Interpréter le contexte dans un corpus oral : fonctions et limites du traitement automatique des données linguistiques

Iris Eshkol

▶ **To cite this version:**

Iris Eshkol. Interpréter le contexte dans un corpus oral : fonctions et limites du traitement automatique des données linguistiques. Le contexte - Rencontres interdisciplinaires sur les systèmes complexes naturels et artificiels, Chemins de tr@verse, pp.67-80, 2015. hal-01174731

HAL Id: hal-01174731
https://hal.science/hal-01174731

Submitted on 9 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Iris ESHKOL
Université d'Orléans, LLL
E-mail : iris.eshkol@univ-orleans.fr

Interpréter le contexte dans un corpus oral : fonctions et limites du traitement automatique des données linguistiques

1. Introduction

Les années 1990 ont constitué un tournant dans l'évolution du traitement automatique du langage (TAL) avec la constitution et l'exploitation de corpus qui ont provoqué une redéfinition des objectifs et un renouvellement des méthodes de la linguistique (cf. Habert et Nazarenko, 1997). L'apparition d'Internet a offert un accès à des données textuelles massives et les outils élaborés par le TAL, par exemple les lemmatiseurs ou les analyseurs syntaxiques, en ont permis une exploitation immédiate. En règle générale, l'analyse linguistique introduit une certaine séparation entre les données textuelles et leur situation de production. Le processus de décontextualisation discrimine les phénomènes langagiers qui sont totalement imbriqués avec la situation dans laquelle ils ont été produits. Il convient de distinguer entre ce qui supporte d'être décrit de manière systématique indépendamment du contexte (certaines relations morphologiques ou syntaxiques) et ce qui requiert l'élucidation des conditions de réalisation, en particulier pour les marques sémantiques ou pragmatiques.

Dans ce cadre général, l'analyse d'un corpus oral diffère de celle requise par un corpus écrit du fait que l'analyse est tenue d'intégrer les paramètres d'une communication in praesentia (rôle de l'intonation et des composants prosodiques, du statut des locuteurs, etc.) ancrée par des marques énonciatives (deixis, usage des temps…) dans le contexte. Dans un discours oral, le sens est l'objet d'une constante renégociation et réévaluation qui s'élabore au fil de l'échange. Selon Corblin et Gardent (2005:15) "Pour l'analyse linguistique, le contexte pertinent est celui qui recouvre l'ensemble des éléments impliqués par l'activité langagière : les connaissances lexicales et encyclopédiques des participants, la situation physique d'énonciation (participants, lieu, temps) et le contexte linguistique, c'est-à-dire une trace du texte ou du dialogue précédant l'énoncé considéré." L'interprétation des participants comme celle des résultats est elle-même paramétrée, obéissant à des appréhensions fondées sur des intentions qui construisent la lecture. Dans cet article, nous montrerons à partir d'un exemple précis les conséquences d'une telle approche en nous fondant sur le programme de recherche ESLO (*Enquête SocioLinguistique à Orléans*) dont l'objectif est d'assurer la disponibilité d'un corpus de référence du français parlé aujourd'hui, exploitant les techniques informatiques. L'une des contraintes, qui sera illustrée dans cet article, concerne l'anonymisation des données afin de rendre publiques les informations tout en protégeant la vie privée des locuteurs.

Nous présenterons dans la deuxième partie un descriptif rapide et non exhaustif des étapes du traitement automatique des corpus qui sollicite le contexte pour résoudre certains problèmes linguistiques liés à l'ambiguïté. Nous montrerons comment le contexte, ou plutôt quels aspects du contexte sont pris en compte dans le traitement automatique du corpus oral et quelles sont les limites de ces techniques. Nous consacrerons la section 3 au projet d'anonymisation du corpus ESLO et nous évoquerons quelques-uns de nos choix techniques. Le projet est en cours de réalisation et donc pas encore finalisé. Cet article présente un certain

nombre de réflexions que soulèvent la tâche d'anonymisation et le rôle du contexte dans la méthodologie choisie.

