

Automatiser le processus d'anonymisation des corpus oraux : le cas d'ESLO

Iris Eshkol-Taravella¹ Olivier Baude¹ Loyal Kanaan-Caillol¹ Denis Maurel²

(1) LLL, UMR 7270, CNRS, UFR LLSH, 10 Rue de Tours 45065 ORLEANS cedex 2

(2) Université François-Rabelais de Tours, LI

iris.eshkol@univ-orleans.fr, olivier.baude@univ-orleans.fr, loyal.kanaan@univ-orleans.fr,
denis.maurel@univ-tours.fr

Résumé. Cet article aborde la question de l'anonymisation automatique des corpus oraux afin de permettre leur utilisation et diffusion sur la Toile. Nous proposons une analyse des éléments permettant l'identification du locuteur que nous appelons les entités dénommantes et qui dépassent par leur diversité et leur hétérogénéité les entités nommées. Nous décrivons ensuite une expérimentation à partir d'un module de reconnaissance automatique dans les transcriptions. Celle-ci permet d'étudier la différence entre un processus de repérage automatique d'informations personnelles et le processus d'anonymisation qui demande en plus une intervention humaine. Cette intervention permet de distinguer les informations personnelles sur le locuteur qui ne nécessitent pas d'être masquées dans la transcription et les informations permettant l'identification et qui donc doivent être anonymisées. Enfin nous détaillerons la procédure actuelle utilisée et qui prend en compte les résultats de l'expérimentation et les contraintes de production dans le workflow de la chaîne de traitement d'un grand corpus oral.

Abstract. This article tackles the question of oral corpus anonymisation in preparation for its diffusion on the Web. We first analyze elements corresponding to personal information about the speaker which we call designating entities and which exceed named entities by their diversity and heterogeneity. Then we describe an experiment based on a module of automatic recognition in the transcriptions that enables us to study the difference between the automatic recognition of personal information process and the anonymisation process that requires an additional human intervention. This latter allows distinguishing the speaker personal information that does not need to be hidden in the transcription and the identifying information that requires anonymisation. Finally, we describe the current procedure we use which takes into account the results of the experiment in the chain of treatments' workflow of a large oral corpus.

Mots-clés : anonymisation, anonymisation automatique, corpus oral, entités dénommantes, information personnelle, information identifiante

Keywords: anonymisation, automatic anonymisation, oral corpus, designating entities, personal information, identifying information

1 Introduction

Grâce au développement des outils informatiques, la mise à disposition de différents corpus a modifié le travail des chercheurs en linguistique, en sciences sociales et humaines et en traitement automatique des langues (TAL). Les initiatives actuelles se développent autour de la diffusion et de la disponibilité de ces ressources en accès –souvent libre- sur la Toile. Les corpus oraux en langues étrangères le BNC¹, le Russian National Corpus² ou encore le National Corpus of Polish³, ou en français, CLAPI⁴, PFC⁵, CRFP⁶, Corpus de la parole, etc. sont apparus sur le Toile et plus récemment la France s'est doté d'un EQUIPEX dédié à la diffusion des ressources linguistiques (EQUIPEX

¹ British National Corpus, <http://www.natcorp.ox.ac.uk/>

² <http://www.ruscorpora.ru/en/index.html>

³ <http://nkjp.pl/index.php?page=0&lang=1>

⁴ Corpus de langues parlées en interaction, <http://clapi.univ-lyon2.fr/>

⁵ Phonologie du français contemporain, <http://www.projet-pfc.net/?accueil:intro>

⁶ Corpus de référence du français parlé, <http://www.up.univ-mrs.fr/delic/crpf>

ORTOLANG). Pour diffuser ces corpus, les questions juridiques dont celle de leur anonymisation se sont avérées primordiales.

La linguistique sur corpus oraux a bénéficié d'un travail précurseur pour la collecte et la diffusion d'enregistrements sonores et de leurs transcriptions. Sous l'égide du Ministère de la Culture et du CNRS un groupe de travail constitué de linguistes, d'informaticiens, de juristes et de conservateurs a réfléchi aux aspects juridiques et éthiques de l'usage des corpus oraux. Ce travail s'est concrétisé par la publication de l'ouvrage *Corpus oraux, guide des bonnes pratiques 2006* (Baude et al. 2006). L'anonymisation est une pratique qui répond à un impératif juridique précis. Sans recueil du consentement de la personne enregistrée, il est obligatoire d'empêcher son identification. L'impossibilité d'identifier est une notion complexe qu'on a trop souvent réduite à l'effacement des noms propres. La tâche est bien plus difficile, mais aussi plus stimulante pour les recherches en linguistique et en TAL.

L'anonymisation relève de procédures différentes selon qu'on traite l'enregistrement sonore, sa transcription ou les métadonnées descriptives. Toutefois dans tous les cas l'objectif reste le même. Si selon certains juristes la voix est une donnée identifiante ce qui nécessiterait de modifier le signal acoustique de tout enregistrement et par là même obérerait toute recherche en linguistique, les pratiques des chercheurs s'orientent plus généralement vers un traitement des données personnelles au sens large. Que ce soit sur l'oral ou sur l'écrit celles-ci sont diverses, il peut s'agir d'une forme nominative, d'une profession, d'un statut, d'une caractéristique physique, etc. et/ou du recoupement de plusieurs de ces informations. Si l'on convient que l'anonymisation ne se réduit pas à l'effacement des noms propres, il est nécessaire de définir avec précision quels sont les traitements à effectuer pour répondre à l'objectif de réduire les possibilités d'identification. Dans le cas de grands corpus, ces traitements deviennent une étape fondamentale du travail de constitution du corpus avec des effets très importants sur la gestion et la diffusion des données.

