ANONYMISATION D’ESLO 1

Diffuser le Corpus d'Orléans selon les techniques actuelles, implique une démarche fondée sur de "bonnes pratiques" juridiques et éthiques. Ainsi, si pour des analyses scientifiques précises, le corpus brut reste le seul objet d'analyse possible, la diffusion par internet requiert un corpus anonymisé. L'objectif du projet de construire un portrait sonore de la ville et de ses habitants implique à un haut degré des propos dont la diffusion demande une extrême prudence (informations personnelles, confidences, avis exprimés, etc.). 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 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. (Olivier *et alii* 2006). 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). Notre objectif est de repérer des éléments dans le discours du locuteur permettant son identification par un éventuel utilisateur du corpus. Nous appelons ces éléments *entités dénommantes* (Eshkol 2007). La méthodologie que nous avons choisie peut être décrite en 5 étapes :

* repérage et balisage des entités nommées
* repérage et balisage des entités dénommantes
* extraction des entités dénommantes et leur analyse manuelle
* remplacement d’une entité dénommante qui identifie le plus le locuteur par son hyperonyme (NPERS, NLIEU, etc.)
* masquage du son

1. Le corpus a été annoté des entités nommées.

La cascade CasEN conçue pour reconnaître les entités nommées de corpus journalistiques (essentiellement *Le Monde*) [Friburger, 2002] a été reprise et adaptée au corpus d'Orléans. De plus, le balisage des entités nommées a été modifié pour correspondre à celui de la campagne Ester 2, à quelques ajouts près.

L'ensemble du corpus Eslo1 a ainsi été annoté et ces annotations ont été vérifiées manuellement ensuite.

moi je suis native de <ENT type="loc.admi"> Pithiviers </ENT> j'aime mieux <ENT type="loc.admi"> Orléans </ENT>

1. Le corpus a été annoté des entités denommantes

Nous avons créé une nouvelle cascade, *CasDen*, pour repérer automatiquement l’information sur le locuteur. Celle-ci passe non pas sur le texte originel, mais sur le texte balisé par la cascade CasEN. Nous avons défini de nouveaux types pour décrire la personne qui parle ou dont on parle (identité, famille, travail, syndicat, âge, origine…). Pour repérer ce type d’information nous nous sommes basés sur la forme des syntagmes mais aussi sur les questions posées :

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>

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>

1. Extraction et analyse des entités dénommantes

Au cours de l’étape suivante, les entités dénommantes ont été extraites et étudiées par les étudiants. Celles qui ont été considérées comme identifiant le mieu la personne ont été repérées et leur temps de prononciation a été marqué. Ce travail a été fait avec les feuilles de style XSLT :

<Sync time="443.775"/>

nous sommes nous avons la chance d'être dans une entreprise où <DE type="identité">monsieur Servier le président directeur général</DE> est particulièrement social

|  |  |
| --- | --- |
| <Sync time="443.775"/> | <DE type="identité">monsieur Servier le président directeur général</DE> |

<Sync time="2754.313"/>

avec <DE type="identité">mademoiselle Le</DE><ENT type="fonc.ari"> Prince</ENT> que vous connaissez peut-être

<Sync time="2757.185"/>

c'est la directrice du centre social de l'<ENT type="loc.admi">Argonne</ENT>

|  |  |
| --- | --- |
| <Sync time="2754.313"/> | <DE type="identité">mademoiselle Le</DE><ENT type="fonc.ari"> Prince</ENT> 🡪 mademoiselle Le Prince |

1. Remplacement

Les entités nommées ou dénommantes les plus identifiantes ont été remplacées dans les fichiers de transcription (.trs) par leur hyperonyme :

monsieur NPERS le président directeur général

L’anonymisation se fait sur la plus petite partie possible (ex : nom sans le prénom « Jean-Louis NPERS ». Les formes de remplacement :

NPERS nom de personne,

NLIEU nom de lieu,

NPROF profession,

1. Masquage du son

L’anonymisation du fichier son se fait avec l’aide du logiciel praat pour

- segmenter le fichier son (création d’intervalles correspondants exactement à la partie du signal qui doit être brouillée)

- annoter ces intervalles (dans une tier de praat on indique un code, en l’occurrence « buzz »puis traiter le signal à l’aide d’un script.