2. Prise en compte du contexte dans le traitement automatique du corpus oral

La langue naturelle est riche en ambiguïtés (lexicale : un mot peut avoir plusieurs sens et/ou appartenir à différentes catégories ; syntaxique : un énoncé peut correspondre à plusieurs structures syntaxiques ; de portée : un opérateur sémantique peut prendre plusieurs valeurs anaphoriques et une expression anaphorique a souvent plusieurs antécédents possibles, etc.) qui multiplient les interprétations possibles d'un énoncé. Ces ambiguïtés ne sont pas relevées par l'auditeur qui généralement n'en a pas conscience car elles sont discriminées avant même que ne soit perçue leur ambivalence. Inversement, elles posent problème à chaque étape pour un système informatique de traitement. On montrera comment ces problèmes sont surmontés, en particulier par l'intégration de données extraites du contexte.

2.1. Analyse morpho-syntaxique

Une des premières étapes du traitement automatique du corpus consiste dans l'analyse morpho-syntaxique des énoncés transcrits. L'objectif premier est, à partir de la segmentation des chaînes graphiques, d'identifier les lexèmes par lemmatisation (c'est-à-dire de ramener l'ensemble des occurrences d'un mot donné à sa forme de base ou « lemme » :

spécial / spéciale / spéciales / spéciaux > spécial

Deux approches sont en concurrence. L'une, fondée sur les dictionnaires, consiste à décrire aussi exhaustivement que possible l'ensemble des mots du lexique ; l'autre repose sur des règles de calcul morphologique. Les mots sont alors corrélés à un jeu d'étiquettes (*tags*) qui récapitulent les informations morpho-syntaxiques (partie de discours, marques de genre, nombre, temps verbal, personne, etc.), ce qui permet de rechercher et/ou compter les différentes occurrences des mots ou des suites de mots et de déterminer leur environnement syntagmatique par l'établissement de concordanciers. L'étiquetage morpho-syntaxique considère les formes du corpus une à une, sans prendre en considération les contextes d'apparition. De ce fait, elle peut produire plus d'une interprétation pour une forme et engendrer des parcours très divergents à partir d'une séquence donnée :

avions
- le pluriel du nom *avion*
- la première personne du pluriel de l'imparfait de l'indicatif du verbe *avoir*

La prise en compte du contexte (appelé aussi « cotexte » pour le distinguer de la situation objective et le contenir dans les bornes du discours) permet de résoudre de nombreux problèmes liés à l'ambiguïté des mots polycatégoriels. Le contexte concerne l'entourage lexical proche (co-occurrence) droite/gauche de l'unité lexicale, c'est-à-dire les deux ou trois mots qui précèdent et qui suivent.

Dans un corpus oral, d'autres problèmes se présentent du fait que les transcriptions ne sont en général pas ponctuées afin d'éviter l'anticipation de l'interprétation (cf. Blanche-Benveniste et Jeanjean 1987). Au nombre des spécificités de l'oral, il faut prendre en compte les disfluences : répétitions, autocorrections, amorces de mots, pauses, etc. En accord avec

Blanche-Benveniste 2005, on considère que l'ensemble de ces phénomènes doit être intégré par l'analyse linguistique même s'ils créent des difficultés pour le traitement. Ils constituent un témoignage irremplaçable du « dynamisme de la composition sémantique dans la langue parlée » (p. 39). Il en va de même pour d'autres éléments :

des ponctuants comme *hein, bon, bien, quoi, voilà, comment dire*, *etc*

qui apparaissent avec une fréquence élevée dans les corpus oraux. Qu'on les désigne comme des *phatiques*, des *particules*, des *marqueurs discursifs* ou des *inserts*, ces formes figées ou invariables, peuvent constituer des énoncés à elles seules ou se manifester à différentes places d'un énoncé sans intégrer sa structure, (c'est-à-dire sans entrer en relation syntaxique avec un autre élément). Or, très souvent, ces unités lexicales sont porteurs d'ambiguïté. Soit *bien* :

- interjection, s'il est isolé
  (i) déjà il est gentil bien mais
  (ii) bien bien je vois ce que tu veux dire

- adverbe, lorsqu'il est modifieur d'un adjectif, d'un verbe ou d'un adverbe :
  (iii) très bien
  (iv) elle est bien gentille
  (v) ça c'est bien
  (vi) il s'est fait mal et bien mal.

