



*Ethique et Traitement Automatique des Langues  
Journée de l'ATALA, INALCO, Paris, le 22 novembre 2014*

# ***PROCÉDURE D'ANONYMISATION ET TRAITEMENT AUTOMATIQUE : L'EXPÉRIENCE D'ESLO***

**Eshkol-Taravella I.<sup>(1)</sup>, Kanaan-Caillol L.<sup>(1)</sup>, Baude O.<sup>(1)</sup>,  
Dugua C.<sup>(1)</sup>, Maurel D.<sup>(2)</sup>**

**(1) LLL UMR 7072, Université d'Orléans**

**(2) LI, Université de Tours**

# PLAN

- Présentation du corpus ESLO
- Questions juridiques et anonymisation des données sonores
- L'expérience d'anonymisation automatique d'ESLO
- Procédure d'anonymisation actuelle



# ESLO ENQUÊTES

## SOCIOLINGUISTIQUES À ORLÉANS

**ESLO1 et ESLO2** : deux collectes à quarante ans d'intervalle (1968 – 2008)

### Un grand corpus

- 700 heures d'enregistrement
- évalué à 10 millions de mots
- + de 500 locuteurs en interaction
- une fiche de métadonnées pour chaque locuteur, enregistrement, transcription

**Un corpus sociolinguistique** : variété des situations (entretiens, repas, conférences, micro-trottoir, etc.)

# OUTIL DE TRANSCRIPTION TRANSCRIBER (SORTIE XML)

The screenshot shows the Transcriber 1.5.1 application window. The title bar reads "Transcriber 1.5.1". The menu bar includes "Fichier", "Edition", "Signal", "Segmentation", "Options", and "Aide". The main text area contains the following transcription segments:

- report
- RC  
monsieur
- report - QP1
- RC  
depuis combien de temps habitez vous Orléans ?
- GJ 131  
oh ça fait neuf ans depuis dix neuf cent soixante
- report - QP3
- RC  
vous vous plaisez à Orléans ?
- GJ 131

Below the text area is a control bar with playback buttons and the file name "008\_A[1].trs" and "008\_Section024\_report\_P11\_856.67.878.64.wav". A waveform is displayed below the control bar. At the bottom, a timeline shows the transcription segments aligned with the audio. The timeline has a scale from 0 to 16 seconds. The segments are color-coded: report (red), RC (blue), and GJ 131 (green).

report	QP1				
RC	RC	GJ 131	RC	GJ 131	RC
monsieur	depuis combien de temps habitez vous Orléans ?	oh ça fait neuf ans depuis dix neuf cent soixante	vous vous plaisez à Orléans ?	oui et non [rire]	pourquo i cela ?

Cursor : 0

```
<Turn speaker="spk1"  
startTime="97.254"  
endTime="213.606">  
<Sync time="99.537"/>  
euh voilà ça marche ou quoi ?  
<Sync time="106.591"/>
```

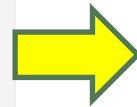
★ Transcription orthographique non aménagée ★

# ESLO : UN GRAND CORPUS DISPONIBLE

<http://eslo.huma-num.fr/>

<http://ortolang.fr>

The screenshot shows the 'Sélection du corpus' section of the ESLO interface. It includes two checkboxes for 'ESLO1' and 'ESLO2'. Below them is a 'Sélection catégorie(s)' section with two dropdown menus. The left menu contains: Entretien, Contact, Ouverture de l'entretien, Cloture de l'entretien, Repas, and Magasin. The right menu contains: Entretien, Conférences, Itinéraire, Cinéma, Diachronie, and Entretien jeunes. Below the dropdowns are two empty text input fields. At the bottom, there is a section for 'Sélection des caractéristiques des locuteurs, des enregistrements, des transcriptions' with three checkboxes: 'Enregistrement', 'Locuteur', and 'Transcription'. A 'Recherche d'une occurrence' button is at the very bottom.



## ORTOLANG

Outils et Ressources pour un Traitement Optimisé de la LANGUE

ORTOLANG est un **équipement d'excellence** validé dans le cadre des **investissements d'avenir**. Son but est de proposer une infrastructure en réseau offrant un **réservoir de données** (corpus, lexiques, dictionnaires, etc.) et d'outils sur **la langue et son traitement** clairement disponibles et documentés qui :

- permette, au travers d'une véritable **mutualisation**, à la recherche sur l'analyse, la **modélisation** et le **traitement automatique** de notre langue de se hisser au **meilleur niveau international** ;
- **facilite l'usage** et le transfert des ressources et **outils** mis en place au sein des **laboratoires publics** vers les **partenaires industriels**, en particulier vers les PME qui souvent ne peuvent pas se permettre de développer de telles ressources et outils de traitement de la langue compte tenu de leurs coûts de réalisation ;
- **valorise** le français et les langues de France à travers un **partage des connaissances** sur notre langue accumulées par les laboratoires publics.