Le travail décrit dans cet article porte sur le corpus oral ESLO (Enquête Sociolinguistique à Orléans). Il s'agit d'un grand corpus de données orales qui regroupe deux enquêtes ESLO 1 et ESLO 2 (Baude et Dugua 2011, Eshkol-Taravella et al. 2012). ESLO 1 a été réalisé entre 1968 et 1974, à l'initiative d'universitaires britanniques avec une visée didactique : l'enseignement du français langue étrangère dans le système public d'éducation anglais. Il a été numérisé et transcrit par l'équipe du Laboratoire Ligérien de Linguistique (LLL). ESLO2 est une nouvelle enquête, débutée en 2008 par le LLL. Réunis, ESLO 1 et ESLO 2 forment une collection de 700 heures d'enregistrement (10 millions de mots), ce qui est considéré aujourd'hui comme une valeur repère pour les investigations projetées. Il s'agit en somme d'un très grand corpus dont l'objectif de mise à disposition (sons – transcriptions – métadonnées) a déclenché une réflexion sur les informations permettant l'identification du locuteur, sur leur repérage automatique, ce qui a abouti à la définition et la mise en place de la phase d'anonymisation dans la chaîne de traitement.

Dans la première partie de l'article, nous présenterons l'expérience d'automatisation de l'anonymisation sur un sous-corpus d'ESLO1. Nous décrirons ensuite le processus qui a finalement été mis au point afin de traiter rapidement les centaines d'heures d'enregistrement au cœur du workflow de transcription. Nous finirons par quelques conclusions concernant les limites des outils du TAL dans ce projet.

2 Anonymisation et TAL

Traditionnellement la tâche d'anonymisation dans le TAL est décomposée en deux étapes : le repérage des entités nommées (noms de personnes, lieux, organisations, âges, etc.) et leur substitution par un hyperonyme ou un élément à référents multiples (le nom de famille *Dupont*, par exemple).

L'anonymisation dans le domaine du TAL concerne souvent le domaine médical (Meyster et al., 2010, Tweit et al., 2004, Raaj, 2012, Uzuner et al., 2007, Grouin, Zweigenbaum, 2011) et porte ainsi sur les documents écrits (rapports, dossiers médicaux, etc.) où les informations à anonymiser sont assez homogènes et présentées d'une manière linéaire propre aux textes écrits. C'est le cas de l'outil Medina (Medical Information Anonymization⁷) disponible gratuitement qui repère automatiquement à l'aide de patrons et de lexiques les noms de personnes, les lieux, les noms d'hôpitaux et les informations numériques comme les adresses, âges, numéros de téléphones, etc. dans les documents cliniques en français.

Les corpus oraux sont différents des corpus écrits car ils n'utilisent pas qu'un seul support mais associent le plus souvent la parole enregistrée à une représentation écrite et/ou codée (transcriptions, traductions, annotations). La présence des phénomènes caractéristiques de l'oral appelés *disfluences* comme les hésitations, amorces, répétitions, reformulations ou les pauses et les chevauchements entre les locuteurs rompent le flux de la parole. Ainsi, anonymiser le discours d'une

⁷ <http://medina.limsi.fr/>

manière automatique est une tâche difficile qui pose des problèmes supplémentaires au TAL comme cela a récemment été constaté (Amblard, Fort, 2014).

3 Expérience de l'anonymisation automatique sur un sous-corpus d'ESLO1

3.1 Contexte

Afin de mettre à disposition le corpus oral ESLO, une réflexion sur son anonymisation a été mise en place. Du point de vue juridique, le corpus ESLO1, a posé deux problèmes. Premièrement, les locuteurs n'ont rempli aucun document pour exprimer leur consentement ; deuxièmement, les locuteurs de la fin des années soixante ne pouvaient pas prévoir que leurs enregistrements pourraient être diffusés par Internet qui n'existaient pas à l'époque. Dans le cas d'ESLO2, les locuteurs signent un document d'un consentement à la 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.

Avant de prendre la décision finale sur la méthodologie à appliquer dans l'anonymisation du corpus entier, un test a été effectué sur un sous corpus en collaboration avec le laboratoire LI (Laboratoire Informatique) de l'université de Tours. Ce test portait sur un sous-corpus d'ESLO1 (112 entretiens face-à-face). Les entretiens ont été choisis pour la richesse de l'information personnelle présente. Cette caractéristique est due à la nature du corpus qui contient les réponses aux 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. Ce sous-corpus est donc composé des données personnelles riches et plus au moins homogènes. L'expérience visait deux objectifs : repérer et étudier les éléments personnels et identifiants dans le corpus, d'une part, et tester le module de l'anonymisation automatique des transcriptions, d'autre part.

