

**“ Propositions de Normalisation pour  
une Base de Corpus Multimédia ”  
à l'ED268**

Frédérique Bénard, Sonia Branca, Maria Candéa, Serge Fleury, Cédric Gendrot,  
Michel Jacobson, Thierry Pagnier, Patrick Renaud, André Salem, Pollet  
Samvelian, Luigi Samsonetti, Jacqueline Vaissière.

# Une base de données

par et pour

l'ED268

# Cette présentation est mise à votre disposition ...

- Site Web du projet :

<http://pi-ed268.univ-paris3.fr>

# Linguistique de corpus

⇒ Nécessité d'archivage (corpus d'origine, enregistrement, annotations, outils à utiliser,...)

⇒ Importance des bases de données pour la linguistique.

# Utilités d'une base de corpus

- Archivage de langues. (site LACITO : <http://lacito.vjf.cnrs.fr/archivage/>)
- Analyse / synthèse de la parole.
- Favoriser les collaborations pluri-disciplinaires.

# Guide des bonnes pratiques

... pour la constitution, l'exploitation, la conservation et la diffusion des corpus oraux.

[http://www.culture.gouv.fr/culture/dglf/corpus\\_oraux.htm](http://www.culture.gouv.fr/culture/dglf/corpus_oraux.htm)

- Séminaire BNF C-Oraux 17 mai 2005

# Guide ...

- 1- Recueil des données
- 2- Aspects juridiques
- 3- Transcription/Annotation

4- Exploitation, conservation, diffusion :

**Notre travail !!**

# Le cœur du projet : objectifs

- **L'élaboration d'une base de corpus (principalement oraux) regroupant des données de langues de différentes natures**
  - Une approche pluridisciplinaire
  - Une **normalisation** pour l'encodage et la **description** de corpus de langue (métadonnées)



- Parmi les membres du projet :

- Sociolinguistique : P. Renaud; S. Branca; T. Pagnier
- Phonétique : C. Gendrot; M. Candéa; J. Vaissière
- TAL / Informatique : S. Fleury; M. Jacobson; A. Salem ;  
F. Bénard
- Syntaxe, Sémantique : P. Samvelian
- Acquisition : L. Samsonetti

# Exploitation, conservation, diffusion

- Nécessité de normaliser :

- pourquoi ?