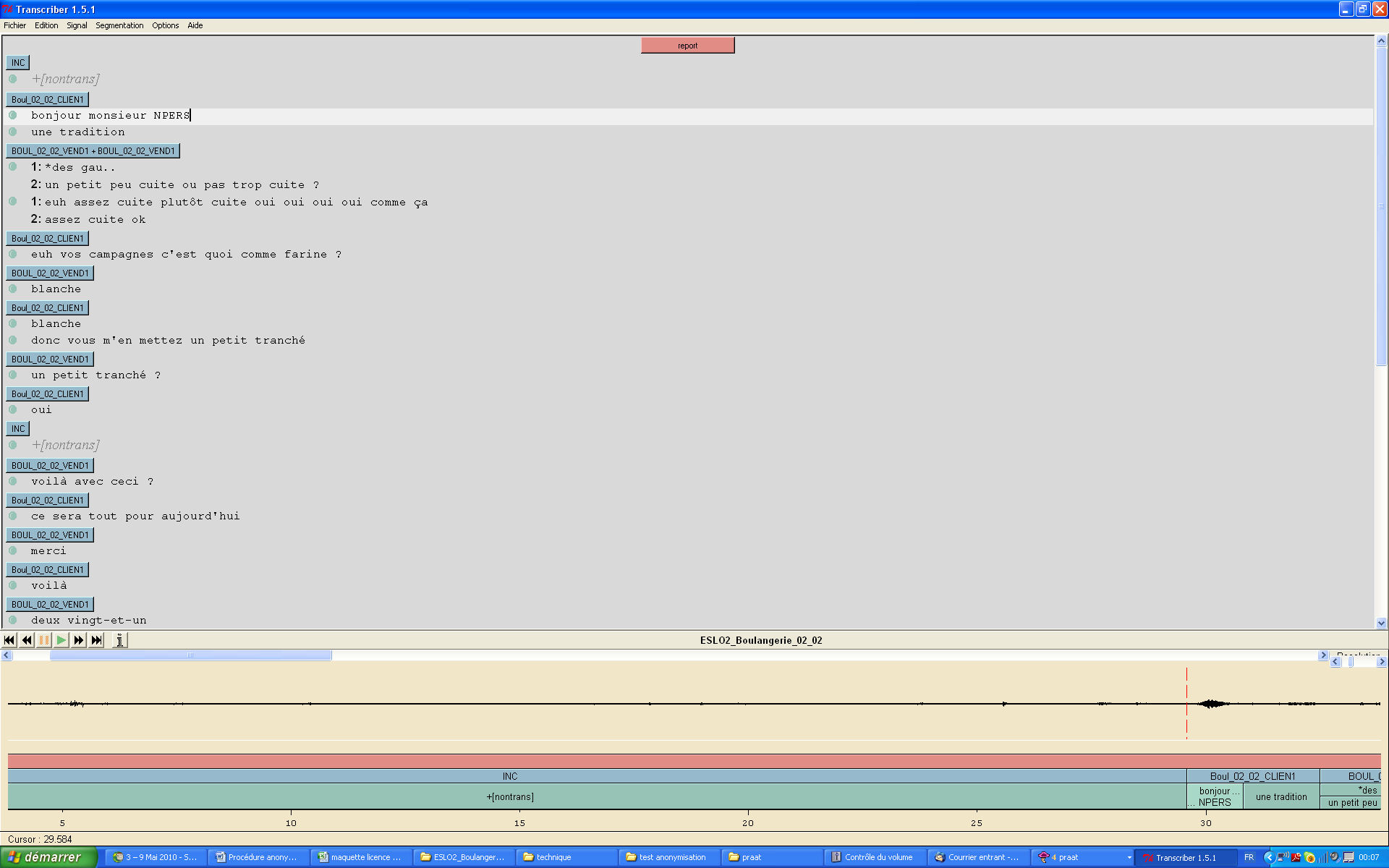
Cette procédure utilise, avec son autorisation, le script développé par D. Hirst. Ce script est disponible à l’adresse suivante :

[*http://uk.groups.yahoo.com/group/praat-users/files/Daniel\_Hirst/anonymise\_long\_sounds.praat*](http://uk.groups.yahoo.com/group/praat-users/files/Daniel_Hirst/anonymise_long_sounds.praat)

*Pour ESLO D. Hirst a modifié son script (gestion de la stéréo et conservation de l’amplitude, nous utilisons donc cette autre version du script.*

ANONYMISATION D’ESLO 2

La phase de remplacement par un hyperonyme se fait dès la transcription (version A).



Bibliographie :

Baude, O. *et alii* (2006), *Corpus oraux : guide des bonnes pratiques 2006.* Paris et Orléans, CNRS-Editions et PUO.

Eshkol I. (2007), Entrer dans l’anonymat. Etude des entités dénommantes dans un corpus oral. Actes du colloque NOMINA2007. Ed. Narr, Tübingen (à paraître).

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.