotations automatiques dont nous nous sommes servis pour l'anonymisation des entretiens d'ESLO1.

18. Les licences Creative Commons sont des contrats-types pour la mise à disposition d'œuvres en ligne. Il s'agit d'autorisations non exclusives données par les titulaires des droits au public. Ces autorisations spécifient les conditions d'utilisation des œuvres.

également l'impact de ce choix sur la structuration de la base de données et sur les opérations de traitement du corpus.

Les sections suivantes sont consacrées à la description des opérations de traitement du corpus entrant en jeu dans la préparation de ressources destinées à une libre diffusion. Comme nous le verrons il ne s'agit pas simplement de problèmes techniques qui pourraient apparaître neutres sur le plan théorique.

4. Annotation

Pour exploiter un corpus oral il est nécessaire de le transcrire et certaines tâches d'annotation deviennent utiles et/ou indispensables. « Mais c'est certainement une erreur que d'imaginer que le modèle suivi pour l'écrit pourrait être transféré à l'oral. En effet, les corpus oraux sont liés à des exploitations extrêmement diversifiées (analyse prosodique, analyse de discours, analyse syntaxique, approches pragmatiques ou sociolinguistiques, etc.) qui nécessitent des informations par nature très disparates. » (Cappeau et Gadet, 2007). Les choix d'annotation diffèrent d'un projet à l'autre suivant des objectifs variés. Ainsi, dans le cadre du projet OTIM[19], le travail d'annotation a porté sur un grand nombre de domaines : phonétique, prosodie, phonologie, syntaxe, discours et gestes. Le corpus EPAC que nous avons mentionné avant a été annoté en prenant en compte divers phénomènes : bruits, musiques, inspirations, prononciations particulières ou erronées, mots étrangers, néologismes... Le projet Rhapsodie[20], quant à lui, met au centre de ses activités les annotations prosodique et syntaxique des données orales existantes.

Les outils d'annotation varient également selon la nature de l'annotation, c'est-à-dire selon les phénomènes que l'on veut distinguer. Ainsi, l'annotation automatique des coréférences, par exemple, pose de nombreux problèmes et nécessite le recours à l'intervention humaine. Toutefois, il existe des outils d'aide à l'annotation manuelle comme Transcriber[21], Praat[22], ANVIL[23], ELAN[24] pour la transcription, et Glozz[25], Gate[26], etc. pour d'autres niveaux d'annotation. L'annotation automatique ou semi-automatique peut se faire avec des méthodes à base de règles linguistiques décrivant le contexte d'emploi de phénomènes à annoter sous forme de grammaires locales ou avec des méthodes d'apprentissage automatique à partir d'un corpus de référence

19. Outils pour le traitement de l'information multimodale, http://www.lpl-aix.fr/~otim.
20. http://rhapsodie.risc.cnrs.fr/fr/index.html
21. http://trans.sourceforge.net/en/presentation.php. Une nouvelle version de ce dernier est disponible depuis juillet 2011.
22. http://www.fon.hum.uva.nl/praat/
23. http://www.anvil-software.de/
24. http://icar.univ-lyon2.fr/projets/corinte/confection/elan.htm
25. http://www.glozz.org/
26. http://gate.ac.uk/

annoté manuellement. C'est la nature des données qui dicte le choix de la méthodologie à adopter. Le corpus ESLO est un exemple de cette démarche. Nous montrerons dans cette partie, comment nous avons adopté les outils et techniques existants à chaque type d'annotation.

D'après Leech (1997), l'annotation est une « valeur ajoutée » aux données brutes, c'est-à-dire un apport d'informations. Toujours selon cet auteur, la transcription possède un statut ambigu car la frontière entre les données brutes, neutres et leur annotation n'est pas clairement délimitée. Tout commentaire (balisage des bruits, notes du transcripteur) appartient également au domaine de l'annotation et peut donc être considéré comme de l'interprétation.

Nous considérons le processus d'annotation comme *porteur d'une interprétation*. Ainsi, il n'y a pas qu'une version de corpus annoté mais plusieurs versions –existantes ou potentielles – du même corpus. Les différents annotateurs humains peuvent interpréter et percevoir différemment les données. Même dans le cas de l'annotation automatique, les annotations diffèrent par les conventions, les techniques, etc. ESLO est un corpus de variations : variations entre le français d'hier et d'aujourd'hui, entre les différents locuteurs, entre les différentes situations d'enregistrement, mais aussi entre les différentes annotations.

Afin de rendre disponible et exploitable notre corpus, nous avons suivi les principes d'annotation suivants pour la transcription :

– lisibilité ;
– conservation des spécificités de l'oral ;
– volonté d'un maximum d'interopérabilité ;
– codage non ambigu ;
– contraintes de comparabilité (d'ESLO1 à ESLO2).

Pour nous, la transcription doit être considérée comme un premier niveau d'annotation : le son étant enrichi d'une information orthographique.

Dans le cadre d'un corpus sociolinguistique, on souhaite également intégrer des descripteurs de la situation d'interaction, notamment les données sociologiques sur les locuteurs et la description de la situation d'enregistrement. Ces données peuvent être décrites dans les métadonnées mais peuvent aussi être contenues directement dans le corpus, comme par exemple dans un entretien ou un récit de vie. L'exploitation sociolinguistique nécessite évidemment la disponibilité de ces informations.

Nous montrerons, dans la partie qui suit, les différentes annotations effectuées sur le corpus afin de répondre à cet objectif.

4.1. *Annotation du niveau zéro : transcription*

4.1.1. *Contraintes et choix de transcription*

La transcription qui est le premier degré d'annotation de l'oral est une étape primordiale dans la constitution du corpus puisque c'est sur ce premier niveau que vont s'ajouter d'autres annotations. Les choix faits à ce stade influencent donc tout le traitement postérieur. La tâche a été d'autant plus difficile qu'il n'y a pas de conventions de transcription admises par la communauté scientifique.

Plusieurs contraintes ont influencé nos choix. Notre volonté première était de mettre à disposition des chercheurs une grande quantité de données transcrites (700 heures d'enregistrement). Le processus de transcription devait donc être effectué rapidement mais avec une bonne efficacité. Il n'existe pas aujourd'hui d'outils de transcription automatique disponible, il s'agit donc de transcriptions manuelles. Ceux qui ont travaillé sur l'annotation manuelle savent que moins on annote d'informations, plus on gagne dans la quantité et la qualité car l'annotateur est moins dispersé et donc plus concentré sur sa tâche. Nous sommes allés dans la même direction et nous avons choisi l'annotation minimale. Il s'agit de la transcription orthographique qui conserve les spécificités de l'oral (amorces, disfluences, répétitions, etc.). Les conventions de transcription ont été réduites ainsi au minimum. Pour éviter l'anticipation de l'interprétation (Blanche-Benveniste et Jeanjean, 1987), les marques typographiques comme le point, la virgule, le point d'exclamation ou encore la majuscule en début d'énoncé sont absentes. La segmentation a été faite soit sur une unité intuitive de type « groupe de souffle » repérée par le transcripteur humain, soit sur un « tour de parole », défini uniquement par le changement de locuteurs.

La synchronisation avec le son était une autre contrainte. On devait pouvoir naviguer dans la transcription et le son en parallèle. L'objectif a été défini de transcrire et rendre disponible l'intégralité du corpus en y associant des jalons temporels.