Besoins de partager, diffuser et rechercher/retrouver

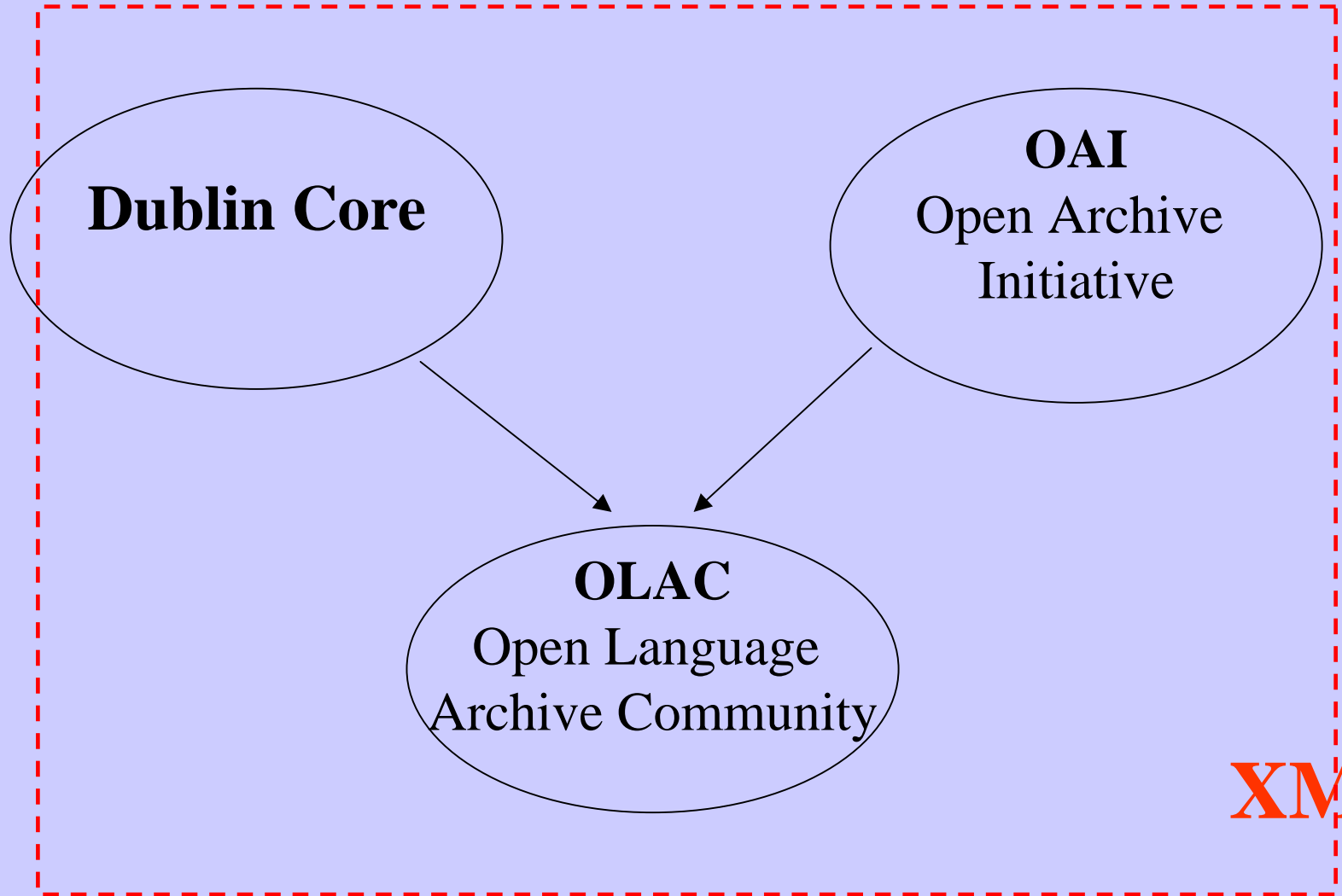
- Comment ?

XML / Dublin Core / OAI / OLAC

# Normalisation des données

- XML: Format de représentation
- Dublin Core: Norme de métadonnées
- OAI: Concept d'interopérabilité
- OLAC: Standard proposé par des linguistes

# Normalisation des corpus oraux



**XML**

# XML

- Langage à balises qui permet d'annoter et de structurer une ressource.
  - libre de droit, multi-plateforme, échangeable

- **par ex:**

```
<balise attribut= ''valeur''>donnée</balise>  
<titre lang=''fr''>norma...</titre>
```

# Dublin Core

- Norme de métadonnées.
- 15 éléments simples mais efficaces pour décrire les ressources :
  - Title, (creator), subject, description, publisher, contributor, date, type, format, identifier, language, relation, coverage, rights, source.

# OAI : Open Archive Initiative

- Concept « d'interopérabilité. »
  - Recherche sur les métadonnées.
  - Retrouver l'emplacement physique des corpus sans les télécharger.
  - Accessible à tous.

# OLAC: Open Language Archive Community

OLAC, communauté de linguistes

« Comment mieux décrire les ressources linguistiques, en partant de la norme du Dublin Core et de l'OAI ? »

⇒ Dublin Core pour les linguistes !!!