ORTOLANG est un service **spécialisé pour la langue**, complémentaire de l'offre générale proposée par Huma-Num (très grande infrastructure de recherche).

En savoir [plus sur le projet...](#)



## Mise à disposition → Anonymisation

# ANONYMISATION : ASPECTS JURIDIQUES

- anonymiser : assurer l'impossibilité d'identifier des personnes
- juridiquement : anonymisation obligatoire en l'absence de consentement

Choix et procédure d'anonymisation (aspects juridiques et éthiques) dans ESLO Corpus oraux : *Guide des bonnes pratiques* (Baude et al\*. 2006)

\*Groupe de travail : linguistes, juristes, informaticiens, conservateurs

# DONNÉES NOMINATIVES ET CONSENTEMENT

## Consentement témoins ESLO

- participation au projet
- conservation des données nominatives pour contact
- utilisation des données pour recherche et diffusion

Le formulaire « consentement » signale l'anonymisation

⇒ **consentement éclairé**

Données nominatives (nom, adresse, numéro de téléphone)

- conservation dans BDD physiquement indépendante

Cas du module « Diachronie »

# ANNOTATION EN ENTITÉS DÉNOMMANTES



# ENTITÉS DÉNOMMANTES : DÉFINITION

- **identifiant direct** : sa présence est nécessaire et suffisante pour la reconnaissance de l'individu.

- nom rare de la personne (*Kanaan, Eshkol*)
- métier rare (*général*) ou statut (*maire*)
- caractéristique rare (*nombre élevé d'enfants, handicap*)

- **identifiant non direct** ou "élément sensible" contribuant à l'identification :

- sa présence est nécessaire mais pas suffisante pour la reconnaissance de l'individu
- sa présence seule ne permet pas l'identification, mais en combinaison avec d'autres identifiants, il peut désigner un référent unique

*le locuteur est patron d'un bar au moment d'enregistrement, et avant il travaillait dans l'aviation militaire*

- plus sensibles à l'anonymisation car apportent une information plus importante et plus spécifique
- plus généraux
  - les noms de famille comme *M.Dupond* ou *M.Durand*
  - les noms de métiers *professeur de physique/enseignant*

# TYPES DES ENTITÉS DÉNOMMANTES

## ○ Entités nommées

## ○ Noms communs :

- métier
  - *comme officier j'ai été obligé de rester 45 ans, je suis enseignant dans dans l'école publique*
- origine
  - *oui je suis orléanaise*
- maladie
  - *ça lui a même occasionné une petite scoliose déformation légère de la colonne vertébrale*
- études
  - *je suis licencié licencié en physique*
- loisirs
  - *je suis scout de France le jeudi soir où j'anime un un atelier photos*
- famille
  - *mon neveu [...] Jean-Pierre et puis malheureusement j'ai perdu mon mari assez tôt aussi et ça fait quinze ans*

# MODULE DE REPÉRAGE AUTOMATIQUE

## Méthodologie choisie : approche « en surface »

- utilisation du système CasSys (Friburger, 2002)

## Corpus

- corpus ESLO1 : 112 entretiens « face-à-face » transcrits
- six fichiers ont été réservés pour les tests et neuf fichiers pour l'évaluation

## Adaptation de CasSys au corpus oral

- prétraitement du corpus
  - le découpage en phrases d'Unitex a été remplacé par un découpage en fonction des balises Transcriber
- prise en compte de disfluences de l'oral
- les *questions sensibles/questions neutres*

## Processus en plusieurs étapes

- repérage des entités nommées
- repérage des entités dénommantes
- validation par un humain
- remplacement par un hyperonyme

# TYPOLOGIE DES ENTITÉS DÉNOMMANTES

- 1) le type *personne* permet de repérer les informations concernant la personne interrogée et celles qu'il donne sur sa famille ;
- 2) le type *identité* marque des informations précises comme la date de naissance ou la date d'arrivée à Orléans, l'âge de la personne dont on parle, son origine, sa date de mariage, etc. ;
- 3) le type *travail* étiquette le métier, le secteur d'activité, le lieu de travail ou le nom de l'entreprise de la personne dont on parle ;
- 4) le type *engagement* concerne la vie associative (y compris syndicale ou parentale) et la vie militaire ;
- 5) le type *voyage* les différents déplacements car il ne faut pas oublier que ceux-ci étaient plus rares à l'époque du questionnaire qu'aujourd'hui ;
- 6) le type *études* indique les diplômes, les lieux ou les établissements

