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

Iris Eshkol-Taravella, Olivier Baude, Denis Maurel, Layal Kanaan-Caillol

Recherche des indices permettant une identification: l'anonymisation des transcriptions du corpus ESLO

Iris Eshkol-Taravella[1] Olivier Baude[1] Denis Maurel[2] Layal Kanaan-Caillol[1]

(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, denis.maurel@univ-tours.fr, layal.kanaan@univ-orleans.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 constituant un « faisceau d'indices » qui, dans un certain contexte, contribue à l'identification. Ces indices dépassent par leur diversité et leur hétérogénéité les entités nommées. Nous décrivons ensuite une expérimentation du repérage automatique de ce faisceau d'indices dans les transcriptions.

Abstract.

Recognizing clues leading to identification: anonymizing the transcriptions of the ESLO speech corpus
This article tackles the question of oral corpus anonymization in preparation for its diffusion on the Web. We first analyze elements constituting a « clues set » which contribute to the identification. Those clues exceed named entities by their diversity and heterogeneity. Then we describe an experiment based on a module of automatic recognition of its clues in the transcriptions.

Mots-clés : anonymisation, anonymisation automatique, corpus oral, faisceau d'indices, données personnelles, identification
Keywords: anonymisation, automatic anonymisation, oral corpus, indications set, personal data, identification

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 BNC[1], le Russian National Corpus[2] ou encore le National Corpus of Polish[3], ou en français, CLAPI[4], PFC[5], CRFP[6], 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 ORTOLANG). Pour diffuser ces corpus, les questions juridiques dont celle de leur anonymisation se sont avérées primordiales.

---

3 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/crfp

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 obèrerait 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, 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 a déclenché une réflexion sur les éléments permettant l'identification du locuteur et de toute autre personne mentionné dans le discours de celui-ci et sur leur repérage automatique.

2 Identification à travers un « faisceau d'indices »

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. Notre objectif est d'étudier, dans le discours oral, des éléments qui permettent de distinguer le sujet parlant et la personne dont on parle des autres et, par conséquent, de les reconnaître. Nous avons appelé l'ensemble de ces éléments un « faisceau d'indices ». On peut identifier la personne en la dénommant, c'est-à-dire en la mentionnant par son nom, ou en la décrivant, c'est-à-dire en représentant certains de ses traits, voire de ses activités. Anonymiser le corpus consiste dans le repérage de ces indices et leur substitution par un hyperonyme ou un élément à référents multiples. Ces indices peuvent être de nature lexicale et sémantique très variée : des entités nommées (section 2.1), d'une part, mais aussi des groupes nominaux fondés sur le nom commun désignant les traits caractéristiques ou des groupes verbaux, énoncés décrivant les habitudes et les activités sociales de la personne (section 2.2).

2.1 Entités nommées identifiantes

Traditionnellement la tâche d'anonymisation s'arrête au repérage des entités nommées (noms de personnes, lieux, organisations, âges, etc.) (Ehrmann, 2008, Nadeau, Sekine, 2004). C'est le cas de plusieurs travaux en TAL dans le domaine médical (Meyster et al., 2010, Tweit et al., 2004, Raaj, 2012, Uzuner et al., 2007, Grouin, Zweigenbaum, 2011) qui portent sur les documents écrits (rapports, dossiers médicaux, etc.) et où les informations à anonymiser sont assez homogènes et souvent regroupées dans un endroit précis. Un de ses outils disponible gratuitement est Medina

(Medical Information Anonymization[7]). Il 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 entités nommées sont effectivement les candidates idéales à l'anonymisation car par définition, « on appelle entité nommée tout expression linguistique qui réfère à une entité unique du modèle de manière autonome dans le corpus » (Ehrmann, 2011), les entités nommées sont donc censées renvoyer vers un référent unique. Or il s'avère que tous les entités nommées mentionnées dans le discours ne sont pas identifiantes du point de vue de l'anonymisation.