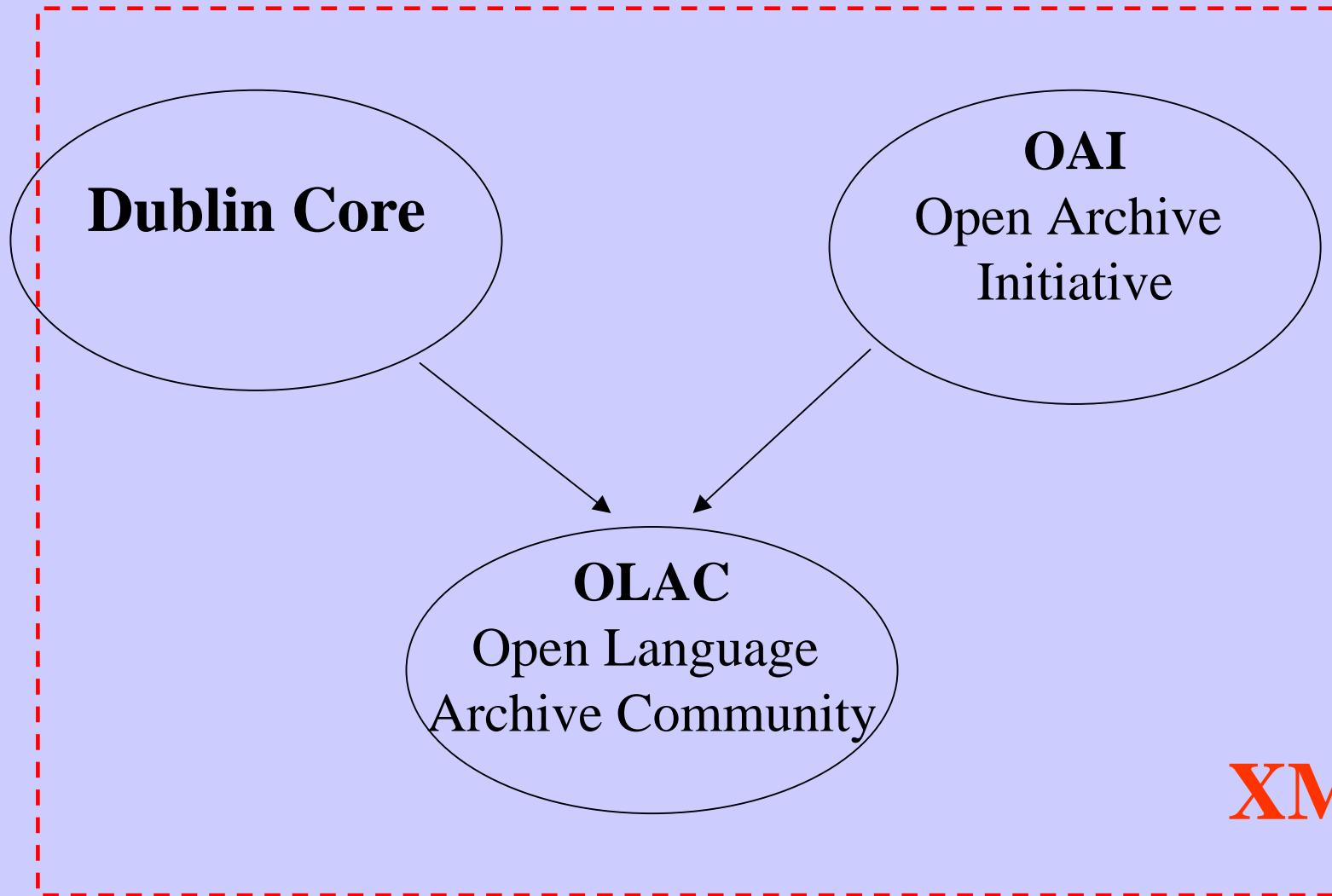
# Extensions OLAC

5 extensions du Dublin Core rattachées à la linguistique:

- Discourse Type : (drama, narrative, language play,...)
- Language Identification : (code ISO: fr, en,...)
- Linguistic Field : (sociolinguistique, phonétique,...)
- Linguistic Data Types : (lexicon, primary-text, language-description)
- Participant Roles : (annotator, author, speaker,...)

⇒OLAC ne remplace pas, mais complète, spécifie le DC par rapport aux attentes de la communauté linguistique.

# Normalisation des corpus oraux



**XML**

# MKM (Make Metadata) S. Fleury

- Comment insérer ces métadonnées de façon conviviale ?
- Outil « fait-maison »...