3.2 Méthodologie

Avant de procéder au développement du module de l'anonymisation automatique, une réflexion sur la nature des données permettant l'identification du locuteur a été lancée. Elle a abouti à une définition d'une nouvelle notion : les *entités dénommantes*.

3.2.1 Entités dénommantes

Selon le Dictionnaire d'analyse du discours, "l'identité résulte, à la fois, des conditions de production qui contraignent le sujet, conditions qui sont inscrites dans la situation de communication et/ou dans le préconstruit discursif, et des stratégies que celui-ci met en œuvre de façon plus ou moins consciente" (Charaudeau, Maingueneau, 2002 : 300). Les auteurs distinguent une identité psychosociale consistant en traits qui définissent le sujet selon son âge, son sexe, son statut, etc. et une identité discursive du sujet énonciateur "qui peut être décrite à l'aide de catégories locutives, de modes de prise de parole, de rôles énonciatifs et de modes d'interventions" (ib.) Nous n'allons pas nous intéresser, dans cette étude, aux stratégies discursives que choisit le sujet parlant pour se construire une identité : sa manière de prendre la parole, de thématiser ses propos, d'organiser son argumentation. De la même manière, nous laisserons de côté, les déictiques qui renvoient à la situation spatio-temporelle du locuteur et la voix.

L'objectif que nous nous sommes fixé est d'étudier des éléments dans le discours du locuteur permettant son identification par un éventuel utilisateur du corpus. Nous appelons ces éléments *entités dénommantes* (Eshkol, 2010). Les entités dénommantes "servent à identifier le locuteur en le mentionnant par son nom ou en représentant certains de ses traits" (p.247). Le processus d'anonymisation du corpus consiste donc pour nous à chercher des indices, des traces "visibles" qui permettront d'identifier le sujet parlant dans le discours.

Plusieurs études linguistiques, parlent des traits caractéristiques et définitoires d'un objet. Nous avons cité Charaudeau qui distingue les deux types d'identité du sujet. En sémiotique narrative, (Hamon, 1977) utilise le terme de *qualification différentielle* pour la série de traits indicateurs de l'importance des personnages dans les romans et qui le distinguent des autres. Les entités dénommantes sont donc les éléments descriptifs qui permettent de distinguer le sujet parlant des autres et, par conséquent, de le reconnaître.

Quand on parle de l'information permettant de reconnaître l'individu on se limite souvent aux noms propres. C'est le cas de beaucoup de travaux en TAL sur l'anonymisation automatique des corpus où la reconnaissance de l'information

personnelle s'arrête aux entités nommées : noms de personnes, de lieux, d'organisations et éléments numériques (âge, dates, etc.). Cependant, d'autres éléments peuvent permettre l'identification du locuteur par recoupement :

- noms de métiers : *je suis enseignant dans dans l'école publique*⁸, *comme officier j'ai été obligé de rester 45 ans*
- origine : *oui je suis orléanaise*
- maladies : *j'ai une maladie du foie, ça lui a même occasionné une petite scoliose déformation légère de la colonne vertébrale*

Les entités dénommantes ne sont pas composées que des entités nommées et peuvent être de nature lexicale et sémantique très variée.

Le processus d'identification du locuteur à travers les entités dénommantes a été analysé dans (Eshkol, 2010). Nous présenterons dans la partie qui suit les grandes lignes de ce travail.

Le processus d'identification peut être direct ou indirect d'où la distinction entre :

- identifiant direct (unicité référentielle) : il permet, à lui seul, de distinguer un individu des autres et renvoie directement vers un référent unique ; sa présence est nécessaire et suffisante pour la reconnaissance de l'individu. Le processus n'est pas progressif, il est ponctuel : nom ou métier rares de la personne (*Eshkol, général, maire*), caractéristique rare (*nombre élevé d'enfants, handicap*)
- identifiant non direct : sa présence seule ne permet pas l'identification, mais en combinaison avec d'autres identifiants, il peut renvoyer vers un référent unique (*le locuteur est patron d'un bar au moment d'enregistrement, et avant il travaillait dans l'aviation militaire*). Le processus d'identification est progressif, il se construit au fur et à mesure de l'accrétion des indices.

Parmi ces identifiants non directs, nous distinguons ceux qui sont les plus sensibles à l'anonymisation, c'est-à-dire ceux qui apportent une information plus importante et plus spécifique, de ceux qui sont plus généraux :

- les noms de famille comme *Dupond* ou *Durand* vs. *Eshkol* ou *Kanaan*
- les noms de métiers *professeur de physique* vs *enseignant*

On peut supposer qu'une entité dénommante (*nom rare, handicap, caractéristique particulière*) ou une série de ces entités (*nom, métier, lieu de travail, loisir, etc.*) est associée à un individu particulier dans la mémoire à l'aide d'un certain lien dénominatif qui sera réactivé lors de leur apparition dans le discours. C'est grâce aux facteurs contextuels, c'est-à-dire grâce aux connaissances que l'utilisateur du corpus maîtrise concernant le locuteur ou la personne mentionné dans le discours de celui-ci, que l'identification peut se faire. Ainsi, pour repérer le référent il faut tenir compte de divers types de connaissances (linguistiques, métalinguistiques et encyclopédiques).