Selon Dister 2007 « Toute forme peut potentiellement devenir une interjection. On assiste alors à une recatégorisation grammaticale […], le phénomène par lequel un mot ayant une classe grammaticale dans le lexique peut, en discours, changer de classe ». (p. 350) Est-il possible de distinguer les deux emplois d'un mot par l'analyse de son environnement ? Dans certains cas, comme dans le cas de *bon,* ce serait envisageable. On crée des grammaires ou des règles de désambiguïsation. On pourra émettre l'hypothèse, par exemple, que l'unité *bon* sera considérée comme :

- un adjectif lorsqu'il suit ou précède un nom (vii) *un bon élève* ;
- et dans les constructions attributives à définir en établissant une liste des verbes dits d'état comme *être, rester, demeurer, etc* : (viii) *il est bon* ;
- une interjection dans les autres cas.

Considérons un autre exemple

(ix) elle s'est opérée en 1966 voyez-vous

qui met en jeu l'expression *voyez-vous*. Celle-ci fonctionne comme un atome et elle peut permuter avec « d'accord, n'est-ce pas, etc. ». Afin de déterminer la nature des éléments, on est souvent obligé de recourir à des tests de substitution qu'il est difficile d'exploiter automatiquement alors que l'analyse linguistique du contexte de l'unité ambiguë ne s'avère pas toujours satisfaisante.

2.2. Analyse sémantique

Le contexte analysé peut être élargi au cours du traitement sémantique, dont l'objectif est l'analyse du sens/signification[1] d'un mot, d'un énoncé, d'un discours... Les significations d'un mot sont répertoriées dans les dictionnaires de langue et la polysémie requiert que soit déterminée la bijection entre l'énoncé et le lexique de référence. Comment procéder ? Une solution consiste à ne pas expliciter les significations en établissant des relations entre les significations des mots : relations d'hyperonymie, de synonymie, etc. Une autre solution recourt aux primitives appelées *sèmes* ou *traits sémantiques*. Depuis les années 90, l'intérêt se porte moins sur la décomposition en atomes de sens qu'au calcul concernant la variabilité du sens en contexte en se conformant à l'hypothèse distributionnelle de Harris (1976) selon laquelle le sens des mots est déterminé par la manière dont ils sont employés. Dans la sémantique de la phrase, on utilise souvent la notion de *prédicat* qu'on appréhende comme une unité lexicale à même d'opérer une sélection sur ses arguments. Discriminer les différentes significations d'une unité lexicale donnée revient à définir chacune de ses positions argumentales. De même, on va spécifier dans les dictionnaires quelle est la nature des différents arguments des prédicats verbaux tels que *manger* et *parler* :

manger (N1=animé, N2=nourriture)
parler (N1=humain, avec N2=humain)

si l'on veut avoir une estimation de la pertinence de sélection entre les interprétations « avocat fruit » et « avocat personne » dans les phrases :

Il mange un avocat.
Il parle avec un avocat.

On peut citer les travaux de Gross 1994 sur les classes d'objets au nombre de ceux qui ont constitué un apport important à cette tâche en français. Cependant des questions importantes restent en suspens : combien doit-on distinguer de significations différentes pour un prédicat donné ? A quel degré de finesse doit-on suspendre la décomposition ? Et surtout, comment parvenir à sélectionner « le » sens pertinent dans un énoncé ? La signification d'une unité dans un énoncé dépend des autres unités présentes dans le même énoncé (collocations et cooccurrences) et leurs relations syntaxiques. Ainsi, l'analyse du contexte dépasse l'environnement immédiat du mot étudié pour se déployer à l'échelle de la phrase et du discours.

2.3. Analyse pragmatique

La pragmatique prend pour objet le sens que les énoncés prennent dans le contexte d'énonciation. Le traitement pragmatique met l'accent sur les aspects suivants :

- mise en jeu de connaissances générales partagées ;
- mise en jeu de règles de bon usage dans la relation de communication ;
- prise en compte du contexte d'énonciation ;
- mécanismes de type inférentiels qui enrichissent ou dépassent l'information « apparemment » portée par le sens du texte.

[1] On distingue souvent les deux termes : *signification/sens linguistique/sens littéral* qui désigne le noyau de sens porté par un mot/énoncé/discours indépendamment de la situation dans laquelle il est émis et *sens* comme interprétation qui en est faite par un sujet donné dans un contexte d'énonciation donné.