La mise à disposition des ressources implique l'utilisation de normes et même si les conventions de transcription ne sont pas normalisées, l'exigence sur le format standardisé des fichiers s'impose. Ce format interopérable devait permettre un traitement plus facile du corpus par les outils du TAL.

Notre choix s'est arrêté sur le logiciel de transcription Transcriber qui répondait complètement à nos attentes : outil facile d'utilisation permettant la synchronisation entre la transcription et le signal sonore et ayant un format de sortie XML, un format normé, facilement exportable et largement utilisé, gage d'interopérabilité. À l'aide de ce logiciel, l'ensemble des données dont l'acoustique était acceptable a été transcrit, ce qui nous permet de disposer d'un panorama varié des différents types d'enregistrements ESLO1 (Baude et Dugua, 2011).

Pour nous, la phase de transcription relève et révèle systématiquement des variations linguistiques. Cette conception de la transcription nous a donc guidés dans l'élaboration de notre méthodologie. Chaque enregistrement est transcrit en trois étapes successives, donnant lieu à trois versions différentes :

– transcription (A), première transcription rapide ;
– transcription (B) qui est la transcription (A) relue et corrigée par un deuxième transcripteur ;
– transcription (C), la transcription (B) relue et corrigée par un troisième transcripteur.

Avec cette méthode, « nous évaluons le temps de transcription à 10 fois pour une première version brute, 5 fois pour une deuxième et autant pour une troisième. » (Baude et Dugua, 2011). Cette méthodologie nous a permis de constater et de confirmer notre hypothèse sur la variation dans l'annotation ; ce sont trois perceptions différentes de l'écoute, qui manifestent toutes trois des types de variations et des opérations particulières.

4.1.2. *Transcriptions et variations*

La comparaison[27] des trois versions de transcriptions a montré d'importantes divergences. Trois cent trente différences ont été relevées en moyenne entre les trois versions. Dans le cadre d'un travail de thèse (Hriba, en cours), leur description, leur analyse et leur catégorisation, à partir d'un corpus constitué de vingt enregistrements et de soixante fichiers de transcription (225 173 mots), ont permis la mise en évidence de trois types de variations :

– des variations graphiques qui regroupent l'ensemble des erreurs qui, d'une part, correspondent aux fautes induites par les outils de saisie (clavier) et d'aide à la transcription (Transcriber) et, d'autre part, qui correspondent au non-respect d'une norme orthographique ou de codage (cf. conventions de transcription) ;

  ESLO1_110B/C : les g**ra**çons/les g**ar**çons
– des variations de segmentation qui concernent des différences d'alignement temporel, la segmentation en sections, en tours de parole et les pauses ;
– des variations de perception manifestant des divergences d'écoute ;

  ESLO1_062A/B : **on** le mettait tout à fait **en bas** à gauche / **il** le mettait tout à fait **en haut** à gauche.

L'étude des variations a montré qu'aux trois types de variations correspondaient des opérations communes, d'une part, et des opérations spécifiques à chacune d'elles, d'autre part. Les opérations correspondent aux interventions que les relecteurs font sur la version précédente (de la B sur la A et de la C sur la B). Nous

27. Comparaison réalisée à l'aide d'un outil spécialisé Beyond Compare 3 : http://www.scootersoftware.com/

présentons ici les résultats généraux en nous limitant aux opérations communes, au nombre de trois : des modifications, des suppressions et des rajouts. Le tableau 1 fait apparaître la proportion de chaque type d'opération pour les deux relectures.

Types d'interventions	**Interventions de la version B sur la version A**	**Interventions de la version C sur la version B**
Modifications	56 %	49 %
Suppressions	11 %	9 %
Ajouts	33 %	42 %

Tableau 1. *Interventions des relecteurs*

Globalement, nous constatons que les types d'opérations présentent une répartition semblable que ce soit pour la relecture B ou pour la validation C.

Parmi les trois types de variations, les variations de perception sont celles qui ont fait l'objet d'une étude particulière. En effet, à travers l'analyse de 130 exemples et des interférences qui ont conduit aux divergences constatées, Hriba a pu relever des mécanismes qui démontrent que la variation est inhérente au système, y compris dans un traitement de description du signal sonore.

ESLO est donc un corpus qui en imbrique d'autres. Nous le considérons plutôt comme une « archive » qui permet l'extraction de « corpus d'études ». Ainsi, les trois versions des transcriptions constituent à leur tour un corpus à part entière qui demande à être analysé. Le choix de garder les trois versions correspond à notre conception d'un corpus dont les traitements sont révélateurs et même porteurs de variations.

4.2. *Annotations et métadonnées*

À toutes les étapes et donc dès la collecte des données, se pose la question des éléments descripteurs de la ressource afin de faciliter son exploitation, sa réutilisation et son archivage. Il est important de prévoir l'annotation des métadonnées qui décrivent et enrichissent le corpus dans la chaîne de traitement dès le début. Cette tâche devient encore plus compliquée pour le corpus oral qui met en jeu les différents types de fichiers (sonores et écrits), les différentes situations d'enregistrement : entretiens en face-à-face, conversations spontanées, réunions de travail, etc., les locuteurs qui se distinguent par leurs âge, sexe, profession, lieu de naissance, milieu social, etc. Les contributeurs du corpus varient également : ceux

qui font les enregistrements ne sont pas toujours ceux qui les transcrivent. Il s'agit donc de décrire d'une manière homogène toute cette variété de données.

Dans le cadre du projet ESLO, nous pouvons distinguer les métadonnées décrivant les enregistrements et les transcriptions et les métadonnées enrichissant la situation de production linguistique et notamment le profil sociologique du locuteur.

Pour les fichiers de transcription, une partie des métadonnées est contenue dans les balises XML proposées par Trancriber au cours de l'étape de transcription. Cette information permet de décrire le fichier de transcription.

Chaque fichier de transcription correspondant à un enregistrement est caractérisé, en premier lieu, par le nom du transcripteur, le numéro de l'enregistrement, la version ainsi que la date de la transcription :

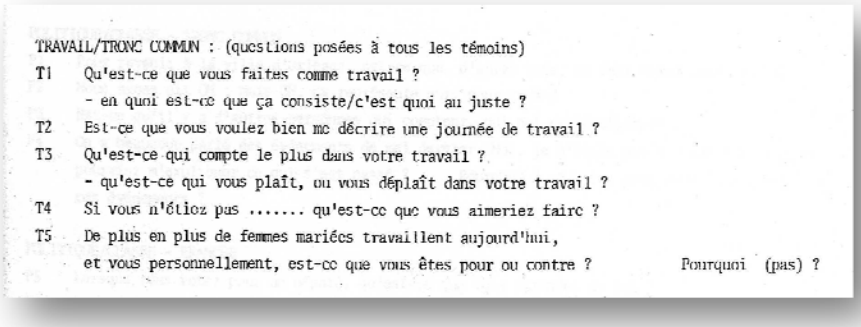
```
<Trans scribe="Panot" audio_filename="001" version="15" version_date="091210">
```

L'attribut *version* marque le nombre d'interventions au cours de la transcription. Cette information peut s'avérer très intéressante si l'on veut analyser la durée de transcription manuelle d'un enregistrement et le nombre d'interruptions que le transcripteur humain fait pendant ce processus.