- Noms de personnes

Le premier cas des entités nommées est les noms de personne. C'est un indice fort pour l'anonymisation car le rôle même de ces noms est de nommer, c'est-à-dire d'indiquer le référent d'une personne mentionnée. Cela est prouvé par l'anonymisation manuelle des transcriptions (section 3). Dans un test sur 112 fichiers d'ESLO1, 168 éléments ont été masqués. Parmi eux, les 159 éléments ont été remplacés par hyperonyme *NPERS.*

Les noms de personne jouent aussi le rôle primordial dans les travaux sur la détection de l'identité du locuteur dans les journaux télévisés (Charhad, Quénot, 2005). Les auteurs reconnaissent le locuteur grâce aux patrons qui détectent la personne qui se présente, qui vient de parler (le locuteur remercie, par exemple, l'orateur précédent en l'appelant par son nom) et la personne qui va parler (le locuteur passe la parole à un autre orateur en le nommant).

Est-ce que tous les noms évoqués dans le discours pointent vers l'identité de la personne ? Les noms de famille ou prénoms rares comme *Eshkol* ou *Kanaan,* dans le cadre de la ville comme Orléans, peuvent éventuellement identifier la personne. Pourtant, dans le cadre de l'anonymisation, les noms de personnalités *Sarkozy, Cotillard*, étant les noms publics ne doivent pas être anonymisés. Les noms de famille très répandus qui renvoient à un nombre élevé de référents comme *Dupont, Durand* ne donnent aucune information sur la personne et ne permettent pas à eux-seuls de l'identifier.

- Fonction

Selon le guide d'annotation des entités nommés Quaero (Rosset et al. 2011), la fonction (*func*) comprend les métiers, les fonctions et les rôles sociaux de la personne.

Nommer la personne par sa fonction *maire d'Orléans, directeur du collège de Saint-Jean de Braye* est un acte qui peut renvoyer à un référent unique. On est de nouveau en présence d'un indice fort. Ce n'est pourtant pas le cas d'un nom de métier. La mention dans le discours du métier *enseignant-chercheur* ne veut rien dire sur l'identité de la personne, mais dans le contexte *je suis enseignant-chercheur* il devient un indice de l'identification.

- Autres entités nommées

Les autres entités nommées présentes dans le discours doivent aussi avoir un lien avec le locuteur ou la personne qu'il mentionne dans son discours pour devenir un indice d'identification. Ce lien est souvent exprimé dans le discours même par le contexte gauche/droite de l'entité ou par la question posée dans le cadre de l'entretien ce qui est le cas du corpus étudié. Ainsi, le nom de lieu tout seul ne dit rien sur la personne mais employé avec la précision *je travaille à ….* ou *mon père est originaire de…,* il devient un indice, une information personnelle. Il le devient aussi dans les réponses à des questions portant sur l'identité de la personne comme *où travaille votre femme ? vous êtes originaire d'où ?*. Les mêmes observations peuvent se faire pour d'autres types d'entités nommées : les dates ou encore les noms d'organisations.

De manière concomitante, il y a dans le discours d'autres éléments qui ne font pas partie des entités nommées mais qui peuvent renvoyer vers l'identité de la personne. Ce phénomène a été déjà mentionné dans (Amblard, Fort, 2014) où les auteurs présentent entre autres le processus d'anonymisation automatique du discours transcrit de schizophrènes. Ils notent l'insuffisance du simple repérage à l'aide de scripts Python des mots commençant par une majuscule dans les extrais du corpus où des sujets relatent un événement « s'inscrivant dans une temporalité et une géographie particulière » et la présence d'autres indices selon lesquelles on peut identifier le locuteur ou ses proches. Cette affirmation se manifeste à travers les chiffres provenant des résultats de l'expérience de l'annotation automatique du sous-corpus ESLO1 en indices permettant l'identification éventuelle du locuteur (section 4). Dans 112 fichiers de transcription d'ESLO1 annotés en entités nommées et en indices, candidats à l'anonymisation, on retrouve 13 909 entités nommées au total et seulement 1 038 autres indices. Ces chiffres confirment que, d'une part, toutes les entités nommées ne

[7] http://medina.limsi.fr/

renvoient pas vers le locuteur et que, d'autre part, il existe dans le corpus d'autres éléments qui peuvent permettre l'identification éventuelle de la personne.