[HOME](#) [HELP-MAKEMETADATA](#) [HELP-DC-OLAC](#) [MKMETA1](#) [MKMETA2](#) [MKMETA3](#) [MKMETA4](#) [MKMETA5](#) [MKMETA6](#) [RESULT](#) [EXPORT](#)

## *Projet Innovant PIED268*

*L'objectif de ce projet est de proposer une réflexion et une démarche pour constituer des ressources linguistiques normalisées (données orales, écrites et vidéo) dans un cadre pluridisciplinaire. Sont en effet apparues ces dernières années de nombreuses tentatives internationales visant à normaliser les ressources électroniques (pour les sciences humaines en particulier (cf TEI, CES) ou plus généralement pour la diffusion des informations sur le web (projet web sémantique, W3C).*

*Ce projet s'inscrit dans cette perspective de constitution de ressources électroniques normalisées dans le cadre des corpus de langue.*

*L'outil [makeMETADATA](#) permet de générer un fichier de métadonnées pour une ressource donnée.*

## *Générateur de METADONNEES*

[ED268 \[U. DE PARIS 3, Sorbonne nouvelle\]](#)

Projet Innovant 2004-2006 PIED268

<http://pi-ed268.univ-paris3.fr>



HOME HELP-MAKEMETADATA HELP-DC-OLAC MKMETA1 MKMETA2 MKMETA3 MKMETA4 MKMETA5 MKMETA6 RESULT EXPORT

## MODE D'EMPLOI

### 1. Génération des métadonnées :

Pour constituer les **métadonnées**, *vous devez remplir l'ensemble des formulaires MKMETA1, MKMETA2, MKMETA3, MKMETA4, MKMETA5, MKMETA6.*

Pour chacun de ces onglets, *compléter la colonne METADONNEES.*  
Pour vous aider dans cette tâche vous pouvez consulter les fichiers d'aide disponibles (sur la colonne la plus à droite de chaque ligne du formulaire).

Dans chacun de ces onglets, vous trouverez soit des cases à cocher, soit des zones de saisie, soit un bouton "Edit" donnant accès à un éditeur.

Les zones de saisie se composent de deux champs (de saisie) :

- un pour entrer la valeur de la métadonnée idoine,
- l'autre pour décrire la langue utilisée dans le premier champ

Par défaut ce second champ est initialisé avec la valeur "fr".

Vous pouvez modifier cette valeur par défaut en utilisant le tableau de codage des langues utilisé par OLAC et disponible à cette adresse : <http://www.language-archives.org/OLAC/1.0/languageCodes.xml>

## Générateur de METADONNEES

ED268 [U. DE PARIS 3, Sorbonne nouvelle]

Projet Innovant 2004-2006 PIED268

<http://pi-ed268.univ-paris3.fr>



Init Import Build Exit

HOME HELP-MAKEMETADATA HELP-DC-OLAC MKMETA1 MKMETA2 MKMETA3 MKMETA4 MKMETA5 MKMETA6 RESULT EXPORT

## Dublin Core

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

Subject

## Raffinement DC

## Attribut OLAC

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

linguistic-field

## Vocabulaire OLAC

*anthropological\_linguistics**applied\_linguistics**cognitive\_science**computational\_linguistics**discourse\_analysis**forensic\_linguistics**general\_linguistics**historical\_linguistics**language\_acquisition**language\_documentation**lexicography**linguistics\_and\_literature**linguistic\_theories**mathematical\_linguistics**morphology**neurolinguistics**philosophy\_of\_language**phonetics*

## METADONNEES

## Générateur de METADONNEES

ED268 [U. DE PARIS 3, Sorbonne nouvelle]

Projet Innovant 2004-2006 PIED268

<http://pi-ed268.univ-paris3.fr>

# code résultant

```
<dc:subject xsi:type="olac:linguistic-  
field" olac:code="phonetics" />
```

# Normalisation

- ... notre travail → normalisation XML
- ... et le vôtre ...



# Guide (...) des corpus oraux

- 1- Recueil des données : recommandations
- 2- Aspects juridiques : recommandations, anonymisation
- 3- Transcription/Annotation : recommandations
- 4- Exploitation, conservation, diffusion :

Notre travail !!