Dans le cas des entretiens, on décrit des différentes thématiques de l'interview :

```
<Topics>
<Topic id="to1" desc="QP1"/>
<Topic id="to2" desc="QP2"/>
```

Les codes décrits font référence à une trame d'entretien établie lors de la constitution du plan expérimental. Un document annexe dont voici un extrait du catalogue tapuscrit original de 1974 (figure 1) détaille les thématiques.



TRAVAIL/TRONC COMMUN : (questions posées à tous les témoins)
T1 Qu'est-ce que vous faites comme travail ?
- en quoi est-ce que ça consiste/c'est quoi au juste ?
T2 Est-ce que vous voulez bien me décrire une journée de travail ?
T3 Qu'est-ce qui compte le plus dans votre travail ?
- qu'est-ce qui vous plaît, ou vous déplaît dans votre travail ?
T4 Si vous n'étiez pas ..... qu'est-ce que vous aimeriez faire ?
T5 De plus en plus de femmes mariées travaillent aujourd'hui,
et vous personnellement, est-ce que vous êtes pour ou contre ? Pourquoi (pas) ?

Figure 1. *Extrait des thématiques de l'interview dans ESLO1*

Chaque question est incluse dans une thématique plus vaste traitant du travail, des loisirs, de la politique, de l'éducation, de la culture ou encore de la langue elle-

même. L'extraction de cette information permet de constituer des sous-corpus selon un sujet traité.

L'information sur les participants de l'enregistrement est intégrée dans la balise <Speakers> :

<Speaker id="spk1" name="HM" check="no" dialect="native" accent="" scope="local"/>

Les métadonnées sont décrites aussi dans la base de données connectée à l'interface Web ce qui donne accès à la fois à l'enregistrement (fichier sonore), à sa/ses transcription(s) (fichier de Transcriber) et aux informations relatives au locuteur principal de chaque entretien : date et lieu de naissance, sexe, profession et appartenance sociale. Ces critères ont été renseignés au moment de l'enquête.

Le corpus des ESLO est également déposé au CRDO[28] avec les métadonnées requises par la procédure d'archivage de celui-ci (quinze étiquettes extraites du standard DUBLIN-CORE OLAC). Voici un exemple extrait du corpus ESLO au CRDO et une description des métadonnées utilisant les étiquettes du DUBLIN CORE OLAC[29] :

dc:title. [1, 1]. Titre.
dc:subject. [1, n]. Description du sujet du contenu de la ressource.
dc:type. [0, n]. Nature ou genre du contenu de la ressource.
dc:source. [0, n]. Référence à une ressource à partir de laquelle la ressource actuelle a été dérivée.
dcterms:spatial. [0, n]. Couverture spatiale. En général le point d'enquête, sauf s'il ne représente absolument pas la couverture spatiale visée
dcterms:temporal. [0, n]. Couverture temporelle de la ressource. A ne pas confondre avec les dates d'enregistrement,
dc:creator. [0, n]. Entité responsable de l'élaboration du contenu de la ressource (individu, institution, organisation)
dc:publisher. [0, n]. Entité responsable de la mise à disposition de la ressource, dans sa forme actuelle.
dc:contributor. [0, n]. Entités ayant contribué à la création du contenu de la ressource. Préciser leurs rôles par choix
dc:rights. [0, 1]. Indiquer la mention de copyright
dcterms:license. [1, 1]. URL de la licence creative-commons choisie
dcterms:created. [1, 1]. Date de création de la ressource.
dcterms:modified. [1, 1]. Date de dernière modification de la ressource.
dc:language. [0, n]. Langue du contenu intellectuel de la ressource. Langue de l'enquêteur dans un document sonore.

Ces métadonnées correspondent à des normes internationales qui permettent non seulement de décrire une situation de production mais aussi de cataloguer les ressources électroniques afin d'en faciliter l'accès ultérieurement.

Enfin, dans le cadre du projet ESLO, nous avons développé notre propre jeu complémentaire de métadonnées structuré dans une base de données dédiée à celui-ci.

28. Centre de ressources pour la description de l'oral – CNRS.
29. http://www.language-archives.org/REC/role.html.

4.3. *Annotation du discours*

L'étape suivante n'est plus descriptive mais a comme objectif la segmentation du flux de paroles, qui se manifeste par les changements de locuteurs, les pauses, les événements comme la toux, les rires, etc. Les tours de parole, le temps d'énoncé, les pauses, les différents événements interrompant le flux de paroles, doivent être marqués pour faciliter le traitement du corpus par les outils informatiques et permettre l'alignement avec le fichier sonore. Ce processus d'annotation a été effectué sous Transcriber.

La transcription est découpée d'abord en sections (ou *Report*) qui correspondent dans les entretiens, par exemple, aux questions de la trame du questionnaire. La deuxième segmentation se fait par le transcripteur soit intuitivement selon le groupe de souffle ou s'il y a une pause dans le discours du locuteur, soit selon le tour de parole, défini uniquement par les changements de locuteurs :

```
<Turn speaker="spk1" startTime="0.449" endTime="3.114">
<Sync time="0.449"/>
vous savez euh
<Sync time="1.317"/>
<Sync time="2.814"/>
enfin
</Turn>
```

Les attributs « *startTime* » et « *endTime* » indiquent le temps de début et de fin des segments de parole. Un des phénomènes de l'oral qui nécessite une segmentation particulière est la pause, elle est notée par un segment vide. Cette segmentation permettra d'avoir précisément la durée de la pause.

Les divers événements au sein du discours oral annotés avec Transcriber concernent les différents bruits : rire, micro, passages non transcrits, bruits de respiration (inspiration, soupir, respiration) ou encore des clics ou bruits de bouche :

```
<Event desc="rire" type="noise" extent="instantaneous"/>
```

Nous ajouterons à ces événements les phénomènes de prononciation :

```
petite <Event desc="pi" type="pronounce" extent="instantaneous"/> moyenne
les techniciens ils <Event desc="i" type="pronounce" extent="previous"/> font
```

Les informations annotées sont un autre exemple de variation. La variation dans le même enregistrement peut être observée entre les différentes questions posées, entre la pause et le discours, entre le discours et l'événement. Les différents enregistrements d'interviews mettent en évidence la variation entre les différents locuteurs. Ainsi, dans (Dupont *et al.,* 2012), les auteurs se sont intéressés au sous-corpus composé des réponses à une question posée aux différents locuteurs sur les événements de mai 1968. Certaines informations comme la durée de pause après la question, la durée de pause par section ou la durée de section ainsi que les renseignements sur le locuteur comme son sexe, son âge, son niveau d'études, etc. ont été analysées avec des méthodes de statistiques descriptives. Il s'agissait

d'étudier des variations et des corrélations entre d'une part, des annotations de temps dans les fichiers de transcription, et, d'autre part, des valeurs sociologiques.

4.4. *Annotations des informations personnelles concernant le locuteur*

L'annotation est toujours liée à la nature des données à annoter et à l'objectif visé. Les enquêtes sociolinguistiques à Orléans contiennent beaucoup d'informations personnelles sur les locuteurs interviewés. Repérer ces données est d'autant plus nécessaire qu'elles peuvent être utilisées dans le processus d'anonymisation.