Le contexte joue ainsi un rôle primordial dans le processus d'identification car il permet de réduire le champ d'application de ces éléments à un seul individu, de le distinguer des autres référents possibles. En premier lieu, on peut mentionner le contexte immédiat (gauche et/ou droite) d'une entité. Le nom de lieu n'aura pas de grand intérêt employé seul, mais employé avec des verbes comme *venir de, travailler à* ou avec des noms comme *collège, hôpital, etc.* il devient identifiant du lieu de travail, d'études ou d'origine de la personne. Ce contexte peut être aussi défini par la question posée. On sort ce faisant des limites de l'énoncé pour étudier un contexte plus large. Le nom de lieu, par exemple, n'est pas signifiant s'il est utilisé pour répondre à la question : *où parle-t-on le mieux le français ?*, par contre il devient un identifiant dans les réponses aux questions concernant les origines du locuteur, ou dans les énoncés décrivant l'emploi du locuteur, pour autant que celui-ci indique le lieu de son travail. De la même manière, les réponses aux questions sur les émissions de télévision, par exemple, n'apportent pas d'information personnelle et les noms de personnes qui apparaissent n'ont pas à être pris en compte. L'entité dénommante repérée doit être étiquetée selon le contexte. Dans la phrase *je travaille au collège de Saint-Jean-de-Bray*, l'entité *collège de Saint-Jean-de-Bray* ne réfère plus seulement à un établissement scolaire en général, c'est une référence à un lieu de travail. Les questions posées peuvent donc jouer un rôle important dans la catégorisation adéquate d'une entité repérée. Enfin, il est nécessaire de prendre en compte le contexte socioculturel de l'époque. Ainsi, les destinations de vacances peuvent être prises en compte car en 1968 très peu de gens voyageaient à l'étranger :

⁸ La répétition de la préposition *dans* dans cet énoncé fait partie des disfluences de l'oral et est transcrite comme telle dans les fichiers de transcription

j'ai vu aussi pas mal de pays j'ai vu l'Espagne le Portugal euh l'Allemagne l'Italie la Sicile qui m'a beaucoup plu également le la Yougoslavie

nous sommes allés par bateau jusqu'au Cap Nord et retour euh par euh jusqu'à la frontière finlandaise jusqu'à Oslo après nous avons vu euh la Suède et le Danemark Canaries et retour par Dakar

Certaines informations doivent être parfois déduites du contexte comme dans l'exemple suivant:

BV: y a longtemps que vous êtes à Orléans ?

MS530: euh oui euh vingt-deux ans

BV: ça fait euh vous êtes née à Orléans

MS530: oui

La prise en compte du contexte est donc nécessaire pour repérer et catégoriser les entités dénommantes.

L'analyse empirique du corpus a permis d'établir la typologie des entités dénommantes et de choisir une méthodologie de leur repérage automatique.

3.2.2 Repérage automatique

Pour repérer et annoter les entités dénommantes, nous avons choisi l'approche symbolique en surface permettant de construire les grammaires locales selon le contexte en utilisant le système CasSys (Friburger, 2002) intégré à la plateforme Unitex (Paumier, 2003) et adapté au corpus oral. L'adaptation de l'outil à ESLO1 s'est faite sur plusieurs niveaux. Tout d'abord, le corpus a été segmenté en tours de parole en fonction des balises Transcriber⁹. Les cascades de CasSys ont été enrichies de nouvelles grammaires locales avec des dictionnaires et des graphes spécifiques pour atteindre l'objectif final. En tenant compte de la nature du corpus traité, les différentes disfluences de l'oral ont été prises aussi en compte comme par exemple dans *je m'appelle euh Patrick Mallon*¹⁰.

Nous avons procédé en deux étapes. Tout d'abord, nous lançons des cascades de transducteurs qui repèrent et annotent les entités nommées (EN). Ensuite, une autre série de cascades appliquée à ce corpus annoté, identifie les entités dénommantes (DE):

- 1^{ère} étape : `<EN type="loc.admi">Pithiviers</EN>`
- 2^{ème} étape : `<DE type="pers.speaker">moi je suis <DE type="identity.origin">native de <EN type="loc.admi">Pithiviers</EN></DE></DE>`

Les règles de reconnaissance des entités dénommantes tiennent compte du contexte gauche et/ou droit de l'entité et du contexte plus large : la nature de la question posée si elle porte sur l'identité du locuteur.

Suite à l'analyse manuelle du corpus et à partir de la typologie de la campagne d'évaluation Ester2 (campagne d'évaluation des systèmes de transcription enrichie d'émissions radiophoniques)¹¹ nous avons élaboré le jeu d'étiquettes pour annoter des entités dénommantes. Ce jeu d'étiquettes a été guidé par la nature et le contenu du corpus. Le sujet sur qui porte l'information est annoté en premier lieu. Nous distinguons entre le locuteur (*pers.speaker*) et les autres membres de sa famille (*pers.spouse*, *pers.parent*, *pers.child*). Nous précisons ensuite la nature de cette information : l'identité, le travail, les études, l'engagement associatif ou syndical, les vacances :

- *il est parti à Paris =>*
`<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 je suis monsieur Gabrion je suis ingénieur chimiste=>*
`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>`

⁹ Méthode recommandée par (Dister, 2007)

¹⁰ L'annotation automatique des entités nommées et dénommantes a été décrite dans (Maurel et al., 2011, Eshkol et al., 2012).