2.2 Autres indices

Si l'on veut anonymiser efficacement le discours, on ne peut pas s'arrêter aux entités nommées car d'autres indices peuvent renvoyer vers le locuteur ou vers la personne dont il parle. Observons les exemples tirés du corpus ESLO1 :

- *j'ai une maladie du foie ça m'a même occasionné une petite scoliose déformation légère de la colonne vertébrale.*
- *mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville*[8]
- *je suis scout de France le jeudi soir où j'anime un un atelier photos*[9]

Cette catégorie des indices est large. Elle inclut des éléments assez hétérogènes désignant les différentes informations personnelles sur la personne : événements, activités sociales, loisirs, maladies, handicap, etc. qui peuvent au même titre que le travail, la famille donner les informations sur le locuteur ou la personne dont on parle.

Ainsi, le « faisceau d'indices » inclut les entités nommées identifiantes, mais peut contenir aussi d'autres éléments qui permettent l'identification soit directement, soit, par combinaison au sein de ce faisceau : la personne est patron d'un bar au moment d'enregistrement, et avant elle travaillait dans l'aviation militaire. Le processus d'identification est progressif, il se construit au fur et à mesure de l'accroissement des indices. On peut supposer qu'un indice identifiant ou une série de ces indices 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 mentionnée dans le discours de celui-ci, que l'identification peut se faire.

Les parties qui suivent sont consacrées à la description de l'anonymisation du corpus ESLO. Le processus d'anonymisation du corpus consiste à repérer un faisceau d'indices qui permet d'identifier le sujet parlant ou toute autre personne mentionnée dans le discours. Dans le processus actuel de l'anonymisation des transcriptions d'ESLO (section 3), ces indices sont repérés manuellement par les transcripteurs. Pour aider ce processus, une expérimentation de l'automatisation de ce processus a été tentée (section 4). Nous finirons par quelques perspectives liées à l'intégration du module automatique développé dans le processus actuel (section 5).

3 Procédure semi-automatique d'anonymisation des transcriptions dans le corpus des ESLO

Nous allons voir dans cette partie la procédure suivie par les gestionnaires du corpus des ESLO afin de procéder à l'anonymisation des données du corpus.

Du point de vue juridique, le corpus ESLO1, a posé deux problèmes (Baude[10]). 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 de 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.

L'anonymisation actuelle dans ESLO est semi-automatique et porte sur deux types d'objets : les données (sons et transcriptions) et les métadonnées. Dans la chaîne de traitement du corpus, la phase d'anonymisation est fractionnée ;

[8] L'emploi des déterminants *un le* dans cet énoncé fait partie des disfluences de l'oral (autocorrection) et est transcrite comme telle dans les fichiers de transcription.

[9] idem.

[10] http://eslo.huma-num.fr/index.php/pagemethodologie?id=69

elle précède la phase de transcription, coïncide avec elle et lui succède. Nous nous contentons, dans cet article, de décrire la phase de d'anonymisation des transcriptions[11].

Le codage des noms propres des locuteurs est l'action la plus classique et attendue dans une procédure d'anonymisation. Les transcriptions comportent des informations issues des métadonnées, à savoir l'identifiant du locuteur. Dans la procédure ESLO, des codes aléatoires sont générés par l'application suite à la création d'une fiche en saisissant les métadonnées du locuteur (ex : DC738). Ces codes sont repris dans les transcriptions. Le traitement des données identifiantes contenues dans les énoncés est effectué au niveau-même de la transcription. Il est demandé aux transcripteurs de remplacer par l'hyperonyme *NPERS* les noms de personnes (Figure 1) et par *NANON*[12] les autres segments du discours permettant d'identifier un locuteur.