Les difficultés sont nombreuses en particulier la prise en compte du contexte extralinguistique qui se révèle difficile à définir et parfois à formaliser[2]. Il s'établit une forme de collaboration entre, d'une part, l'information portée linguistiquement par le discours et, d'autre part, les connaissances générales du lecteur (ou de l'auditeur) incluant les nouvelles connaissances acquises au fil de la lecture (ou de l'échange verbal).
Le traitement automatique du langage apparaît comme une tâche fastidieuse, dont la résolution suppose le recours à des connaissances contextuelles impliquant un modèle linguistique et intégrant à ces procédures l'application visée et le domaine traité. Comment décrire et traiter cette articulation ? Le projet d'anonymisation du corpus ESLO soulève plusieurs questions en lien avec cette problématique et montre quels obstacles demeurent.

3. L'anonymisation ou le calcul d'identification en contexte

3.1. Présentation du corpus et du projet

L'enquête ESLO a été réalisée à la fin des années 1960. En 2005, le laboratoire CORAL (aujourd'hui LLL) a entrepris de mettre le corpus à la disposition de la communauté scientifique, dans le respect des pratiques et outils actuels, et a engagé une nouvelle enquête *ESLO2*. Réunis, *ESLO1* et *ESLO2* constitueront une collection de 700 heures d'enregistrement. L'objectif des ESLO est de construire un portrait sonore de la ville et de ses habitants de sorte que la diffusion des documents requiert une certaine prudence (informations personnelles, confidences, opinions, etc.). Par la force des choses les personnes enregistrées en 1968-69 n'ont pas donné leur autorisation pour une exploitation de leurs paroles telle qu'elle est prévue maintenant (diffusion en ligne notamment). Bien sûr il ne s'agit pas de rendre totalement impossible l'identification d'un locuteur (il faudrait alors brouiller la voix sur l'ensemble de l'enregistrement, ce qui rendrait toute analyse linguistique impossible) mais il convient de concevoir des corpus aux formes variables adaptables à différents contextes d'exploitation.

3.2. Réflexions sur le processus cognitif de la reconnaissance du locuteur

Les mécanismes cognitifs intervenant dans le processus de reconnaissance d'un individu sont complexes. Dans le cas présent, ils sont paramétrés en fonction des informations disponibles socialement puisque l'objectif est de masquer, auprès de l'utilisateur potentiel du corpus, l'identité du locuteur telle qu'elle pourrait être reconstituée à partir de l'écoute de l'enregistrement ou de la lecture de la transcription[3]. Enjalbert 2005 pose la question de la compréhension du texte lu et il interroge la façon dont le sens se construit à travers la lecture. La lecture d'un texte conduirait à la construction de représentations mentales relevant tout aussi bien de l'imaginaire, de l'affectif que des concepts abstraits. Le discours « déclenche et contraint l'interprétation pour exprimer qu'il ne peut fonctionner seul, mais plutôt en coopération avec un ensemble de connaissances et d'attentes du récepteur, dont il stimule et oriente l'activité ». (pp. 37-38) Ainsi, les deux sources d'information, le discours (la lecture

[2] Schank et Abelson 1977 ont cherché à produire une formalisation des connaissances sur le monde avec les notions de *script* et de *plan*.
[3] Cet article est centré exclusivement sur les transcriptions.

dans ce cas) et les connaissances extralinguistiques sont mobilisés conjointement pour construire ces représentations. Le processus de reconnaissance du locuteur à partir de la lecture d'une transcription mettrait en œuvre la construction de certaines représentations mentales liées à l'image que l'auditeur est conduit à se faire du locuteur. Anonymiser le corpus, c'est s'approprier le rôle du lecteur afin d'essayer de retrouver toutes ces représentations possibles. L'imprécision et l'infinité de la nature des connaissances nécessaires posent un problème majeur.

Comment l'esprit humain peut-il regrouper, rapprocher et croiser un ensemble d'expériences singulières dans sa définition d'un seul individu ? L'identification peut être réalisée à partir de connaissances que l'utilisateur extrait du corpus. Certaines propriétés caractéristiques :

nom rare, handicap, caractéristique particulière

ou une série corrélée de ces propriétés

nom, métier, lieu de travail, loisir, etc.