¹¹ http://www.afcp-parole.org/ester/docs/Conventions_EN_ESTER2_v01.pdf

L'annotation a été réalisée sur 112 fichiers Transcriber (35,75 Mo). L'évaluation des résultats a été effectuée sur 9 fichiers (6 fichiers ont été réservés pour les tests). Les entités dénommantes ont été reconnues avec la précision estimée à 94,2 % et le rappel de 84,4 % (Maurel et al., 2009).

3.2.3 Etape finale : validation et anonymisation

La multitude d'éléments personnels, biographiques annotés dans notre corpus soulève une autre question concernant leur pertinence. Sont-ils tous identifiants et donc susceptibles de révéler l'identité du locuteur ou de ses proches ? Est-ce que le fait de repérer tous ces éléments dans le discours suffit à l'anonymiser ?

Le travail du repérage automatique des entités dénommantes a montré que l'anonymisation ne peut pas s'arrêter à cette étape. Tous les éléments annotés ne nécessitent pas d'être anonymisés. Pour décider si telle ou telle information identifie le locuteur, il est nécessaire de procéder à une validation manuelle. C'est pourquoi l'étape finale du module développé consiste à filtrer les entités dénommantes. Les entités qui renvoient vers les informations personnelles concernant le locuteur et sa famille, et qui peuvent éventuellement permettre sa reconnaissance, sont validées manuellement et remplacées par un hyperonyme : *NPERS* pour un nom de personne, *NLIEU* pour un nom de lieu, *NPROF* pour un nom de profession, etc.

3.2.4 Bilan

Plusieurs objectifs ont guidé notre travail :

- étudier les éléments identifiants,
- essayer de les repérer automatiquement,
- élaborer une procédure d'anonymisation qui prend en compte les résultats de ce module automatique pour l'anonymisation finale du corpus ESLO.

Le travail effectué a montré que si l'on veut anonymiser le corpus oral, il ne suffit pas de reconnaître les entités nommées car d'autres éléments peuvent permettre l'identification du locuteur. Nous avons analysé ces *entités dénommantes* et nous avons réussi de les repérer automatiquement avec une bonne précision et un bon rappel.

Malgré ce succès, plusieurs difficultés et limites du travail ont été mises en évidence. En premier lieu, la présence de multiples disfluences (hésitations, répétitions, reformulations, amorces, etc.) qui peuvent intervenir à différents moments dans le discours et qui rendent cette tâche difficile comme dans *je m'appelle euh Patrick Mallon*.

Le corpus comprend aussi des informations difficiles à catégoriser comme par exemple :

- *mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville*
- *je suis scout de France le jeudi soir où j'anime un un atelier photos*

La catégorie des informations occasionnelles comme ci-dessus semble "imprévisible" en raison de son manque d'homogénéité et elle ne peut être repérée que par l'analyse manuelle du corpus.

Ensuite, les informations de nature personnelle varient d'une manière non homogène dans le corpus. Chaque type d'information peut être présenté à travers un groupe nominal ainsi qu'avec des expressions plus étendues. "Ce passage se manifeste par l'ajout de propriétés supplémentaires à la classe présentée par le groupe nominal minimal, ce qui diminue l'extension de la classe et rapproche le groupe d'une référence plus individualisante" (Eshkol, 2010 : 258). Ainsi, le locuteur peut décrire son métier de manières diverses :

- *je suis enseignant dans l'école publique*
- *je suis maître auxiliaire*
- *j'enseigne des mathématiques modernes des mathématiques classiques de la chimie et de la technologie*

On ne peut donc jamais atteindre une liste exhaustive de toutes les reformulations possibles.

Il faut enfin mentionner le contexte extralinguistique qui est très difficile à prendre en compte automatiquement. Il s'agit du contexte socioculturel de l'époque (le cas, par exemple, d'ESLO1 constitué dans les années 60-70) et la déduction de nouvelles informations.

Le travail a montré aussi une distinction qui est nécessaire de faire entre les informations personnelles sur le locuteur qui sont repérables automatiquement et les informations identifiantes qui demandent une intervention humaine pour être validées. Nous reviendrons sur cette distinction qui permet également montrer mieux les limites de l'automatisation de l'anonymisation dans la section 5.

Enfin, le processus de l'anonymisation mise en place dans notre expérience demande, dans sa phase finale, un travail manuel de relecture de chaque transcription pour remplacer les informations personnelles annotées par leur hyperonyme dans le cas où elles sont jugées identifiantes. Le temps de relecture induit étant trop important par rapport à un traitement manuel intégré à la phase de transcription, l'équipe a préféré ne pas utiliser de détection automatique.

4 Anonymisation actuelle d'ESLO

L'anonymisation actuelle dans ESLO est semi-automatique et porte sur deux types d'objets : les données et les métadonnées. Dans la chaîne de traitement du corpus, la phase d'anonymisation est fractionnée ; elle précède la phase de transcription, coïncide avec elle et lui succède.

4.1 Métadonnées

La première étape d'anonymisation concerne les métadonnées. Elle repose sur le codage des noms propres, l'extraction de l'adresse (ces deux informations étant conservées mais non disponibles) et la correction des données de géolocalisation.