Ce type d'annotation a été réalisé sur une partie du corpus ESLO1[30]. Nous avons sélectionné 112 entretiens en face-à-face. Les entretiens contiennent beaucoup de questions du type : « *Depuis combien de temps habitez-vous Orléans ?* » « *Quel âge avez-vous ?* » « *Qu'est-ce que vous faites comme métier ?* » « *Où travaillez-vous ?* » « *Qu'est-ce que fait votre époux(se) ?* », etc. Les réponses des locuteurs montrent comment les Orléanais parlent d'eux-mêmes et représentent une masse de données mettant en valeur une variation intéressante à analyser. L'intérêt des entretiens pour annoter automatiquement ces informations est qu'ils présentent des données riches et homogènes. Dans les discours spontanés, les énoncés contenant des informations personnelles permettant une identification sont plus rares et surtout moins structurés, ce qui rend plus difficile une annotation automatique. C'est la raison pour laquelle nous nous sommes limités à ce sous-corpus.

4.4.1. *Méthodologie adoptée*

L'annotation s'est faite en deux étapes. Nous avons repéré et annoté, en premier lieu, les entités nommées comme le nom de la personne, son âge ou son lieu de travail. Nous avons recherché ensuite les éléments plus personnels concernant le locuteur comme son métier, le nombre d'enfants qu'il avait, le métier de son conjoint, etc. que nous avons appelés « entités dénommantes » (Eshkol, 2010). Le processus de reconnaissance de ces informations s'est effectué sur le corpus annoté en entités nommées.

L'annotation des entités nommées et dénommantes a été décrite dans (Maurel *et al.,* 2011). Nous nous contenterons dans cet article de présenter une synthèse de ce travail.

Pour repérer et annoter les entités nommées et dénommantes, nous avons choisi l'approche en surface permettant de construire les grammaires locales selon le contexte en utilisant le système CasSys (Friburger, 2002) intégré à la plate-forme Unitex (Paumier, 2003).

---

30. À l'origine de cette étude les transcriptions ESLO2 n'étaient pas disponibles.

Notre choix a été guidé d'une part, par le corpus déjà constitué des entretiens dans lesquels les mêmes questions avaient été posées aux différents locuteurs. L'analyse de ce corpus a permis de créer des règles d'extraction (patrons) fondées sur ces questions et sur les structures répétées dans les réponses. Le système CasSys, d'autre part, nous a permis de ne pas développer un nouvel outil. Il s'agissait d'adapter CasSys à nos données, ce qui nous a semblé être plus économique et approprié à notre tâche.

L'annotation a été réalisée sur cent douze fichiers Transcriber, soit un total de 35,75 Mo. Six fichiers ont été réservés pour les tests et neuf fichiers pour l'évaluation.

L'adaptation de l'outil à ESLO1 s'est faite sur plusieurs niveaux :

– prétraitement du corpus : le découpage en phrases d'Unitex a été remplacé par un découpage en fonction des balises Transcriber, c'est-à-dire en général par un découpage en tours de parole[31] ;
– enrichissement des cascades par des dictionnaires et des graphes spécifiques ;
– élaboration de la typologie des entités adéquates à notre besoin, à partir de la typologie de la campagne d'évaluation Ester2 (*campagne d'évaluation des systèmes de transcription enrichie d'émissions radiophoniques*)[32].

Il était indispensable de prévoir aussi la présence éventuelle de disfluences et de reprises syntaxiques, ainsi que d'insertions et d'amorces, comme par exemple dans :

– *je m'appelle euh Patrick Mallon*

4.4.2. *Jeu d'étiquettes choisi*

Le contenu d'annotation a été guidé par les questions posées au locuteur sur lui ou sa famille, questions qui portaient sur :

– l'identité (date de naissance, date d'arrivée à Orléans, âge, origine, date de mariage, etc.) ;
– le travail (métier, secteur d'activité, lieu ou nom d'entreprise, etc.) ;
– l'engagement (associations, syndicats, etc.) ;
– les études (diplômes, lieux ou établissements) ;
– les voyages.

La typologie ainsi définie concerne les informations sur la personne interrogée (*pers.speaker*), son conjoint (*pers.spouse*), ses enfants (*pers.child*) et d'autres membres de la famille (*pers.parent*) : voir le tableau 2[33].

---

31. Méthode recommandée par Anne Dister (2007).
32. http://www.afcp-parole.org/ester/docs/Conventions_EN_ESTER2_v01.pdf

Personne (+pers)	la personne interrogée (+speaker)
	son conjoint (+spouse)
	ses enfants (+child)
	les autres membres de la famille (+parent)
Identité (+identity)	le nom (+name)
	l'adresse (+addr)
	l'âge (+age)
	le mariage (+wedding)
	l'origine (+origin)
	la naissance (+birth)
	l'arrivée à Orléans (+arrival)
	le nombre d'enfants (+children)
Travail (+work)	métiers (+occupation)
	secteur d'activité (+field)
	lieu de travail (+location)
	entreprise (+business)
Engagement (+involvement)	association (+voluntary)
	militaire (+military)
	scolaire (+school)
	syndical (+tradeunion)
Voyage (+trip)	études (+study)
	vacances (+holiday)
	professionnel (+work)
Etudes (+study)	lieu (+location)
	diplôme (+degree)
	établissement (+edu)

Tableau 2. *Typologie des entités dénommantes*

33. Pour plus de détails, voir (Maurel *et al.*, 2011).

Ainsi, nous annotons tout d'abord le sujet sur qui porte l'information : le locuteur ou les autres membres de sa famille ; nous précisons ensuite la nature de cette information : l'identité, le travail, les études, l'engagement associatif ou syndicale, les vacances.

Voici quelques exemples d'annotations, tout d'abord pour les entités nommées[34] :

– chez moi<ENT type="pers.hum"> Bérénice Nutal</ENT>
– moi je suis native de<ENT type="loc.admi"> Pithiviers</ENT> j'aime mieux <ENT type="loc.admi">Orléans</ENT>

Puis pour les entités dénommantes :

– <DE type="pers.child"> il est parti <DE type="work.location"> à <ENT type="loc.admi"> Paris </ENT></DE> il travaille dans les <Sync time="1526.195"/> <DE type="work.field"> dans les assurances </DE></DE>
– alors <DE type="pers.speaker"> <DE type="identity.name"> je suis <ENT type="pers.hum"> monsieur Gabrion </ENT></DE></DE> <DE type="pers.speaker"> je suis <DE type="work.occupation"> ingénieur chimiste </DE> </DE>
– <DE type="pers.speaker"> <DE type="identity"> je m'appelle euh <ENT type="pers.hum"> Patrick Mallon </ENT></DE></DE>

Le corpus annoté a été vérifié ensuite manuellement. L'évaluation des résultats a été détaillée dans (Maurel *et al.*, 2009). Les entités dénommantes ont été reconnues avec la précision estimée à 94,2 % et le rappel de 84,4 %.