sont associées en mémoire et leur recollection est activée et enrichie à chaque apparition dans le discours. Cette réactivation peut se produire immédiatement, dès l'apparition d'un élément à forte valeur discriminatoire, et on parlera dans ce cas d'« identifiants directs »

dans ma classe quelquefois ils ne sont pas obéissants ...on m'appelle la maîtresse des fous, mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville

qui permettent, à eux seuls, de distinguer un individu. En d'autres occasions, le processus d'identification se produit progressivement, se construisant par accrétion d'indices. On appellera ces indices « identifiants non directs »

le locuteur est patron de café au moment de l'enregistrement et il travaillait auparavant dans l'aviation militaire

et leur présence seule ne permet pas de circonscrire qui est le locuteur. En revanche, par combinaison avec d'autres identifiants, ils peuvent renvoyer vers un individu singulier. Il s'agit, dans ce cas, d'attributs dont chacun est partagé par plusieurs individus mais dont la combinaison ne se rencontre que chez un seul. On peut dans ce cas comparer le processus d'identification avec celui de désambiguïsation en ce qu'ils répondent l'un et l'autre à deux principes : (i) minimalité, qui privilégie les interprétations faisant intervenir un nombre minimal d'objets et (ii) maximalité qui incline le choix, entre plusieurs interprétations possibles d'un énoncé, vers l'interprétation logique la plus forte en prenant en compte la compatibilité avec nos connaissances encyclopédiques.
Pour anonymiser le corpus, il faut en premier lieu repérer les éléments qui permettent l'identification du locuteur par un utilisateur du corpus n'ayant à sa disposition que des méta-données banalisées et le discours. Nous appelons ces éléments les « entités dénommantes » et elles mobilisent le contexte qui contribue à réduire le champ d'application de ces éléments spécifiants à un seul porteur, le distinguant des autres porteurs possibles.

3.3. Rôle du contexte dans le processus de l'anonymisation

Pour comprendre le processus d'identification et mener à bien l'anonymisation, il est indispensable de prendre en considération les facteurs contextuels.

En premier lieu, on s'attache à l'information linguistique contenue dans le texte et à la prise en compte de son contexte immédiat (gauche et/ou droite). Ainsi, un nom de lieu n'aura pas grand intérêt employé seul, mais avec des verbes tels que *venir de*, *travailler à* ou des noms tels que *collège, hôpital*, *etc.* il se transforme en identifiant d'un lieu de travail, d'études ou d'origine de la personne. De la même manière, les noms de personnes inconnues que mentionne le locuteur sont souvent introduits par les civilités *M./Mme* ou par le statut, *le recteur Antoine*, ce qui les différencie des noms de personnes connues du locuteur qui ne sont pas toujours précédés dans le discours par ces formes introductives. Le contexte gauche/droite de l'unité est largement utilisé pour la reconnaissance automatique des éléments par une approche linguistique fondée sur la description syntaxique et lexicale des syntagmes recherchés.

En deuxième lieu, on a recours à la structure du corpus. Dans la partie du corpus ESLO1 contenant des entretiens en face à face, le questionnaire s'ouvre sur des questions concernant le lieu *Depuis combien de temps habitez-vous Orléans ?*, *Qu'est-ce qui vous a amené à vivre à Orléans ?*, *Est-ce que vous vous plaisez à Orléans ?*, etc.), puis le travail et les loisirs du locuteur et des membres de sa famille, enfin sont abordés les sujets plus « généraux » concernant l'enseignement (*Qu'est-ce qu'on devrait apprendre surtout aux enfants à l'école ?*, *Dans quelles matières aimeriez-vous que vos enfants soient forts ?*, etc.), la politique (*Est-ce que, d'après vous, on fait assez pour les habitants d'Orléans ?*, *Que pensez-vous des événements de mai 68 ?*, etc.), la langue et les habitudes culturelles (*Un étranger veut venir en France pour apprendre le français. Dans quelle région est-ce qu'il doit aller d'après vous, dans quelle ville ?*, etc.). Le contexte sera défini, dans ce cas de figure, par la question posée, en élargissant la notion de contexte. 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 ?*, en revanche il fonctionne comme un identifiant dans les réponses aux questions concernant les origines du locuteur, ou dans les énoncés décrivant son activité professionnelle pour autant que celui-ci mentionne son lieu de 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 :

> *il a surpris beaucoup de personnes Edgar Faure certainement*
> *j'ai entendu parler euh Michel Couaste on dit ç- ça ? on dit on prononce comme ça Michel Couaste?*