# Recueil des données (p.97)

- Méthode – matériel
  - (mini) K7
  - Mini-Discs (compression)
  - DATs
  - Mini-Discs (nouvelle génération)
  - enregistreur numérique (MP3 – WAV)

# Aspects juridiques (p.81)

- Non négligeable ...
- Précautions à prendre ...
- Anonymisation

# Outils d'annotations

(par ordre alphabétique)

- AGTK TableTrans - MultiTrans
- CLAN
- ELAN
- Praat
- SoundIndex
- Transcriber
- WinPitch
- **Word (!!)**

# Un petit résumé

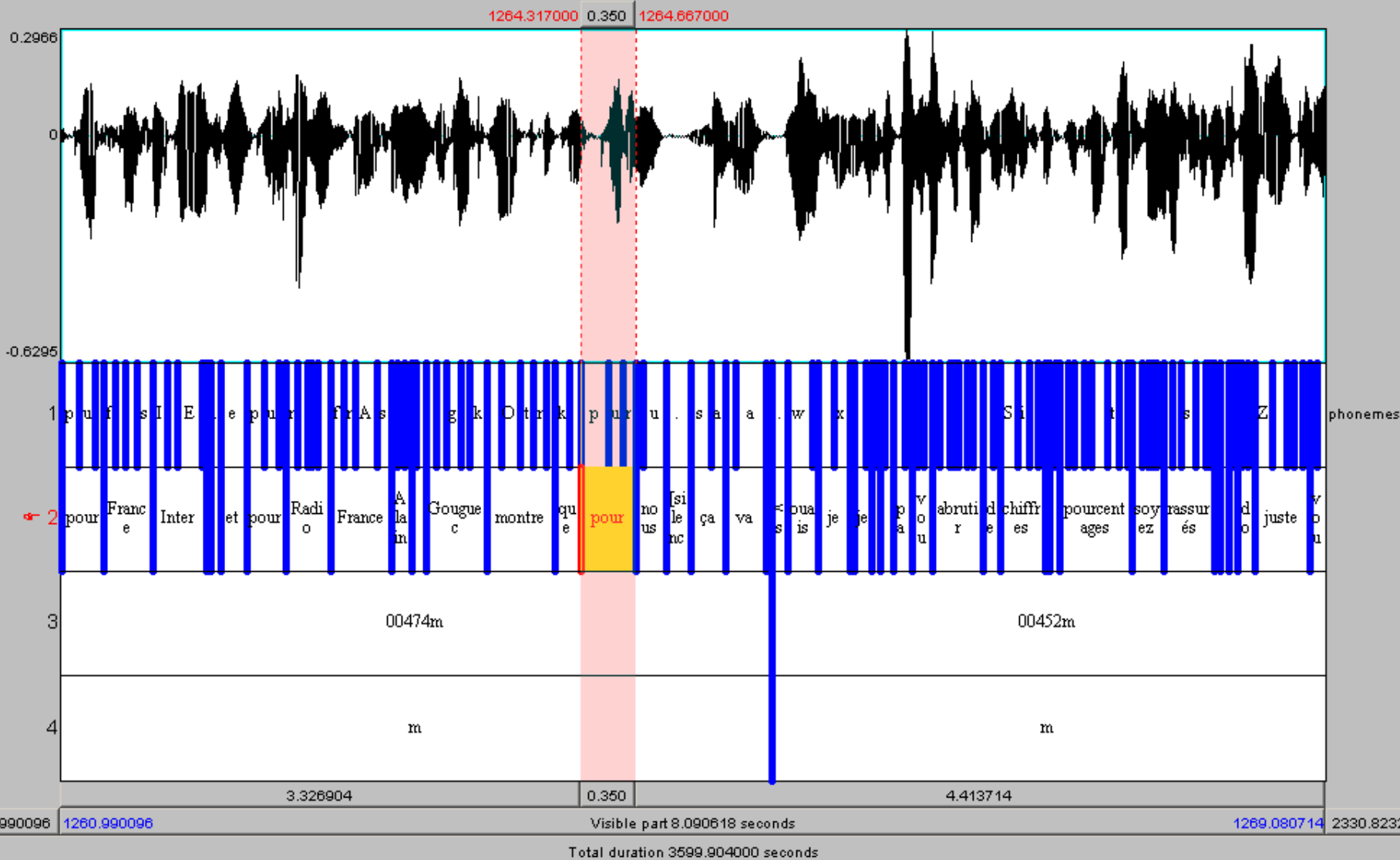
1. Aucun outil d'annotation du signal n'est évidemment parfait.
2. Tout dépend du type d'annotation que l'on souhaite effectuer (prosodique, phonétique, syntaxique, morphologique, orthographique, ...)
3. Libre de droit !!
  - Praat
  - Transcriber
  - ELAN
  - SoundIndex