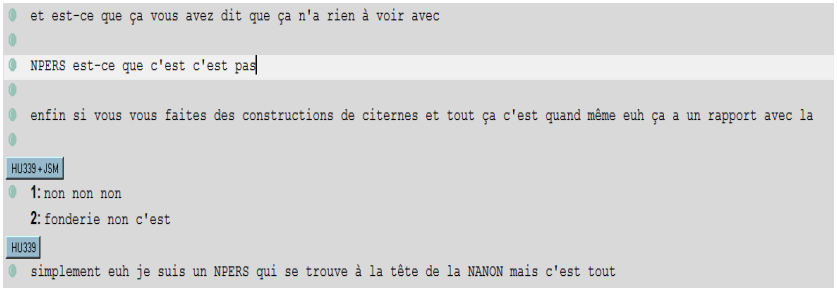

Figure 1 : Anonymisation dans la transcription

L'anonymisation manuelle des fichiers de transcription a soulevé la question d'automatisation de ce processus grâce aux outils du TAL. L'expérience a été menée afin de repérer automatiquement les indices permettant l'identification éventuelle du locuteur ou de toute autre personne mentionnée dans les transcriptions.

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

L'expérimentation décrite dans cette partie a été effectuée en collaboration avec le laboratoire LI (Laboratoire Informatique) de l'université de Tours. Le test portait sur un sous-corpus d'ESLO1 (112 entretiens face-à-face) contenant de nombreuses données personnelles sur le locuteur car il s'agissait d'un questionnaire concernant la vie des témoins : « *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.

4.1 Repérage automatique

Lorsqu'on parle de l'anonymisation automatique tout le monde s'accorde sur la nécessité de repérer les entités nommées. Comme nous l'avons évoqué, ces éléments font très souvent partie, à juste titre, des indices recherchés. C'est la raison pour laquelle, pour faire une expérimentation d'anonymisation automatique des transcriptions d'ESLO1, il a été décidé de partir de l'outil permettant d'identifier les entités nommées. La collaboration avec le LI de Tours a permis d'exploiter le système CasSys développé dans le cadre de la thèse par Nathalie Friburger (Friburger, 2002) et intégré à la plate-forme Unitex (Paumier, 2003). Il s'agit d'une approche symbolique en surface permettant de construire les grammaires locales selon le contexte sous forme des cascades de transducteurs qui repèrent et annotent les entités nommées dans le discours médiatique.

Le système CasSys a été adapté au corpus traité. Tout d'abord, le corpus a été segmenté en tours de parole en fonction des balises Transcriber[13]. Les cascades de CasSys ont été ensuite enrichies de nouvelles grammaires locales avec des dictionnaires et des graphes spécifiques pour reconnaître dans les transcriptions de l'oral en plus des entités nommées

[11] Pour une présentation de la procédure : Baude et Dugua Guide d'Anonymisation (en ligne http://eslo.huma-num.fr/index.php/pagemethodologie?id=69)

[12] Nom anonymisé

[13] Méthode recommandée par (Dister, 2007)

d'autres indices. Enfin, en tenant compte de la nature du corpus, les différentes disfluences de l'oral ont été prises aussi en compte comme par exemple dans *je m'appelle euh Patrick Mallon*[14].

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 indices-candidats à l'anonymisation (*DE*[15]). Dans cet exemple, l'entité nommée *Pithiviers* a été reconnue au cours de la première étape, cette entité devient identifiante au cours de la deuxième étape car elle se trouve dans le contexte indiquant son lien avec le locuteur *moi je suis native de Pithiviers* :

– 1ère étape : <***EN*** *type="loc.admi">***Pithiviers**</**EN**>
– 2ème étape : <***DE*** *type="pers.speaker">***moi je suis** <***DE*** *type="identity.origin">***native de** <*EN type="loc.admi">Pithiviers</EN></**DE**></**DE**>

La Figure 1 présente le graphe permettant la reconnaissance d'une origine géographique : ce graphe appelle un sous-graphe (*NELoc*) qui reconnaît un toponyme identifié par la cascade des entités nommées.