On pourrait, en partant de là, opérer une distinction entre les questions sensibles dont les réponses peuvent contenir certaines informations personnelles :

> *Qu'est-ce que vous faites comme travail?*
> *Et votre femme, est-ce qu'elle travaille aussi? Pourquoi (pas)?*
> *Et vos enfants, que font-ils?/ métier?*
> *Qu'est-ce que vous faites de votre temps libre - soirées, week-end?*
> *Etc.*

et les questions neutres où la présence des entités nommées ne renvoie pas nécessairement au locuteur :

A votre avis, qu'est-ce qu'on devrait apprendre surtout aux enfants à l'école ?
Pour revenir à la ville d'Orléans, est-ce que, d'après vous, on fait assez pour les habitants d'Orléans?
Ecoutez-vous la radio ? nombre d'heures par semaine/jour ? Votre chaîne préférée ?
Etc.

Seules les questions sensibles requièrent d'être prises en compte si l'on veut distinguer l'information neutre de l'information personnelle. L'entité dénommante repérée doit être étiquetée selon le contexte. Dans la phrase

je suis au collège de Saint-Jean-de-Braye,

l'entité *collège de Saint-Jean-de-Braye* ne réfère plus seulement à un établissement scolaire en général, mais c'est une référence à un lieu de travail et le nombre de personnes correspondant à ce statut est limité à une cinquantaine. Les questions posées pourront donc jouer un rôle important dans la catégorisation adéquate d'une entité repérée.

Enfin, il est nécessaire de prendre en compte le contexte extra-linguistique. La situation d'énonciation est « circonstanciée » dans un lieu (Orléans) et à un moment donné (1968). On pourrait, par exemple, distinguer entre les métiers les moins répandus comme *vitrailliste, géomètre, etc.* et ceux plus courants comme *professeur, commerçant, etc.* Les listes de l'INSEE permettent de préciser ces critères. De même, les destinations de vacances peuvent être prises en compte, dans un second temps, car en 1968 très peu d'Orléanais voyageaient beaucoup à l'étranger :

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

La prise en compte du contexte extralinguistique montre les limites de l'automatisation. La difficulté réside dans la définition, l'inventaire et la description formelle de ce type de connaissances. La tâche de l'anonymisation du corpus oral met donc en jeu des compétences fondées sur :
- des connaissances linguistiques ;
- des connaissances sur le contexte dans lequel le corpus a été produit ;
- des connaissances encyclopédiques.

3.4. Méthodologie choisie

Pour procéder à l'anonymisation du corpus, nous avons préféré une approche « en surface » qui ne recourt pas à des représentations « profondes » de la sémantique du discours. Dans le cadre du projet VARILING[4], nous collaborons avec le Laboratoire Informatique (LI) de Tours (Denis Maurel, Marie-Aimée Gazeau) qui utilise le système CasSys (Friburger, 2002). CasSys

[4] VARILING Projet ANR 2006

exploite des cascades de transducteurs[5] en utilisant les outils fournis par Unitex (Paumier, 2003). Le processus d'anonymisation se déroule en plusieurs étapes.

En premier lieu, les cascades de transducteurs reconnaissent et annotent les entités nommées :

le ***<ENT type="org.pol">****ministère de l'Education Nationale****</ENT>***

Il s'agit des noms de lieux, de personnes, d'organisations, etc.[6] Des règles de grammaire créées décrivent le syntagme et son contexte immédiat en utilisant des marqueurs lexicaux (mots déclencheurs, comme le mot « ministère » dans cet exemple), des dictionnaires de noms propres et des dictionnaires spécifiques. Ces indices permettent de repérer un élément mais aussi de le catégoriser automatiquement[7].

Selon le même principe s'effectue la deuxième étape qui consiste à repérer et à annoter des entités dénommantes[8], c'est-à-dire des informations au sujet du locuteur. La cascade des transducteurs opère sur le corpus annoté :

<DE type="pers.speaker"> ***moi je suis*** *<DE type="identity.origin">* ***native de*** *<EN type="loc.admi">* ***Pithiviers****</EN> </DE> </DE>*

L'information recherchée concerne les origines, l'âge, le travail, les études, etc. du locuteur. Pour repérer ces entités, sont également prises en compte les questions posées au locuteur notamment sur son travail et sa date d'arrivée en ville.. Les entités dénommantes repérées ne sont que des candidats potentiels à l'anonymisation. C'est dans un contexte particulier qu'elles deviennent identifiants et doivent alors être masquées. Ce contexte se définit par une combinaison aléatoire de ces entités qu'il est complexe de calculer et qui est donc repérée automatiquement. D'où la troisième étape définitive du travail consistant en un filtrage de l'information annotée, ultime étape qui se fera manuellement.