**=> Savoir-faire**

# Annotations (1)

- Praat : <http://www.fon.hum.uva.nl/praat>
  - transcription fine
  - convertible dans une bonne mesure en un format XML
  - 8 niveaux d'analyse possibles (niveaux indépendants, ancrage temporel nécessaire)
  - analyse du signal

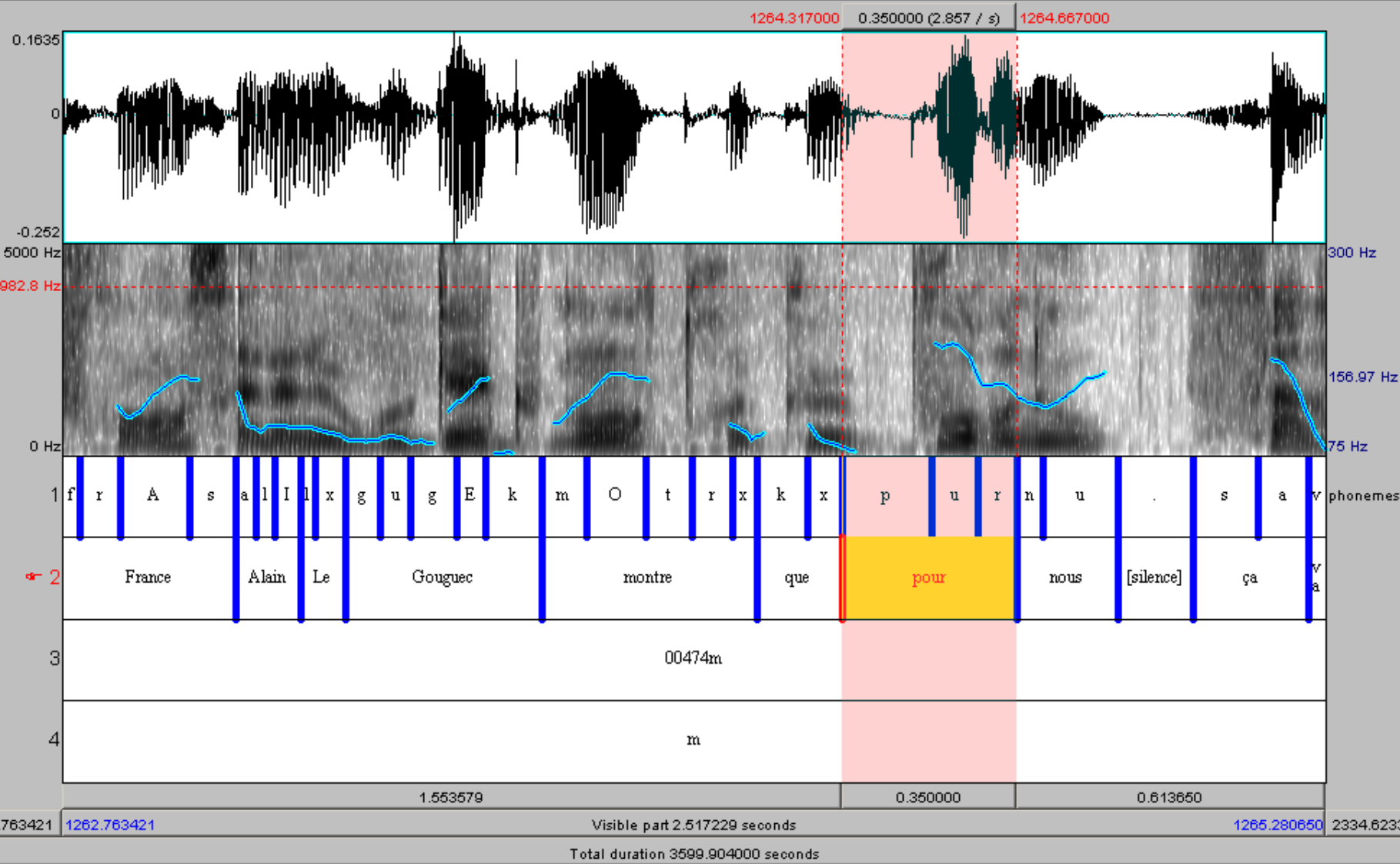
pour



all in out sel

Group

pouf





# Annotations (2)

Transcriber : <http://www.etca.fr/CTA/gip/Projets/Transcriber/IndexFr.html>

- XML (dtd spécifique),
- très pratique pour les dialogues, pas de transcription fine
- un seul niveau d'analyse possible.

filler

Simon Tivolle + Patricia Martin

1: ouais .

2: sûr ?

Simon Tivolle

ah bon ?[rire] non . blague , blague de Patricia .

[i] France-Inter , [rire-] il est 7 heures [-rire] .

nontrans

(no speaker)

♪ ♪

filler

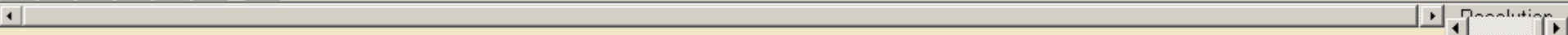
Patricia Martin

le journal , Simon Tivolle :

report - les titres



frint980428



		music		music					
	filler	nontrans	filler	les titres					
S	Simon Tivolle		(no speaker)	Patricia.	Simon Tivolle				
...	ah bon ?[rire] non...	[i] ...		le ...	[i] ...	mardi.	la consultation ...	[i] grand débat...	pour tirer.
...	... Patricia .	....		..Tivolle :	... r !	...	... lycées :	... demain à Lyon	... ents du
0		5		10			15		20

Cursor : 0

Démarrer



Poste de... C:\Docu... C:\Docu... partie\_e... C:\Docu... C:\Docu... Microsoft... Elan - el... 7% Transcr...



01:59

# Annotations (3)

- ELAN : <http://www.mpi.nl/tools/elan.html>
  - sortie XML (dtd spécifique)
  - très complet (vidéo)
  - Plusieurs niveaux d'analyse possibles (niveaux dépendants/indépendants, nécessité de spécifier l'ancrage temporel pour chaque niveau).

