« Procédure d'anonymisation et traitement automatique : l'expérience d'ESLO »

Eshkol-Taravella I.[(1)], Kanaan-Caillol L.[(1)], Baude O.[(1)], Dugua C.[(1)], Maurel D.[(2)]

(1) LLL UMR7270, Université d'Orléans
(2) LI, Université de Tours

ESLO (Enquête Sociolinguistique à Orléans), projet du Laboratoire Ligérien de Linguistique, est 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 fait suite à l'initiative d'universitaires britanniques qui, entre 1968 et 1974, ont collecté à Orléans des documents sonores (entretiens, appels téléphoniques, réunions, repas, divers) avec une visée didactique : l'enseignement du français langue étrangère dans le système public d'éducation anglais. En partant des acquis d'ESLO 1, une nouvelle enquête a été mise en chantier en 2008 par le LLL : ESLO 2. Il s'agit, à quarante années de distance, de constituer un corpus comparable dans le produit attendu et dans les modalités de la collecte : l'objectif a été fixé à 400 heures environ de documents sonores répondant à une approche variationniste et conçus à travers des programmes spécifiques. 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 aspects juridiques qui a abouti à la définition et à la mise en place d'une phase d'anonymisation dans la chaîne de traitement.

Bien que l'on parle souvent d'anonymisation, la question légale concerne principalement l'assurance qu'il sera impossible d'identifier des personnes. Juridiquement l'anonymisation, qui n'est obligatoire qu'en substitut du consentement, sert à qualifier l'opération par laquelle se trouve supprimé dans un ensemble de données, recueilli auprès d'un individu ou d'un groupe, tout élément qui permettrait l'identification de ces derniers. (Baude et al. 2006). Dans le cas des corpus oraux il faut prendre en compte, d'une part les moyens permettant l'identification dans un contexte de développement d'outils de fouille de masses de données, et d'autre part, l'ambigüité sur le statut juridique de la voix comme données identifiantes.

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) 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. Le langage oral est différent de l'écrit. Anonymiser le discours d'une manière automatique est une tâche difficile qui pose des problèmes supplémentaires au TAL. Cela a récemment été constaté dans (Amblard et Fort, 2014)

Dans le cadre d'ANR Variling nous avons effectué, entre 2009 et 2011, une expérience d'anonymisation automatique d'un sous-corpus transcrit d'ESLO1. En premier lieu, nous avons étudié des éléments permettant l'identification du locuteur que nous avons appelés *entités dénommantes* (Eshkol 2010a).

Il est difficile de rendre compte du mécanisme cognitif en jeu dans le processus de reconnaissance d'un individu. 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. Il faut prendre en considération les facteurs contextuels qui entourent l'énonciation de ces entités. C'est le contexte qui permettra de réduire le champ d'application de ces éléments à un seul porteur, de le distinguer des autres référents possibles. Le locuteur peut être identifié directement (nom *Kanaan ou Eshkol*, métier *général, maire* ou caractéristiques rares *mon père a fondé le plus grand cabinet ophtalmologique de la ville*) ou à travers la combinaison de divers identifiants (*le locuteur est patron d'un bar au moment de l'enregistrement et avant il travaillait dans l'aviation militaire*). On voit dans ces exemples que la reconnaissance des entités nommées ne suffit pas à repérer toutes les informations concernant le sujet parlant et que certaines informations décrivant le locuteur ou sa famille ne sont pas repérables automatiquement.

L'expérience de l'annotation automatique s'est portée sur un sous-corpus d'ESLO1[1] : les 112 fichiers de transcription contenant les entretiens « face-à-face » riches des informations personnelles sur les locuteurs interviewés. Les questions posées portent beaucoup sur l'identité du locuteur et des membres de sa famille : « *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. Ces questions nous ont aidés à établir une typologie des éléments à annoter. 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*). Les questions posées ont été aussi utilisées pour désambiguïser certaines informations. Ainsi, la mention du lieu est importante dans la réponse à la question portant sur les origines du locuteur que dans la question sur la région en France où on parle le mieux français.

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). Ce choix nous a permis de ne pas développer un nouvel outil mais d'adapter un outil existant à nos données. L'annotation s'est faite en deux étapes. Nous avons repéré et annoté, en premier lieu, les entités nommées. Nous avons recherché ensuite les entités dénommantes. L'annotation des entités nommées et dénommantes a été décrite dans (Eshkol-Taravella et al. 2012, Eshkol et al. 2010b, Maurel et al. 2011, 2009).

Comme il a été mentionné ci-dessus, la reconnaissance des entités dénommantes prend en compte le contexte d'emploi de ces éléments mais aussi la nature orale des données.