4. Conclusion

En conclusion, le processus d'anonymisation consiste pour nous dans la recherche des indices qui permettent d'identifier le sujet parlant dans le discours. L'objectif est donc de localiser et de repérer ensuite dans le corpus certains types précis d'informations. Malheureusement, le processus ne peut se limiter au gommage des noms de personnes, ces éléments étant très peu présents dans notre corpus. Il reste en effet beaucoup d'autres marqueurs qui, employés seuls ou en combinaison avec d'autres, peuvent amener l'utilisateur à reconnaître le locuteur. Cette tâche n'est pas envisageable sans prendre en considération les facteurs contextuels divers comme l'information linguistique contenue dans le texte avec son contexte immédiat (gauche et/ou droite), la structure du corpus et évidemment le contexte extra-linguistique.

---

[5] Un transducteur est un automate à nombre fini d'états dont les transitions sont étiquetées par un couple de symboles : un symbole reconnu en entrée et un symbole produit en sortie. "Une cascade de transducteurs est une succession de transducteurs appliqués sur un texte, dans un ordre précis, pour le transformer ou en extraire des motifs. " (Friburger 2002 : 49). Chaque transducteur utilise les résultats des transducteurs précédents.

[6] Notre annotation est conforme à la typologie de la base de données Prolex (Maurel, 2008) adaptée dans le cadre de la campagne d'évaluation Ester (http://www.afcp-parole.org/ester/).

[7] Aujourd'hui le système reconnait les entités nommées du corpus avec la précision de 91,1% et le rappel de 87,5% (Maurel et al. 2009).

[8] Evaluer l'annotation des entités dénommantes est très difficile. Le système a été évalué sur un corpus de test de 7 fichiers avec la précision de 94,2% et le rappel de 84,4% (Maurel et al. 2009).

BLANCHE-BENVENISTE C., JEANJEAN C. (1987), *Le Français parlé. Transcription et édition*, Paris, Didier Érudition.

CORBLIN F., GARDENT C. (2005), « Contexte et interprétation », *Interpréter en contexte.* Lavoisier, Hermès, Paris, 15-28.

DISTER A. (2007), *De la transcription à l'étiquetage morphosyntaxique. Le cas de la banque de données textuelle orale VALIBEL,* Thèse de Doctorat, Université de Louvain.

ENJALBERT P. (2005), « Sémantique et traitement automatique du langage naturel : première approche. », *Sémantique et traitement automatique du langage naturel.* Lavoisier, Hermès, Paris, 27-52.

ESHKOL I. (2010), « Entrer dans l'anonymat. Etude des "entités dénommantes" dans un corpus oral. », *Actes du colloque NOMINA2007* (22-23 novembre 2007, Basel). Narr, Tübingen. (à paraître).

FRIBURGER N. (2002), *Reconnaissance automatique des noms propres. Application à la classification automatique de textes journalistiques.* Thèse de Doctorat, Université de Tours.

GROSS G. (1994), Classes d'objets et description des verbes. *Langages,* 115: 15-30.

HABERT B., NAZARENKO A. (1997), *Les linguistiques de corpus*, Paris, A. Colin.

HARRIS Z. (1976), *Notes du cours de syntaxe.* Paris, Le Seuil.

MAUREL D. (2008), Prolexbase. A multilingual relational lexical database of proper names, *Sixth language resources and evaluation conference* (*LREC 2008*), Marrakech, Maroc, 28-30 mai.

MAUREL D., FRIBURGER F., ESHKOL I. (2009), « Who are you, you who speak? Transducer cascades for information retrieval », *4th Language & Technology Conference*, Poznań, Poland.

PAUMIER S. (2003), *De la Reconnaissance de Formes Linguistiques à l'Analyse Syntaxique*, Thèse de Doctorat en Informatique, Université de Marne-la-Vallée.

SCHANK R., ABELSON R. (1977), *Scripts, plans, goals and understanding,* Lawrence Erlbaum, Hillsdale, NJ.