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)[16] nous avons élaboré le jeu d'étiquettes pour annoter des indices. L'enquête correspond essentiellement à des questions concernant la personne interrogée et sa famille : origine, âge, naissance, arrivée à Orléans, travail et même syndicat. Pour cela nous avons défini une typologie avec trois types principaux, personne, identité et travail, eux même divisés en sous-types, comme présenté dans la Figure *3*. 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 syndicale, 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**>*
– *je peux vous demander quel est votre syndicat ? </Turn> <Turn speaker="spk5" startTime="5071.106" endTime="5072.466"> <Sync time="5071.106"/> <Sync time="5071.22"/> oui c'est la <**DE** type="pers.speaker"> <**DE** type="involvement.tradeunion"> <**ENT** type="org"> **CGT** </**ENT**></**DE**></**DE**>*
– *de ce fait <**DE** type="pers.speaker">**nous sommes allés euh** <**ENT** type="time.date.rel"> **trois jours** </**ENT**> <**DE** type="trip.work"> **à** <**ENT** type="loc.admi"> **Londres** </**ENT**> </**DE**> <**ENT** type="time.date.rel"> **trois jours** </**ENT**> <**DE** type="trip.work"> **à** <**ENT** type="loc.admi"> **Vienne** </**ENT**> </**DE**> **nous avons été** <**ENT** type="time.date.rel"> **trois jours** </**ENT**> <**DE** type="trip.work"> **en** <**ENT** type="loc.admi"> **Hongrie** </**ENT**></DE>*

On voit dans ces exemples, que l'entité nommée *monsieur Gabrion* est bien annotée en tant qu'indice car elle se trouve dans un contexte qui concerne le locuteur *je suis monsieur Gabrion*. C'est le cas pour une autre entité nommé, le nom du syndicat *CGT*, car elle se trouve dans la réponse à la question concernant le locuteur. Les cascades annotent également les syntagmes fondés sur les noms communs comme le métier *ingénieur chimiste* ou le domaine d'activités professionnelles *dans les assurances*. A cela s'ajoute l'annotation d'autres indices comme les vacances *nous sommes allés trois jours à Londres* ou des autres actions *il est parti de Paris*.

La reconnaissance des indices est fondée sur le contexte qui joue 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'indice. Le nom de lieu

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

[15] Le terme que nous avons utilisé à l'époque de cette expérimentation pour désigner les indices annotés est celui d'« entité dénommante » (Eshkol, 2010). Le travail complémentaire entrepris depuis sur la définition de cette notion nous amène maintenant à préférer la terminologie de « faisceau d'indices » telle que nous l'avons présentée dans la section 2.

[16] http://www.afcp-parole.org/ester/docs/Conventions_EN_ESTER2_v01.pdf

n'aura pas 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. L'indice repéré doit être étiqueté aussi selon le contexte. Dans la phrase *je travaille au collège de Saint-Jean-de-Braye*, l'entité nommée *collège de Saint-Jean-de-Braye* ne réfère plus seulement à un établissement scolaire en général, c'est une référence à un lieu de travail du locuteur. 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 indice 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. Les questions posées peuvent donc jouer un rôle important dans la catégorisation adéquate d'un indice repéré. 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 peu de gens à Orléans voyageaient à l'étranger, c'est le cas du dernier exemple ci-dessus.

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 indices ont été reconnus avec la précision estimée à 94,2 % et le rappel de 84,4 % (Maurel et al., 2009).