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.

et qu'est-ce que vous faites comme travail ?
<Turn speaker="spk1" startTime="40.394" endTime="43.041">
<DE type="pers.speaker">je suis<DE type="work.occupation"> contrôleur divisionnaire<DE type="work.occupation"> au <ENT type="org.com"> PTT </ENT></DE></DE></DE>

depuis combien de temps habitez-vous <ENT type="loc.admi">Orléans</ENT> ?
<DE type="pers.speaker"><DE type="identity.origin">
<Turn speaker="spk1" startTime="6.754" endTime="10.88">
oh ça fait <ENT type="time.date.rel">neuf ans</ENT> depuis dix neuf cent soixante</DE></DE>

Pour l'anonymisation finale, les entités repérées et étiquetées sont validées manuellement. Celles qui identifient directement le locuteur sont remplacées par un hyperonyme.

Pour évaluer la cascade des entités dénommantes, nous avons utilisé les mêmes enregistrements que ceux des entités nommées[2]. Parmi 112 fichiers annotés, 6 fichiers ont été réservés pour les tests et 9 pour l'évaluation. Sur les 77 entités dénommantes, nous en avons reconnu 69, nous avons fait 4 erreurs et oublié 12 entités. Notre précision est donc de 94,2 % et notre rappel de 84,4 %.

Cette expérience sur un sous-corpus d'ESLO1 a permis, d'une part, de définir, de repérer et d'annoter les éléments permettant l'identification du locuteur mais aussi de montrer les limites du TAL dans l'automatisation de ce processus sur les données orales. Après l'évaluation du coût de l'automatisation de l'anonymisation, l'équipe a décidé de procéder à une anonymisation semi-automatique.

L'anonymisation actuelle dans ESLO est donc 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.

La première étape d'anonymisation concerne les métadonnées. Elle correspond au codage des locuteurs. 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, et les codes répondant à un plan de nommage (ex : BA725).

[1] Au moment de cette étude les transcriptions ESLO2 n'étaient pas disponibles.
[2] Une évaluation de cette cascade est présentée dans (Maurel et al., 2009)

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 (cf. 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.
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 contenues dans les énoncés. La deuxième intervention se situe donc au niveau de la transcription. Il est demandé aux transcripteurs de remplacer par l'hyperonyme NPERS les noms de personnes 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.
Cette dernière étape est effectuée avec le logiciel Praat et grâce à un script réalisé par Daniel Hirst (LPL, Aix-en-Provence). 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) –, une isolation 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.
La spécificité de ce script réside dans le fait qu'il garde un aspect de la parole notamment la courbe intonative.
L'anonymisation est une étape souvent nécessaire et toujours délicate d'un traitement du corpus oral. Le test effectué sur un sous-corpus a permis de définir, décrire et repérer automatiquement les éléments permettant l'identification du locuteur que nous avons appelé *entités dénommantes*. Il a démontré que le traitement automatique des entités dénommantes ne peut pas se satisfaire des techniques linguistiques éprouvées de repérage des entités nommées : d'une part, les entités dénommantes dépassent par leur diversité les entités nommées et, d'autre part, les entités nommées repérées doivent fournir des informations sur le locuteur, ce qui n'est pas toujours le cas. Cette expérience a aussi montré les limites du traitement automatique dans le cadre de l'anonymisation du corpus oral où certaines informations apparaissant d'une manière non régulière ne peuvent pas être

repérées automatiquement. La procédure d'anonymisation semi-automatique mise en place par ESLO nous parait la plus fiable actuellement.

Bibliographie

AMBLARD M., Fort K. (2014). Étude quantitative des disfluences dans le discours de schizophrènes : automatiser pour limiter les biais, *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*.

ESHKOL I., (2010a). « Entrer dans l'anonymat. Etude des « entités dénommantes » dans un corpus oral », in *Eigennamen in der gesprochenen Sprache*, p.245-266.

ESHKOL I., MAUREL D., FRIBURGER N. (2010b). « Eslo: from transcription to speakers' personal information annotation », *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 », in *Ressources linguistiques libres, TAL*. Volume 52, n° 3, p. 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.

MAUREL D., FRIBURGER N., ESHKOL I. (2009). « 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, 220-223.

MAUREL D., FRIBURGER N., ANTOINE J.-Y., ESHKOL-TARAVELLA I., NOUVEL D., (2011). « Cascades de trans-ducteurs autour de la reconnaissance des entités nommées ». *Varia TAL*, vol. 52, n° 1, p. 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, Imperiel College, London[3].

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.

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