Le codage des noms propres des locuteurs est l'action la plus classique et attendue dans une procédure d'anonymisation. Deux types de codages sont mis en place : les codes aléatoires qui sont générés par notre application suite à la création d'une fiche en saisissant les métadonnées du locuteur (ex : DC738, Figure 1 : *Fiche du locuteur*), et les codes répondant à un plan de nommage (ex : DC738FEM).

Fiche locuteur	
Identifiant locuteur : CT418	
Anonyme:	OUI
Année de naissance:	2005
Tranche d'âge:	5/10
Lieu de naissance:	
Sexe:	Homme
Niveau d'études:	Primaire
Commentaire:	Elève en CE1
Age de fin d'études:	
Catégorie Professionnelle (INSEE):	Autres personnes sans activité professionnelle
Profession en termes propres:	
Langue(s):	
Commentaire niveau langue:	
Situation de famille:	Célibataire
Année d'arrivée:	
Domicile:	Olivet
Nombre d'enfants:	
Information sur les enfants:	
Remarques diverses:	
Fiche modifiée par:	lkanaan
Enregistrements et transcriptions:	<ul style="list-style-type: none"> ■ Enregistrement ESLO2_REPAS_1258 <ul style="list-style-type: none"> • Transcription ESLO2_REPAS_1258_A • Transcription ESLO2_REPAS_1258_B • Transcription ESLO2_REPAS_1258_C

Figure 1 : Fiche du locuteur

Le plan de nommage ESLO propose des combinaisons permettant de marquer des relations ou des catégories. En effet, pour marquer les relations certains locuteurs possèdent des codes construits sur le code aléatoire d'un autre locuteur (BA725FIL). Un autre type de codes du plan de nommage repose sur les numéros des enregistrements et permet de marquer une catégorie (653CLI, 653VEN, 308STAN dans un appel téléphonique). De plus, les locuteurs non

identifiés, non attendus dans un enregistrement et pour lesquels aucune information n'est repérable ni fournie sont aussi codés en lien avec l'enregistrement (452INC).

Toujours au niveau de métadonnées, nous intervenons au niveau du lieu de l'enregistrement. L'anonymisation s'effectue à travers la transformation de l'adresse en coordonnées GPS de manière à délimiter un périmètre correspondant au "pâté de maisons".

Notons que les données nominatives (nom, adresse, numéro de téléphone) des témoins sont conservées dans une BDD physiquement indépendante, conformément aux recommandations de la CNIL (Baude et al., 2006). Ces informations sont recueillies par le chercheur dans ESLO1 et renseignées par les locuteurs eux-mêmes dans ESLO2 qui complètent un "Formulaire témoin". Pour ESLO2, et qui n'est pas le cas pour ESLO1, les témoins signent un "Formulaire de consentement" concernant : leur participation au projet scientifique, la conservation (mais non diffusion) de leurs informations personnelles dans le seul but d'être recontactés, l'utilisation des enregistrements et de leurs transcriptions pour la recherche et pour la diffusion. Il leur est précisé que enregistrement et transcription seront rendus anonymes (nom remplacé par un code et son brouillé pour la séquence concernée). Il s'agit donc là d'un consentement éclairé (ibid.) c'est-à-dire que le document signé décrit clairement la manière dont les données seront traitées et les différents types d'utilisation dont elles feront l'objet. En ce sens, les témoins sont informés des "risques" de leur participation au projet.

La conservation des données nominatives s'est révélée particulièrement intéressante dans le cas d'ESLO puisque cela a permis, quarante ans après la première enquête (ESLO1) de retrouver des témoins et de mener des entretiens de nouveau avec eux. Un module d'ESLO2 qui offre des possibilités riches et intéressantes pour des recherches diachroniques a ainsi été constitué.

4.2 Transcriptions

La question du codage du locuteur ayant déjà été traitée au niveau des métadonnées, les codes sont repris dans les transcriptions. Il reste alors à traiter les données identifiantes contenues dans les énoncés. La deuxième intervention se situe donc au niveau de la transcription. Il est demandé aux transpositeurs de remplacer par l'hyperonyme NPERS les noms de personnes (Figure 2 : *Anonymisation dans la transcription*) et par NANON les autres segments du discours permettant d'identifier un locuteur – *i.e.* les entités dénommantes telles qu'elles ont été définies plus haut –, ou encore des propos "sensibles". Ces opérations sont par la suite vérifiées par un chercheur avant qu'elles ne soient traitées au niveau de l'enregistrement, ce dernier constituant la troisième étape d'anonymisation. Il s'agit donc d'une action faite par un humain sans aucun repérage automatique car elle ne représente pas une charge de travail significative de par son intégration à la phase de transcription.