4.4.3. *Exemple d'utilisation des données annotées : l'anonymisation*

Nous avons déjà mentionné les enjeux importants de l'anonymisation des données orales lorsqu'elles sont mises en libre distribution. Pour anonymiser ESLO1, les entités dénommantes qui renvoient vers les informations personnelles concernant le locuteur et sa famille et qui peuvent éventuellement permettre sa reconnaissance, ont été repérées et étiquetées. On procède ensuite à l'analyse manuelle consistant en la validation des éléments annotés dans un contexte. Ceux qui identifient directement le locuteur sont remplacés par un hyperonyme.

Les premières formes de remplacement choisies sont : *NPERS*[35] pour un nom de personne, *NLIEU* pour un nom de lieu, et *NPROF* pour un nom de profession. En deuxième lieu, nous les remplacerons par des identifiants uniques et anonymes ce

34. La cascade utilisée pour les entités nommées est disponible sous licence LGPL-LR à l'URL : http://tln.li.univ-tours.fr/Tln_CasEN.html

35. L'anonymisation du fichier texte a été réalisée sur la plus petite partie possible (par exemple, le nom mais pas le prénom).

qui devrait permettre de traiter les phénomènes tels que les coréférences. Pour le fichier son, nous avons utilisé le logiciel Praat[36] et un script développé par D. Hirst[37]. La procédure consiste à segmenter l'enregistrement (création d'intervalles correspondant exactement à la partie du signal qui doit être brouillée), puis nous avons annoté ces intervalles avec un code, « buzz », et le script opère automatiquement un traitement du signal.

Comme il a été mentionné au début de l'article, la méthodologie a été modifiée pour ESLO2. L'impossibilité de rendre le processus totalement automatique nous a conduits à la simplification de l'anonymisation pour ESLO2 qui consiste dans le remplacement manuel par un hyperonyme d'un élément identifieur et se fait dès la transcription (figure 2).

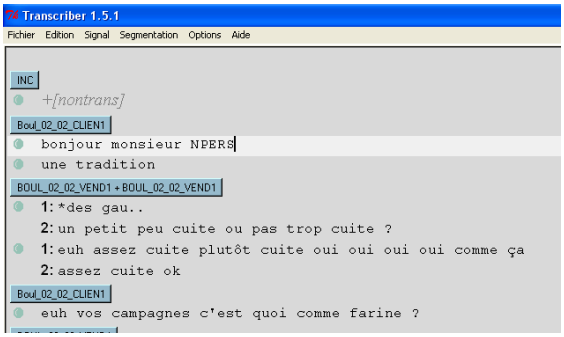


Figure 2. *L'anonymisation du corpus ESLO2*

4.4.4. *Information annotée versus information contenue dans la base de données*

ESLO contient de nombreuses métadonnées permettant une catégorisation sociologique des locuteurs et de la situation. Comme nous l'avons évoqué, celles-ci sont stockées dans une base de données et sont donc accessibles pour un traitement informatique, notamment sous la forme de requêtes croisées.

La base de données répertorie les différentes informations sur le locuteur, comme le montre la figure 3.

Ces informations sont remplies manuellement par la personne qui a fait un enregistrement, qu'on appelle un contributeur chercheur. Certains champs comme le sexe, la catégorie professionnelle INSEE, la situation de famille, etc. sont normalisés

36. http://www.fon.hum.uva.nl/praat/
37. http://uk.groups.yahoo.com/group/praat-users/files/Daniel_Hirst/anonymise_long_sounds.praat

et permettent leur interrogation dans les requêtes par les utilisateurs[38]. D'autres champs sont libres comme les remarques, les informations sur les enfants, la profession en termes propres, etc. et ne peuvent être interrogés.

L'annotation effectuée automatiquement par CasSys reprend certaines informations contenues dans la base de données, comme l'année et le lieu de naissance, par exemple, mais, contrairement à celle-ci, l'information annotée pourrait être interrogée. L'annotation réalisée apporte des informations complémentaires non contenues dans la base de données comme celles sur les loisirs, les vacances ou encore la vie associative du locuteur. Ces informations sont parfois indiquées dans la base de données dans un champ « remarques diverses », mais d'une façon non systématique et ne permettant pas leur étude.

Fiche locuteur

Identifiant locuteur : BA725

Anonyme:	OUI
Année de naissance:	1912
Tranche d'âge:	55/65
Lieu de naissance:	loiret
Sexe:	Homme
Niveau d'études:	CEP
Commentaire:	Enseignement primaire à Orléans
Age de fin d'études:	14
Catégorie Professionnelle (INSEE):	Artisans, commerçants et chefs d'entreprise
Profession en termes propres:	boucher, gérant boucherie supermarché
Langue(s):	Français
Commentaire niveau langue:	
Situation de famille:	Marié
Année d'arrivée:	1912
Domicile:	Orléans centre
Nombre d'enfants:	2
Information sur les enfants:	fils 1 : coiffeur, fils 2 ?
Remarques diverses:	Famille : femme sans activitéfils, brevet, coiffeurfilsEnseignement : primaire à orléansdiplôme : CEPProblème : non-renseigné
Fiche modifiée par:	obaude
Enregistrements et transcriptions:	▪ Enregistrement ESLO1_ENT_001 • Transcription ESLO1_ENT_001_A • Transcription ESLO1_ENT_001_B • Transcription ESLO1_ENT_001_C

Figure 3. *Fiche du locuteur de l'enregistrement 008*

Le corpus annoté des entités nommées et dénommantes n'étant pas anonymisé, il ne peut pas être mis à disposition librement sur le Web, l'accès en est restreint à la communauté scientifique sous réserve de signature d'une convention.

38. Les quatre champs : nom, prénom, nom de jeune fille et adresse, sont stockés dans une base de données séparée et ne sont plus accessibles pour le public (voir la section 3 de cet article).

Nous envisageons ultérieurement de relier les deux processus, pour que l'information annotée automatiquement soit stockée directement dans la base de données. Cette tâche est prévue dans la suite des travaux sur ESLO.

Parallèlement à l'annotation des informations personnelles concernant le locuteur, nous avons pu procéder aux premiers tests de la phase de traitement consacrée à l'annotation morphosyntaxique.

4.5. *Annotation morphosyntaxique*

Le travail développé dans cette partie est encore en cours de réalisation. Il s'agit de l'annotation morphosyntaxique, qui consiste à attribuer à chaque unité lexicale du corpus une étiquette apportant certaines informations (sa catégorie syntaxique, ses éventuels genre, nombre, temps verbal, etc.) dans le contexte où elle apparaît. Cette étape est précédée par une phase de segmentation dans laquelle le logiciel doit reconnaître et séparer les unités lexicales les unes des autres.

L'étiquetage morphosyntaxique est important pour la mise à disposition et la consultation du corpus car il permet de faire des requêtes précises. On pourrait ainsi vouloir extraire tous les noms communs employés par un certain locuteur, tous les verbes d'un autre, ou encore analyser les différentes prépositions utilisées après un certain verbe, etc. Comme le travail décrit ici n'est pas fini, la consultation d'ESLO en utilisant ce type de critères n'est pas encore possible, mais est envisagée ultérieurement.

Il existe des outils variés pour l'étiquetage morphosyntaxique : libres (TreeTagger[39], Sem[40], Melt[41], LGTagger[42]) ou payants (Cordial[43]). Le problème majeur est que ces outils ne sont pas adaptés à l'oral, caractérisé par les phénomènes de disfluences : répétitions, autocorrections, amorces de mots, etc. Cette tâche est d'autant plus difficile que les transcriptions ne sont pas ponctuées (voir la section 4.2).