# DIFFICULTÉS

## ○ Informations difficiles à cataloguer

- *on a monté une association d'élèves infirmières*
- *nous louons une villa à Royan*
- *mon père a fondé un le plus grand cabinet d'ophtalmologiste de la ville*
- *j'attends un deuxième bébé*
- *je suis scout de France le jeudi soir où j'anime un un atelier photos*

## ○ Variation d'une même information

- *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*

## ○ Disfluences

# EVALUATION

- 112 fichiers annotés
- 6 fichiers réservés pour les tests et 9 pour l'évaluation
- total : 77 entités dénommantes au total; reconnu : 69 entités
- 4 erreurs; 12 entités non reconnues
- précision est 94,2 % , rappel est 84,4 %

# RÉSULTATS

**moi je suis de Pithiviers**

## 1. Repérage des entités nommées

`<EN type="loc.admi"> Pithiviers</EN>`

## 2. Repérage des entités dénommantes

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

# RÉSULTATS

- *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>*



L'ANONYMISATION ACTUELLE DANS ESLO  
*UNE PROCÉDURE SEMI-AUTOMATIQUE*

# ANONYMISATION ACTUELLE DANS ESLO

Anonymisation de 2 types d'objets

- Données : transcriptions et enregistrements
- Métadonnées : locuteurs et enregistrements

# L'ANONYMISATION DANS LA CHAÎNE DE TRAITEMENT

Phase 1 : Exhaustivité, représentativité, proportionnalité

Phase 2 : Techniques de collecte : formats d'enregistrement et numérisation

Phase 3 : Formation des enquêteurs et information des témoins

Phase 4 : Recueil des données

Phase 5 : Codage et catalogage / Anonymisation



Phase 6 : Transcription et alignement / Anonymisation



Phase 7 : Anonymisation du son



Phase 8 : Stockage, archivage et indexation

Phase 9 : Mise à disposition

Phase 10 : Données partagées : interopérabilité et protections

Phase 11 : Applications et développements

Phase 12 : Mise en place du suivi (maintenance, jouvence et sécurité) et applications

# ANONYMISATION DES MÉTADONNÉES

## Codage des locuteurs

- code aléatoire généré par la base : BA725
- code selon plan de nommage
  - porte les marques d'une relation : BA725FIL, CT418BBSIT
  - d'une catégorie : 653CLI, 653VEN, 308STAN
  - ou le numéro d'enregistrement lié : 412PERS, 068INC, 416PERS

# ANONYMISATION DES MÉTADONNÉES



## Fiche locuteur

Identifiant locuteur : CT418



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

# ANONYMISATION DES MÉTADONNÉES

Référence enregistrement: ESLO2\_REPAS\_1258

<b>Fichier son:</b>	ESLO2_REPAS_1258.wav
<b>Corpus:</b>	ESLO2
<b>Catégorie:</b>	Repas
<b>Précisions sur la catégorie:</b>	Enregistrements lors de repas
<b>Sujet:</b>	(text_and_corpus_linguistics) Français (Ethnologue: fra)
<b>Sommaire:</b>	Repas avec une baby-sitter et 2 enfants (4 et 7 ans). Le papa arrive à la fin de l'enregistrement.
<b>Editeurs:</b>	LLL Université d'Orléans
<b>Créateurs:</b>	LLL Université d'Orléans - ESLOs
<b>Chercheurs:</b>	■ Baude, Olivier
<b>Chercheurs locuteurs:</b>	
<b>Participants:</b>	
<b>Description des participants:</b>	
<b>Descriptions annexes:</b>	
<b>Remarques:</b>	
<b>Fiche modifiée par:</b>	lkanaan
<b>Date d'enregistrement:</b>	03/12/2012
<b>Droits:</b>	Copyright (c) 2014 Université d'Orléans/LLLFreely available for non-commercial use. This file is licensed under a Creative Commons License.
<b>Format:</b>	(IANA MIME Media Type: audio/x-wav)
<b>Durée:</b>	00:17:19
<b>Acoustique:</b>	Bonne
<b>Précisions acoustiques:</b>	
 <b>Lieu spatial:</b>	Olivet
<b>Lieu TGN:</b>	1034729
<b>Lieu Point:</b>	east=1.891; north=47.855
	
<b>Locuteurs:</b>	■ CT418 ■ CT418BBSIT ■ CT418SOE

# ANONYMISATION DANS LA TRANSCRIPTION

- NPERS pour les noms de personnes
- NANON autres segments permettant l'identification (entités dénommantes ou propos sensibles)

et est-ce que ça vous avez dit que ça n'a rien à voir avec

NPERS est-ce que c'est c'est pas

enfin si vous vous faites des constructions de citernes et tout ça c'est quand même euh ça a un rapport avec la