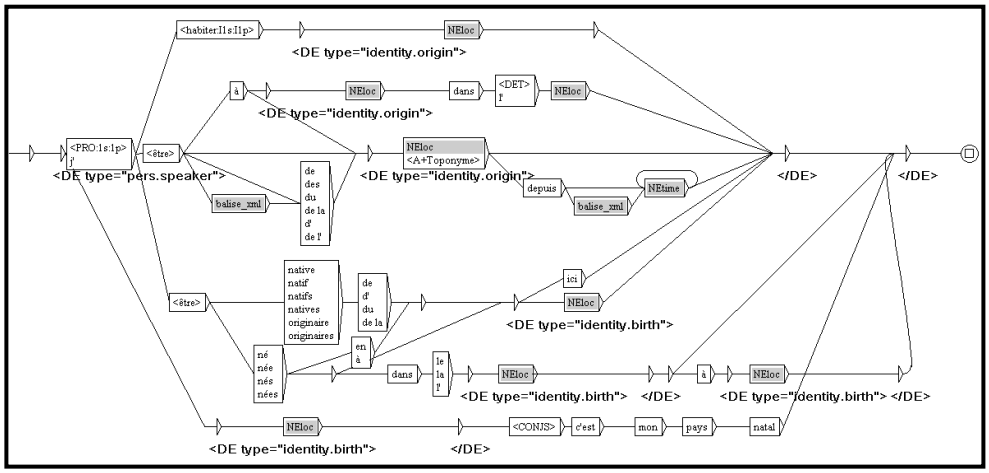

Figure 2 : Un graphe pour l'origine géographique

4.2 Difficultés rencontrées

Malgré ce succès, plusieurs difficultés 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 comme dans *je m'appelle euh Patrick Mallon* rendent la tâche difficile. Le graphe contenant la liste de disfluents possibles a été créé ce qui a permis de résoudre ce problème dans beaucoup de cas.

Ensuite, dans le discours oral, les informations apparaissent d'une manière parfois aléatoire. Ainsi, des informations sur le témoin ne se trouvent pas nécessairement dans la partie questionnaire, mais peuvent surgir à des endroits inattendus. Par exemple, la description de la recette de l'omelette peut être l'occasion de glisser son origine géographique :

– *enfin on assaisonne sel poivre euh <DE type="pers.speaker"> nous en <DE type="identity.origin"> <ENT type="loc.admi"> Lorraine </ENT></DE></DE> on on découpe des petits des petits morceaux de lards qu'on fait frire avant*

Certaines informations doivent être aussi 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

Une autre difficulté provient de la variation linguistique. 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. 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*

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)

Figure 3 : Typologie des indices

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

Enfin, le corpus peut comprendre 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*

Cette catégorie du faisceau d'indices comprenant les actions, les événements, les activités sociales du locuteur semble « imprévisibles » en raison de son manque d'homogénéité.

Malgré toutes ces difficultés, les indices peuvent être reconnus automatiquement avec une bonne précision et un bon rappel. La première catégorie, les entités nommées identifiantes, est bien reconnue par le module développé. La deuxième indiquant les actions, événements, activités sociales du locuteur est identifiée mais pas d'une manière exhaustive.

Cependant la multitude d'éléments personnels annotée dans le corpus soulève une autre question concernant leur pertinence. Tous les éléments annotés ne nécessitent pas d'être anonymisés. Actuellement, la décision d'anonymiser un tel ou tel indice, ne peut se faire que manuellement par un humain. C'est seulement l'humain qui peut décider aujourd'hui, souvent d'une manière assez subjective, lequel des éléments personnels renvoie le plus vers le locuteur ou ses proches et doit donc être masqué. Le principe respecté est de garder le maximum d'informations pour pouvoir permettre l'analyse du corpus. Ainsi, dans les 112 fichiers contenant 1 038 indices annotés, seulement 168 ont été remplacés par leur hyperonyme (159 *NPERS* et 9 *NANOM*[17]).