Pour étiqueter ESLO, nous avons choisi de développer notre propre étiqueteur en utilisant la technique d'apprentissage automatique la plus adaptée pour cela à l'heure actuelle : les CRF (Lafferty *et al.*, 2001 ; Sutton et McCallum, 2006 ; Tellier et Tommasi, 2011), qui permettent de construire un modèle statistique à partir de données étiquetées fournies en exemple. L'objectif est de développer un étiqueteur libre et gratuit adapté au jeu d'étiquettes établi, en tenant compte des spécificités du corpus traité.

---

39. http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
40. http://www.univ-orleans.fr/lifo/Members/Isabelle.Tellier/SEM.html
41. (Denis et Sagot, 2010).
42. http ://igm.univ-mlv.fr/˜mconstan/research/software
43. http://www.synapse-fr.com/Cordial_Analyseur/Presention_Cordial_Analyseur.htm

En 2010, nous avons réalisé des expériences préliminaires exploitant les CRF, en partant d'un corpus d'apprentissage déjà segmenté et annoté manuellement. Le programme développé permettait d'attribuer une étiquette morphosyntaxique à une unité lexicale selon trois niveaux : POS[44], informations morphologiques (genre, nombre, etc.), informations sémantiques et/ou syntaxiques, comme dans les exemples suivants :

oui	ADV	ADV	ADV
en_effet	ADV	ADV	ADV
on	P	P3I	P3IPER
peut	V	V3SINDP	V3SINDP
commencer	V	VINF	VINF
rire	V	VINF	VINF

Cette structure en trois niveaux autorise une certaine souplesse, suivant la nature et la qualité de l'information attendue : le premier niveau est plus simple à étiqueter et donc plus fiable, le troisième niveau inclut des informations linguistiques plus riches mais est potentiellement plus sujet à l'erreur d'étiquetage. On pourrait ainsi faire des requêtes plus ou moins précises, faisant appel spécifiquement à un certain niveau. Les premières expériences ont été décrites dans (Eshkol *et al.*, 2010), la phase de segmentation étant supposée être déjà réalisée. Elles utilisaient un ensemble d'entraînement assez réduit de 1 723 « énoncés » (tours de parole ou séquences entourées de « blancs ») étiquetés comportant 18 424 unités lexicales. Elles ont montré qu'il était possible d'apprendre un étiqueteur atteignant de 89 % d'exactitude (pour le troisième niveau) à 94 % (pour le premier niveau). Ce qui est moins bon que les performances annoncées des meilleurs étiqueteurs actuels du français, qui obtiennent entre 97 et 98 % d'exactitude (Denis et Sagot 2010, Constant *et al.*, 2011) avec un jeu d'étiquettes comparable à notre premier niveau. Mais ces derniers étiqueteurs ont été appris et testées sur le French Treebank, avec plus de 300 000 unités lexicales en entraînement. Ce corpus extrait d'articles du journal *Le Monde* est rédigé dans une langue beaucoup plus normée que le nôtre. L'oral présentant plus d'irrégularités, il est à prévoir que les modèles statistiques entraînés sur ESLO requièrent encore plus d'exemples pour parvenir à des résultats équivalents.

Dans la suite du travail, nous avons réfléchi à quelques modifications dans le jeu d'étiquettes, en prenant encore plus en compte les spécificités de l'oral. Certaines nouvelles étiquettes telles que « marqueurs discursifs » (MD) incluant trois sous-classes : MD (marqueurs discursifs propres), MDEUH (*euh* d'hésitation) et MDINT (interjections) et « présentatif » (PRES) pour les structures comme « c'est », « voici », etc. ont été ajoutées. Pour essayer d'éviter l'ambiguïté comme, par exemple, celle du participe passé dans « je suis prise/je suis partie »[45], nous avons aussi introduit une étiquette pour le passif.

---

44. *Part of speech*.

45. Dans les deux cas, le participe passé du verbe est précédé du verbe auxiliaire « être ».

Nous prévoyons de procéder à la segmentation préliminaire du texte par un segmenteur qui sera également appris avec un CRF. Il est aussi possible d'apprendre conjointement à segmenter et à étiqueter un texte, comme cela a été fait dans (Constant *et al.*, 2011), mais nous préférons pour l'instant réaliser les deux étapes indépendamment, parce que notre jeu d'étiquettes morphosyntaxiques est plus riche que le leur, ce qui risque d'augmenter les erreurs d'étiquetage. Pour apprendre à segmenter, le plus simple est de découper au maximum les unités du texte et de « recoller les morceaux » des mots composés par une phase d'annotation. Pour cela, nous segmentons tout d'abord le texte au maximum sur une base formelle, en prenant l'espace, l'apostrophe et les ponctuations comme séparateurs. Puis le texte est mis en format tsv[46] et des connaissances externes (comme l'étiquette fournie par un autre étiqueteur) sont ajoutées. L'étiquetage en B (pour « *Begin* », désignant le début d'une unité lexicale) et I (pour « *In* », désignant la suite d'une unité lexicale commencée précédemment) marquant les frontières du mot est suivi par leur fusion dans le cas des mots composés.

```
Pomme    N       B
De       PREP    I    =>    pomme_de_terre    N_PREP_N
terre    N       I
```

Comme connaissances externes (la colonne intermédiaire dans notre exemple), nous utilisons les résultats de l'étiqueteur libre SEM et des ressources linguistiques libres comme le lexique du Lefff et les tables de verbes et de noms prédicatifs du Lexique-Grammaire, passées au format alexina du Lefff. Cette étape est en cours de finalisation.

La deuxième phase de l'étiquetage morphosyntaxique s'appuiera sur cette segmentation. Les expériences d'apprentissage tenant compte des nouvelles étiquettes sont en cours. Le nouveau modèle d'annotation sera appris à partir d'un corpus de référence annoté manuellement ainsi que de connaissances externes (les mêmes que pour la segmentation).

Nous travaillons aussi sur la compatibilité de l'étiquetage avec les fichiers XML de Transcriber pour permettre la synchronisation avec les fichiers sonores. Le fichier de transcription, est prétraité pour donner un fichier en texte brut qui sera segmenté et étiqueté. Le résultat sera ensuite fusionné avec le fichier de départ pour donner un fichier étiqueté compatible avec le format original. On conservera ainsi les performances de notre étiqueteur et le même procédé pourra être applicable à n'importe quel autre format d'entrée et de sortie.

```
<Turn speaker="spk2" startTime="5.0" endTime="7.533">
<Sync time="5.0"/>
<Sync time="5.03"/>
et qu'est-ce qui vous a amené à vivre à Orléans
</Turn>
=>
```

46. *Tab separated values.*

```
<Turn speaker="spk2" startTime="5.0" endTime="7.533">
<Sync time="5.0"/>
<Sync time="5.03"/>
<w total="CONJCOO"> et </w>
<w total="PIINT"> qu'est-ce qui </w>
<w total="P2PPERCOMPL"> vous </w>
<w total="V3SINDPAUX"> a </w>
<w total="VMSPP"> amené </w>
<w total="PREP"> à </w>
<w total="VINF"> vivre </w>
<w total="PREP"> à </w>
<w total="NP"> Orléans </w>
</Turn>[47]
```

Les étiquettes de cet exemple sont celles du troisième niveau, et donc les plus précises. Elles intègrent, en quelque sorte, les niveaux précédents, permettant de retrouver si besoin tous les verbes (dont les étiquettes sont de la forme V*) indépendamment de leurs flexions. Mais si les étiquetages des différents niveaux sont de précisions notablement différentes, on peut aussi garder trois attributs différents pour chaque mot, correspondant à chacun de ces trois niveaux. Dans ce cas, pour rechercher les verbes, il suffira d'interroger uniquement l'attribut de premier niveau.