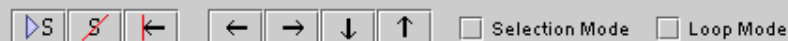


Grid Text Subtitles Controls

Empty				
Nr	Annotation	Begin Time	End Time	Duration

00:00:19.737

Selection: 00:00:22.512 - 00:00:22.930 418



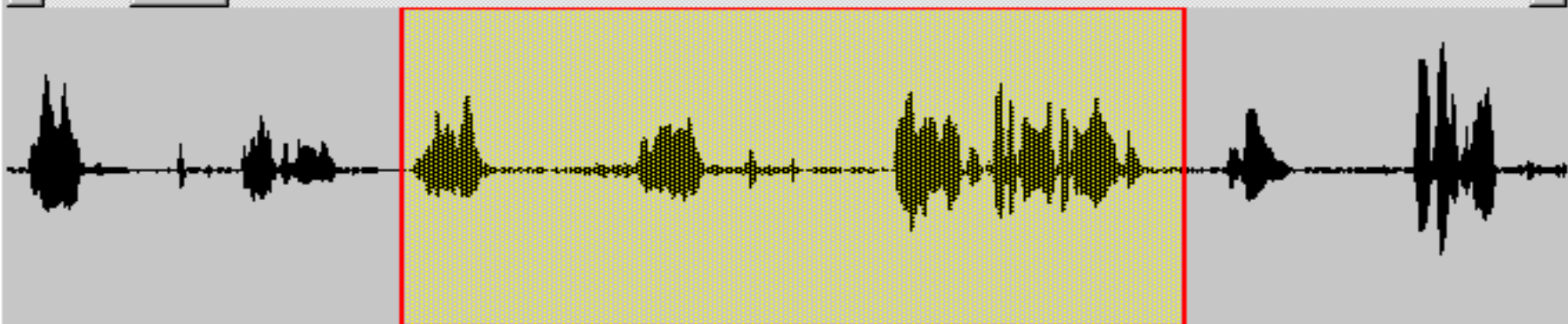
	00:00:16.000	00:00:17.000	00:00:18.000	00:00:19.000	00:00:20.000	00:00:21.000	00:00:22.000	00:00:23.000	00:00:24.000													
K-Spch	en you follow the sign kleeF																					
W-Spch	en you follow the sign kleeF		you come down		you know eh after this trajanus plein			you come down to rhine eh valley														
W-Words	en	you	follo	th	sign	KleeF	yo	co	down	yo	know	eh	after	this	Trajanus	Plei	yo	co	dow	t	the	Rhine
W-POS	dv	pro	v	a	n	n	pro	v	adv	pro	v	post	prep	dem	n	n	pr	v	adv	p	art	n
W-IPA	en ju: foləʊ ðə saɪn kle:f		ju: kʌm daʊn		ju: nə: ə aftə ðɪs trəˈdʒənəs pleɪn			ju: kʌm daʊn tə ðə		raɪn ə væli												
W-RGU	ju: nə: ə aftə ðɪs trəˈdʒənəs pleɪn																					
W-RGph	ration	hold	str	hold	preparati	stroke	hold	preparation	stroke	hold	stroke	preparati	stroke	hold	hold	pre						

# Annotations (4)

- **SoundIndex** : <http://michel.jacobson.free.fr/>
  - Sortie XML (pas de dtd spécifique),
  - Plusieurs niveaux d'analyse possibles (niveaux dépendants/**indépendants**, possibilité de spécifier l'ancrage temporel pour chaque niveau mais pas indispensable)...
  - Mais peu ergonomique.

- Ancrage temporel des transcriptions ... pas toujours nécessaire : annotation morpho-syntaxique par exemple...

```
<TEXT id="hayu1" lang="hayu">
<HEADER>
  <TITLE>Two sisters.</TITLE>
  <SOUNDFILE href="SOEURS.mp3">
</HEADER>
<S id="nepal1s1">
  <AUDIO start="2.3656" end="7.9256"></AUDIO>
  <FORM>nakpu nonotso siŋ pa laʔnatshe-m are.</FORM>
</S>
<S id="nepal1s2">
  <AUDIO start="7.9256" end="23.2255"></AUDIO>
  <FORM>siŋ pa lat-noŋ ban-noŋ... ban-noŋ bilɔ ɔxtotshe-m are.</FORM>
</S>
<S id="nepal1s3">
  <AUDIO start="23.2255" end="33.8056"></AUDIO>
  <FORM>bilɔ!</FORM>
</S>
<S id="nepal1s4">
```



# Conclusion :

- Offrir aux membres de l'ED268 la possibilité de :
- Mettre à disposition leurs corpus.
  - Dès la prochaine rentrée universitaire
- Utiliser pour leurs recherches les données stockées dans cette base de données



# ET VOUS ???

- Vos préférences ...
- Vos attentes ?

## Contacts et Liens ...

- Site Web :

<http://pi-ed268.univ-paris3.fr>

- Serge Fleury :

[serge.fleury@univ-paris3.fr](mailto:serge.fleury@univ-paris3.fr)

- Cédric Gendrot :

[cgendrot@univ-paris3.fr](mailto:cgendrot@univ-paris3.fr)