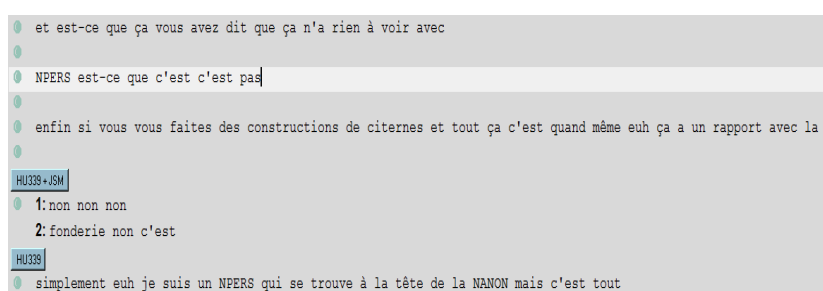


Figure 2 : Anonymisation dans la transcription

Le fichier de transcription contient néanmoins des données nominatives. L'en-tête du fichier XML indique en effet le nom de la personne qui a effectué la transcription. Cette information est conservée avec l'accord de l'auteur de la transcription afin de permettre la traçabilité de l'activité scientifique et de respecter la propriété intellectuelle du travail fourni (cette démarche correspond à l'une des recommandations de la charte Ethique et Big Data¹²).

¹² <http://wiki.ethique-big-data.org/index.php?title=Accueil>

4.3 Son

Cette dernière étape est effectuée avec le logiciel Praat et grâce à un script réalisé par Daniel Hirst (LPL, Aix-en-Provence) qui permet de modifier automatiquement les segments du fichier son à partir d'un repérage dans les annotations. Les segments sonores concernés sont modifiés afin que l'on ne puisse pas comprendre ce qui est prononcé tout en conservant certaines caractéristiques comme, notamment, la courbe intonative. Après un repérage temporel des NPERS et des NANON dans la transcription – un travail manuel qui est devenu automatique grâce à une application développée par Flora Badin (LLL, Orléans) –, l'isolement des segments concernés est effectuée sous Praat après la création d'un textgrid. Le code "buzz" marqué dans chacun des segments délimités permet au script d'opérer à l'intérieur des balises temporelles et de brouiller le signal (Figure 3 : *Anonymisation du son*).

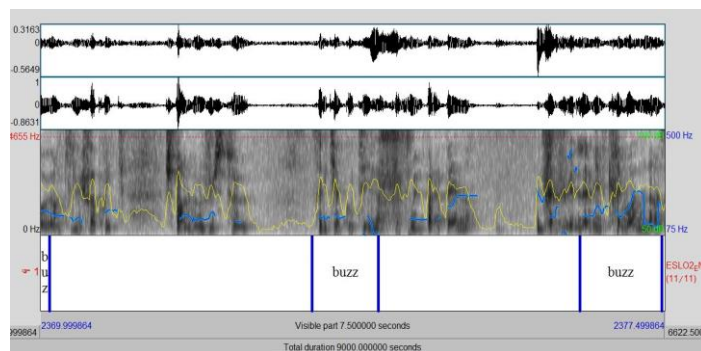


Figure 3 : Anonymisation du son

5 Conclusion

L'anonymisation est une étape souvent nécessaire et toujours délicate d'un traitement du corpus oral. Toutes les informations personnelles n'identifient pas le locuteur mais c'est une combinaison de certaines entre elles qui permettent sa reconnaissance. Le travail sur l'anonymisation du corpus ESLO a permis ainsi de distinguer entre deux types d'informations : les informations personnelles qui sont repérables automatiquement et les informations identifiantes qui permettent de reconnaître le locuteur et qui doivent être validées manuellement. Cette distinction entre les deux types d'information permet donc de définir les limites du traitement automatique de la tâche de l'anonymisation du corpus oral.

On a présenté, dans cet article, un test effectué sur un corpus "sacrifié" afin de pouvoir définir et décrire les éléments permettant l'identification du locuteur. Nous avons appelé ces éléments les *entités dénommantes* par opposition aux *entités nommées* qui ne répondent pas aux mêmes critères. Ainsi, le traitement automatique des entités dénommantes ne peut pas se satisfaire du repérage des entités nommées. D'une part, les entités dénommantes dépassent par leur diversité les entités nommées (*noms de métiers, de maladies, etc.*) et, d'autre part, les entités nommées repérées doivent fournir des informations sur le locuteur, ce qui n'est pas toujours le cas (*Sarkozy, Cotillard*). Le module d'anonymisation automatique testé utilise la méthode symbolique fondée sur les cascades des transducteurs. Le corpus est annoté d'abord en entités nommées et ensuite en entités dénommantes. Ensuite les informations sensibles qui risquent dévoiler l'identité du locuteur sont remplacées manuellement par l'hyperonyme. Si nous reprenons la distinction entre les éléments personnels et identifiants définies ci-dessus, le module développé annoté les éléments personnels qui sont remplacés par un humain dans le cas où ils deviennent identifiants.

Plusieurs résultats ont été obtenus grâce à cette expérience et ses apports sont multiples. Premièrement, la nature des entités dénommantes a été étudiée. Cette étude a permis de développer la typologie de ces entités que nous avons utilisée pour leur annotation. Deuxièmement, le module développé ne s'arrête pas à la reconnaissance des entités nommées mais va plus loin en annotant d'autres éléments qui permettent au même titre que les entités nommées d'identifier le locuteur. Enfin, nous avons travaillé sur le corpus oral ce qui distingue aussi notre recherche par rapport aux autres dans le domaine du TAL qui traitent plutôt les corpus écrits.