5. Consultation

Cette dernière partie est consacrée à la consultation d'ESLO. L'objectif du projet est de rendre le corpus disponible pour une large communauté scientifique mais aussi pour le grand public. L'application Web est disponible sous licence Creatives Commons[48].

La mise à disposition du corpus pose des questions sur sa consultation par les différents profils d'utilisateurs : chercheurs, contributeurs ou grand public. L'accès au corpus ne peut pas être le même pour chacune de ces catégories. Un corpus oral comme ESLO rend la tâche encore plus compliquée en raison de la masse et de la diversité des données et des métadonnées pouvant être consultées.

L'application Web de consultation des corpus ESLO se décompose en trois parties : la partie institutionnelle destinée à publier des informations sur le projet, la partie publique destinée à un large public et la partie administration.

Les quatre types d'utilisateurs sont pris en compte : administrateur, contributeur qui peut en outre ajouter ou modifier des informations d'enregistrement ou de

---

47.CONJCOO = Conjonction de coordination, PIINT = Pronom invariable interrogatif, P2PPERCOMPL = Pronom 2e personne pluriel personnel complément, V3SINDPAUX = Verbe 3e personne singulier indicatif présent auxiliaire, VMSPP = Verbe masculin singulier participe passé, PREP = Préposition, VINF = Verbe infinitif, NP = Nom propre.

48. Application réalisée par le prestataire ARES-GFI.

transcription, chercheur invité qui aura les mêmes accès qu'un utilisateur non authentifié avec en plus la possibilité de voir les données non anonymisées et toutes les versions de transcription et enfin l'utilisateur grand public.

On peut consulter le catalogue selon trois axes indépendants : les enregistrements, les transcriptions et les locuteurs auxquels sont associées les métadonnées (pour les enregistrements : type, durée, lieu, etc. ; pour les transcriptions : nom du transcripteur, problèmes et remarques, etc. ; et pour les locuteurs, leurs caractéristiques sociologiques : date et lieu de naissance, sexe, profession, etc.).

Des recherches simples et des recherches avancées sont possibles en précisant le corpus (ESLO1 ou ESLO2), les descripteurs de métadonnées de l'axe considéré ou encore selon l'occurrence (figure 4). La sélection de la version des transcriptions (brute, relue ou validée) est possible pour les profils administrateur, contributeur et chercheur invité. L'utilisateur peut faire des requêtes sur une occurrence ou en utilisant quelques expressions régulières. L'exploration du lexique d'une transcription choisie est affichable également, il s'agit d'une liste de mots avec leur fréquence. Parmi les autres actions réalisables sur le site, on peut naviguer synchroniquement dans la transcription et le fichier son, ou bien télécharger un enregistrement ou une transcription.

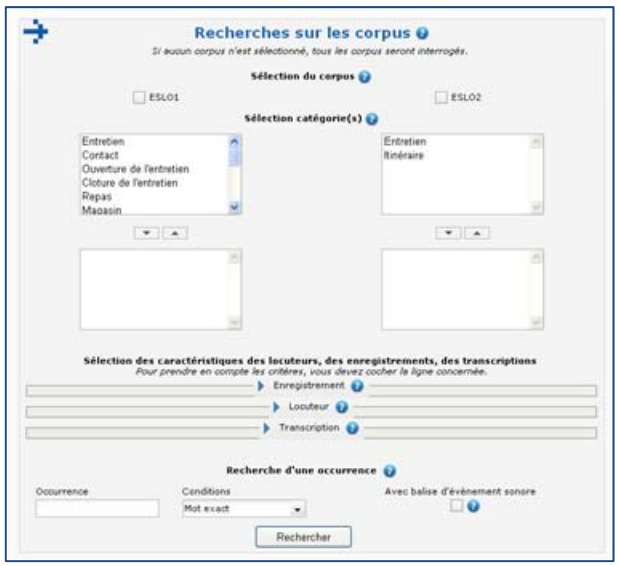


Figure 4. *Interface Web*

Pour des raisons juridiques, une partie non anonymisée des données n'est accessible qu'aux chercheurs et ce, après identification et convention (non diffusion et respect de la confidentialité) ce qui est le cas du corpus annoté des entités nommées et dénommantes. Les transcriptions rendues disponibles au grand public seront les troisièmes versions. Les utilisateurs auront d'ailleurs la possibilité de proposer des transcriptions alternatives avec des corrections.

Les données anonymisées sont également disponibles par l'entrepôt du CRDO-TGE Adonis et par le site Corpus de la parole de la DGLFLF – ministère de la Culture.

Dans le travail sur la consultation du corpus, nous avons été guidés par les différents profils d'utilisateurs potentiels avec la volonté de leur donner accès à des données variées et riches constituant ESLO.

6. Conclusion

En conclusion, nous rappelons quelques objectifs du projet ESLO qui ont guidé notre équipe tout au long du travail.

ESLO est un travail de longue haleine. Il s'agit tout d'abord de constituer un grand corpus de français parlé de quelque 700 heures en préservant prioritairement l'hétérogénéité maximale des données observées et contenant de nombreuses informations complémentaires (profil du locuteur, caractéristiques de la situation, etc.).

La mise à disposition de ce corpus a nécessité d'anticiper les contraintes juridiques et techniques qui ne peuvent être dissociées des cadres théoriques du projet de recherche. Les différentes opérations de traitement des données ont un impact fort sur la constitution de l'objet scientifique et ne peuvent être considérées comme des opérations de prétraitements des données en préalable au travail d'analyse. Ainsi ce projet aborde résolument une démarche réflexive. Il ne s'agit pas seulement de recueillir et de rendre disponibles des données et métadonnées langagières mais aussi de rendre explicite l'ensemble de la chaîne qui permet d'y arriver, de la collecte à l'analyse, en passant par la transcription et les autres opérations d'annotation. On se rend alors compte que toutes les opérations sont liées les unes aux autres, et que les choix qui sont faits doivent l'être en prenant en compte l'ensemble des objectifs et des contraintes, à tous les niveaux.

Le projet ESLO n'est pas achevé. Le corpus ESLO2 est en cours puisque des enregistrements d'Orléanais d'aujourd'hui dans des situations variées continuent. Les travaux sur l'annotation se poursuivent. Comme nous l'avons mentionné, nous sommes en train de travailler sur l'étiquetage morphosyntaxique par apprentissage automatique avec les CRF. Le nouveau modèle d'annotation sera appris à partir d'un corpus de référence annoté manuellement ainsi que de connaissances externes comme les résultats de l'étiqueteur libre SEM et des ressources linguistiques libres

comme le lexique du Lefff et les tables de verbes et de noms prédicatifs du Lexique-Grammaire, passées au format alexina du Lefff. L'étiquetage sera compatible avec les fichiers XML de Transcriber pour permettre la synchronisation avec les fichiers sonores. D'autres couches d'annotations sont programmées sur ESLO, comme l'annotation syntaxique en *chunks*[49], par exemple. On citera également le projet ANCOR[50] décrit dans (Schang *et al.,* 2011) consistant dans l'annotation et l'étude des coréférences à l'oral et se fondant majoritairement sur le corpus ESLO.