5 Conclusion et perspectives

Le travail effectué a montré que si l'on veut anonymiser un corpus d'enquêtes sociolinguistiques, il ne suffit pas de reconnaître les noms propres et les autres entités nommées car d'une part, d'autres éléments peuvent aussi permettre l'identification du locuteur ou de la personne mentionnée dans le discours notamment quand il existe une combinaison de ces éléments au sein du corpus et, d'autre part, tous les entités nommées ne sont pas sensibles à l'anonymisation et ont besoin d'un contexte pour devenir identifiantes.

Le module développé pour le repérage automatique des indices-candidats à l'identification potentielle de la personne tient compte des spécificités de l'oral (la présence de disfluences, l'absence des signes de ponctuation dans les transcriptions, la segmentation en tours de parole) et permet d'obtenir des résultats encourageants.

La difficulté majeure de l'anonymisation automatique des discours transcrits de l'oral est que toutes les informations personnelles n'identifient pas la personne mais qu'en revanche une combinaison de certaines d'entre elles constituent un faisceau qui dans un certain contexte, le plus souvent extralinguistique, contribuent à l'identification. Actuellement la décision sur la pertinence de masquer certains éléments du faisceau ne peut se faire que par une intervention humaine. Pour aider cette validation manuelle, la distinction pourrait se faire entre les éléments les plus sensibles à l'anonymisation, c'est-à-dire ceux qui apportent une information plus importante et plus spécifique, et ceux qui sont plus généraux. Ainsi, pour distinguer entre les noms de famille rares comme *Eshkol* ou *Kanaan* et très répandues *Dupond* ou *Durand,* on pourrait s'appuyer, dans le cas du corpus des ESLO, sur une information concernant la fréquence d'un nom propre, éventuellement pondérée par des critères géographiques. De la même manière, le locuteur peut désigner son métier par un seul mot *enseignant* ou en précisant *professeur de physique.* Ce passage d'un seul nom à un groupe nominal plus étendu grâce aux modifieurs « 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). Ce processus concerne n'importe quelle caractéristique (maladie, loisir, etc.). On pourrait ainsi attribuer plus de poids à ces éléments sensibles à l'anonymisation ce qui diminuerait le nombre des indices candidats à l'anonymisation et de cette manière aiderait la validation manuelle.

Dans le faisceau d'indices, la deuxième catégorie comprenant les éléments ne faisant pas partie des entités nommées comme actions, événements, activités sociales, doit être approfondie d'autant plus qu'elle permet d'apporter une information sur le profil sociologique du locuteur. Le module développé tient compte de ces indices mais leur liste n'est pas exhaustive. Pour un travail futur, nous envisageons d'étudier avec précision dans le corpus ESLO, tous les éléments anonymisés par une procédure manuelle afin d'affiner la typologie du faisceau d'indices.

[17] L'étiquette *NANOM* signifie le nom anonymisé.

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.

CHARHAD M., QUENOT G. (2005). Approche par patrons linguistiques pour la détection automatique du locuteur : application à l'indexation par le contenu des journaux télévisés. *Compression et Représentation des Signaux Audiovisuels (CORESA'05)*, Rennes.

DUBOIS J. (1973). *Dictionnaire de linguistique*. Paris, Larousse.

EHRMANN M. (2008). Les entités nommées, de la linguistique au TAL : statut théorique et méthodes de désambiguïsation. Thèse de doctorat, Université Paris 7 - Centre de recherche Xerox, Grenoble (XRCE).

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. *RSTI-RIA*, 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.

NADEAU N., SEKINE S. (2009). *A survey of named entity recognition and classification*, Satoshi Sekine and Elisabete Ranchhod, ed., John Benjamins publishing company, 3-28.

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, Imperiel College, London[18].

[18] http://www.comp.leeds.ac.uk/mscproj/reports/1112/raaj.pdf

ROSSET S., GROUIN C., ZWEIGENBAUM P. (2011). *Entitées nommées structurées : guide d'annotation Quaero.* Notes et documents LIMSI N°2011-04.

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.