L'expérience effectuée a montré également les limites du traitement automatique de l'anonymisation du corpus oral : les informations occasionnelles, imprévisibles et non homogènes impossibles à être repérées d'une manière systématique automatiquement ; la prise en compte du contexte extralinguistique ; la variation infinie des reformulations possibles à l'oral. Ces limites ainsi que la nécessité de procéder à la lecture manuelle des fichiers annotés pour valider l'information et remplacer celle permettant d'identifier le locuteur par l'hyperonyme ont été la raison pour laquelle l'équipe s'est

ournée aujourd'hui vers le processus moins automatisé qui porte sur les métadonnées, les transcriptions et le son. L'anonymisation des métadonnées repose sur le codage des noms propres, l'extraction de l'adresse et la correction des données de géolocalisation. Les données nominatives des témoins sont conservées dans une BDD physiquement indépendante, conformément aux recommandations de la CNIL (Baudé et al., 2006). Les fichiers de transcriptions sont anonymisés manuellement par le transcripteur pendant le processus-même, il remplace les éléments identifiants par leurs hyperonymes. Les fichiers son sont traités par le script qui met un buzz dans les segments annotés en gardant leur courbe intonative.

En conclusion, les outils du TAL sont utiles et efficaces pour le repérage des informations personnelles. Cette tâche réussit dans les corpus écrits où ces informations sont présentées d'une manière plus au moins homogènes. La tâche se complique pour le discours oral. Si l'on réussit efficacement de repérer automatiquement les éléments personnels, il est nécessaire ensuite de choisir parmi ces éléments ceux qui permettent l'identification du locuteur. Ce filtrage ne peut se faire que manuellement à l'heure actuel. La perspective d'une automatisation de cette détection est un défi que les recherches en TAL pourraient relever.

Références

- AMBLARD M., FORT K. (2014). Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais. Actes de *TALN2014*, Marseille, France.
- BAUDE O. (2006) *Corpus oraux : guide des bonnes pratiques*. CNRS-Editions et Presses universitaires d'Orléans, 2006.
- BAUDE O., DUGUA C. (2011). (Re)faire le corpus d'Orléans quarante ans après : quoi de neuf, linguiste ?, vol. 10, *Corpus, Varia*.
- CHARAUDEAU P., MAINGUENEAU D. (2002). *Dictionnaire d'analyse du discours*. Paris, Éditions du Seuil.
- DUBOIS J. (1973). *Dictionnaire de linguistique*. Paris, Larousse.
- ESHKOL I. (2010a). Entrer dans l'anonymat. Etude des "entités dénommantes" dans un corpus oral. *Eigennamen in der gesprochenen Sprache*, 245-266.
- ESHKOL I., MAUREL D., FRIBURGER N. (2010b). Eslo: from transcription to speakers' personal information annotation. Actes de *Seventh Language Resources and Evaluation Conference (LREC 2010)*, Malte.
- ESHKOL-TARAVELLA I., BAUDE O., MAUREL D., HRIBA L., DUGUA C., TELLIER I., (2012). Un grand corpus oral "disponible" : le corpus d'Orléans 1968-2012. *Ressources linguistiques libres, TAL*. 52 : 3, 17-46.
- FRIBURGER N. (2002). *Reconnaissance automatique des noms propres ; application à la classification automatique de textes journalistiques*, Thèse de doctorat d'informatique, Université François Rabelais Tours.
- GROUIN C, ZWEIGENBAUM P. (2011). Une approche à plusieurs étapes pour anonymiser des documents médicaux. *RSTIRIA*, 25 :4, 525-549
- HAMON P. (1977). Pour un statut sémiologique du personnage. *Poétique du récit*. Barthes R. et alii, Points-Seuil, Paris.
- MAUREL D., FRIBURGER N., ESHKOL I. (2009). Who are you, you who speak? Transducer cascades for information retrieval. Actes de *4th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*. Poznań, Poland, 220-223.
- MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL-TARAVELLA I., NOUVEL D., (2011). Cascades de transducteurs autour de la reconnaissance des entités nommées. *Varia TAL*, 52 :1, 69-96.
- MEYSTRE S., FRIEDLIN B S., SHUYING S., SAMORE M. (2010). Automatic de-identification of textual documents in the electronic health record: a review of recent research. *BMC Medical Research Methodology* 10.70.
- PAUMIER S. (2003). *De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique*. Thèse de Doctorat en Informatique, Université de Marne-la-Vallée.

RAAJ N. (2012). *Automated Tool for Anonymization of Patient Records*. Report. MSc Computing and Management, Imperial College, London¹³.

TRAN M., MAUREL D. (2006). Prolexbase : Un dictionnaire relationnel multilingue de noms propres. *TAL*, 47 : 3, 115-139.

TVEIT A., EDSBERG O., BROX RØST T., FAXVAAG A., NYTRØ Ø., NORDGÅRD T., THORSEN RANANG T., GRIMSMO A., (2004). Anonymization of General Practitioner Medical Records. *Second HelsIT Conference at the Healthcare Informatics*, Trondheim.

UZUNER O., LUO Y., SZOLOVITS P., (2007). *Evaluating the state-of-the-art in automatic de-identification*. J Am Med Inform Assoc 14.550-63.

¹³ <http://www.comp.leeds.ac.uk/mscproj/reports/1112/raaj.pdf>