Cent vingt et un entretiens d'ESLO1 annotés des entités nommées et dénommantes ou un corpus de 1 723 « énoncés » (tours de parole ou séquences entourées de « blancs ») comportant 18 424 unités lexicales étiquetées morphosyntaxiquement représentent des ressources riches et très utiles pour les techniques d'apprentissage.

Le corpus ESLO peut être aujourd'hui exploité de manières diverses. Nous avons mentionné un exemple du travail qui a croisé l'information annotée avec les métadonnées. Il s'agit de l'analyse de la variation des valeurs numériques de temps dans les fichiers de transcription en fonction de différents profils sociologiques des locuteurs à l'aide des méthodes statistiques (pour plus de détails voir Dupont *et al.*, 2012). La richesse des données que permet d'étudier le corpus ESLO est importante. Sa mise à disposition permettra ainsi leur meilleure exploitation.

7. Bibliographie

Abney S., Parsing by chunks. In Berwick R., Abney R. et Tenny C., éditeurs : Principlebased Parsing. Kluwer Academic Publisher, 1991.

Baude O., *Corpus oraux : guide des bonnes pratiques*, CNRS-Editions et Presses universitaires d'Orléans, 2006.

Baude O., Dugua C. « (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ? », vol. 10, *Corpus, Varia,* 2011.

Bergounioux G., Baraduc J., Dumont C., « L'étude socio-linguistique sur Orléans (1966-1991): 25 ans d'histoire d'un corpus », n°93, *Langue française*, 1992, p. 74-93.

Blanche-Benveniste C., Jeanjean C., *Le français parlé, transcription et édition*, Paris, Didier érudition, 1987.

Cappeau P., Gadet F., « Où en sont les corpus sur les français parlés ? » *Revue Française de Linguistique Appliquée*, Vol. XII, 2007, p. 129-133.

49. Les *chunks* sont des constituants continus et non récursifs (Abney, 1991) qui définissent la structure syntaxique des phrases ou énoncés.

50. http://tln.li.univ-tours.fr/Tln_Ancor.html

Constant M., Tellier I., Duchier D., Dupont Y., Sigogne A., Billot S., « Intégrer des connaissances linguistiques dans un CRF : application à l'apprentissage d'un segmenteur-étiqueteur du français », *TALN2011,* Montpellier, 2011.

Cresti E., Moneglia M. *C-ORAL-ROM Integrated Reference Corpora for Spoken Romance Languages*, Studies in Corpus Linguistics 15. Amsterdam : Benjamins, 2005.

Debrock M., Mertens P., Truyen F., Brosens V., *ELICOP, Étude LInguistique de la COmmunication Parlée: Constitution et exploitation d'un corpus de français parlé automatisé*, K.U.Leuven: Departement Linguïstiek, 2000.

Denis P., Sagot B. « Exploitation d'une ressource lexicale pour la construction d'un étiqueteur morpho-syntaxique état-de-l'art du français », *TALN2010*, Montréal, 2010.

Dister A,. De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelles orales VALIBEL, Thèse de doctorat, Université catholique de Louvain, 2007.

Dupont A., Eshkol-Taravella I., Delsol L., « Étude d'application des méthodes et des outils statistiques sur les données du corpus ESLO : cas de la question sur mai 68 », *11es Journées Internationales d'analyse statistique des données textuelles JADT 2012,* Liège, 13-15 juin, 2012 (à paraître).

Eshkol I., « Entrer dans l'anonymat. Etude des "entités dénommantes" dans un corpus oral », *Eigennamen in der gesprochenen Sprache*, Narr Francke Attempto Verlag GmbH, Germany, 2010, p. 245-266.

Eshkol I., Tellier I, Taalab S., Billot S., « Étiqueter un corpus oral par apprentissage automatique à l'aide de connaissances linguistiques », *10es Journées Internationales d'analyse statistique des données textuelles JADT 2010*, Rome, 9-11 juin, 2010.

Friburger N., Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques, Thèse de doctorat, Université François Rabelais Tours, 2002.

Hriba L., Identification automatique des locus de variation dans un corpus de français parlé, Thèse de doctorat. Université d'Orléans, Orléans, en cours.

Kaufmann A., « The Santa Barbara Corpus of Spoken American English. Part 1 », *Journal of Pragmatics* 34, 2002, p. 1309-1316.

Lafferty J., McCallum A., Pereira F., « Conditional random fields : Probabilistic models for segmenting and labeling sequence data », *Proceedings of ICML'01*, 2001, p. 282–289.

Leech G., « Introduction corpus annotation ». In Garside R., Leech G., McEnery A., (Eds.), *Corpus annotation: Linguistic information from computer text corpora*. London: Longman, 1 :18, 1997.

Lonergan J., Kay J., Ross J., *Etude sociolinguistique sur Orléans, catalogue des enregistrements*, Colchester: Multigraphié, 1974.

Maurel D., Friburger N., Antoine J.-Y., Eshkol-Taravella I., Nouvel D., Cascades autour de la reconnaissance des entités nommées, *TAL* 52-1, 2011.

Maurel D., Friburger N., Eshkol I., « Who are you, you who speak? Transducer cascades for information retrieval », *4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, Poznań, Poland, 6-8 novembre, 2009, p. 220-223.

Mertens, P., « Les corpus de français parlé ELICOP : consultation et exploitation », in Binon, J.*, et al.* (éd.) *Tableaux Vivants. Opstellen over taal-en-onderwijs aangeboden aan Mark Debrock*. Leuven: Universitaire Pers., 2002.

Mettouchi, A., C. Chanard, « From Fieldwork to Annotated Corpora: the CorpAfroAs Project », *Faits de Langue-Les Cahiers*, 2 : 255-265, 2010.

Nazarenko, A., « Le point sur l'état actuel des connaissances en traitement automatique du langage (TAL) », *Compréhension des langues et interaction*, Lavoisier, 2006, p. 31-70.

Paumier S., De la reconnaissance de formes linguistiques à l'analyse syntaxique, Thèse de Doctorat, Université de Marne-la-Vallée, 2003.

Schang E., Boyer A., Muzerelle J., Antoine J-Y., Eshkol I., Maurel, D. « Coreference and Anaphoric Annotations for Spontaneous Speech Corpora In French », *The 8th Discourse Anaphora and Anaphor Resolution Colloquium (DAARC2011)*, Faro, Algarve, Portugal, 6-7 October, 2011.

Sutton C., McCallum A., « An Introduction to Conditional Random Fields for Relational Learning », In L. Getoor and B. Taskar, Eds., *Introduction to Statistical Relational Learning*. MIT Press, 2006.

Tellier I., Tommasi M., « Champs Markoviens Conditionnels pour l'extraction d'information », dans *Modèles probabilistes pour l'accès à l'information textuelle*, Hermès, 2011, p. 223-267.