HU339 + JSM

1: non non non


2: fonderie non c'est

HU339

simplement euh je suis un NPERS qui se trouve à la tête de la NANON mais c'est tout

# ANONYMISATION DU SON

Outils :

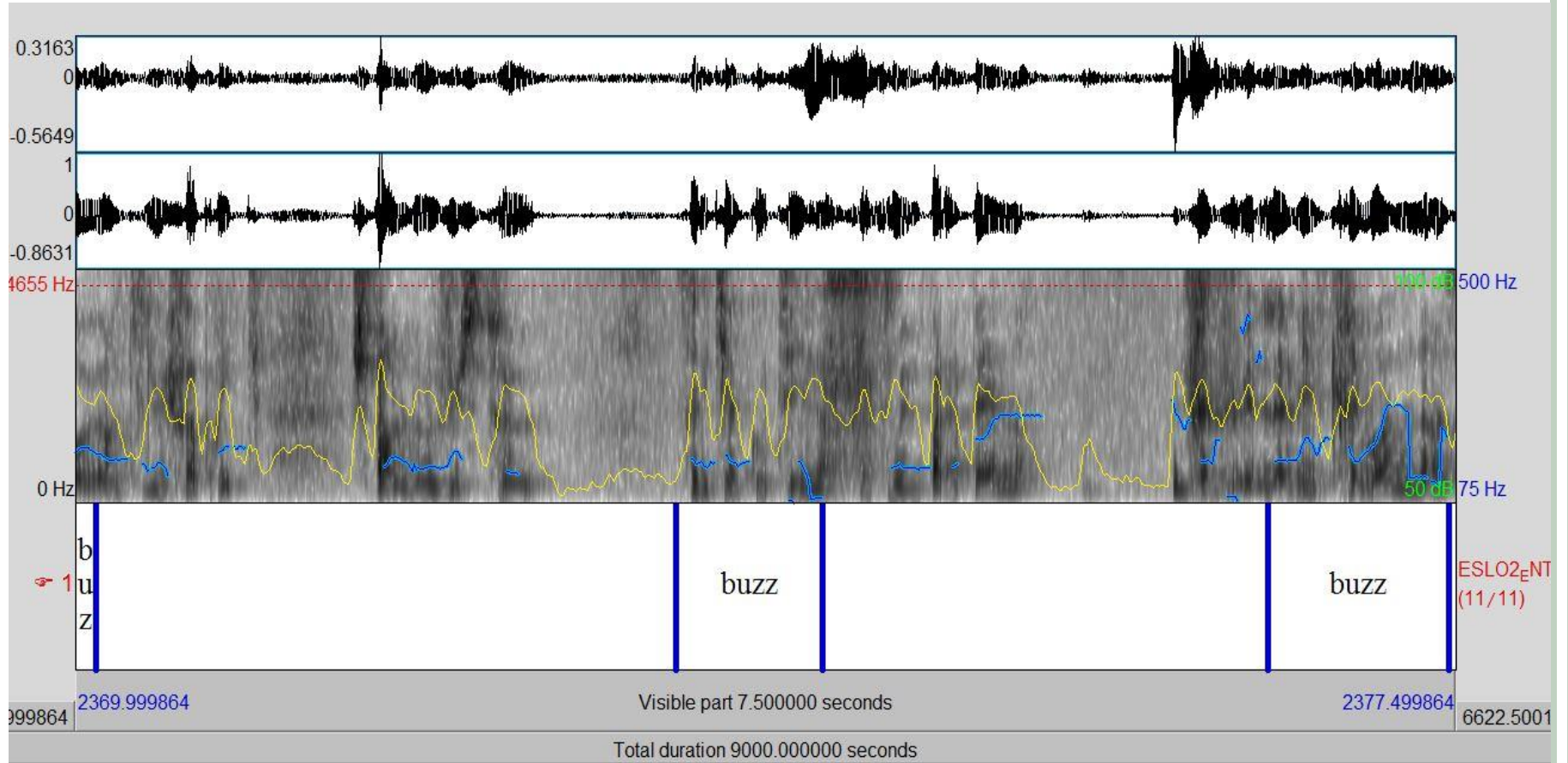
- Praat 
- Script D. Hirst (LPL, Aix-en-Provence)
- Application permettant le repérage automatique des NPERS et NANON + balise temps du segment dans les transcriptions (F. Badin, LLL, Orléans)

Sous Praat, isolation des segments/codage par BUZZ/  
exécution du script ➡ création fichier son  
« nomfichier\_anon »

★ Son brouillé mais courbe intonative conservée ★



# ANONYMISATION DU SON



**DEMO**

# CONCLUSION

## Anonymisation et TAL

- repérage et analyse des entités dénommantes
- anonymisation partielle
- limites du TAL