VIII[èmes] Rencontres de l'ED268
21 mai 2005

" 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, Luiggi 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

- 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

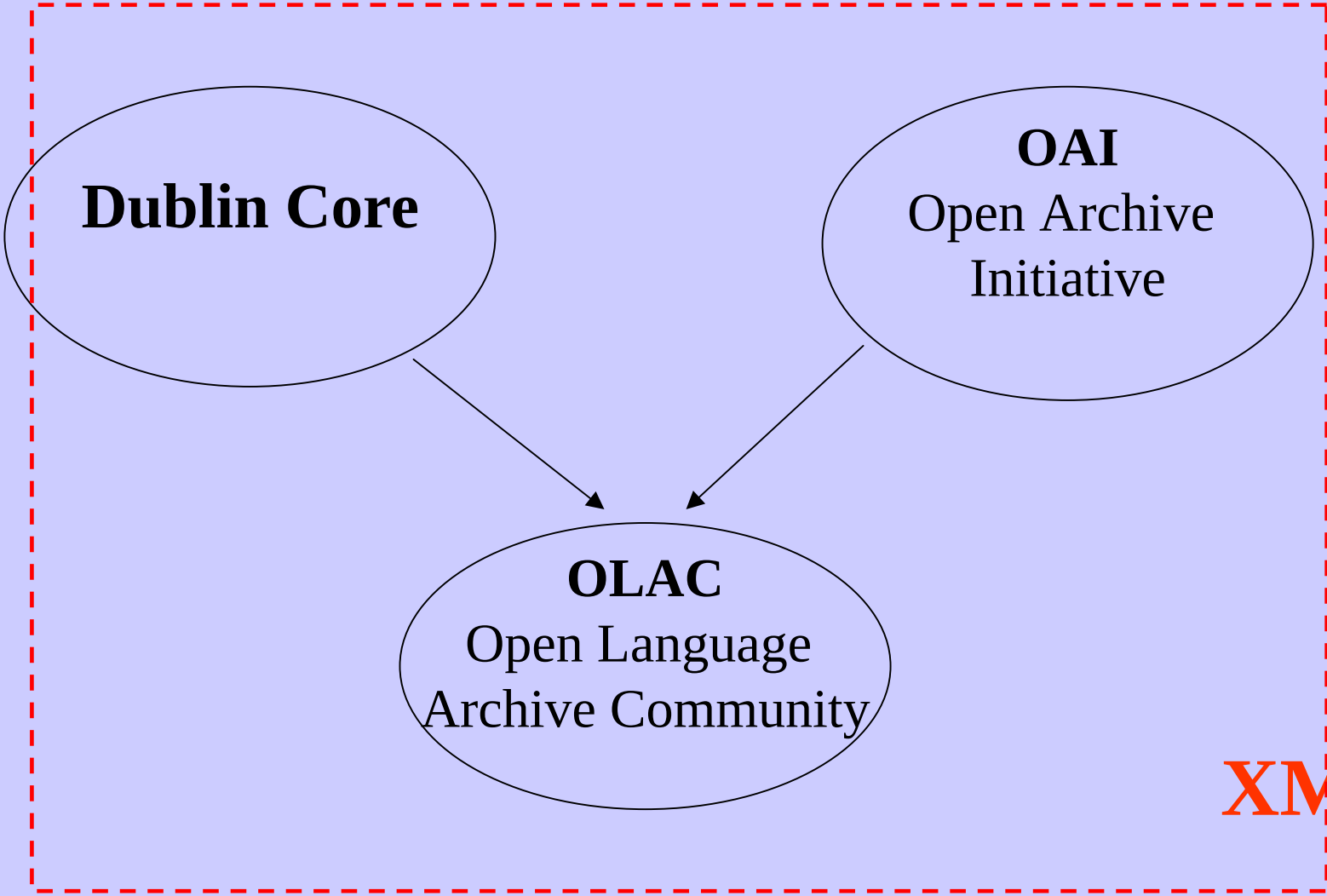
  - **Comment ?**

    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

- 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 !!!

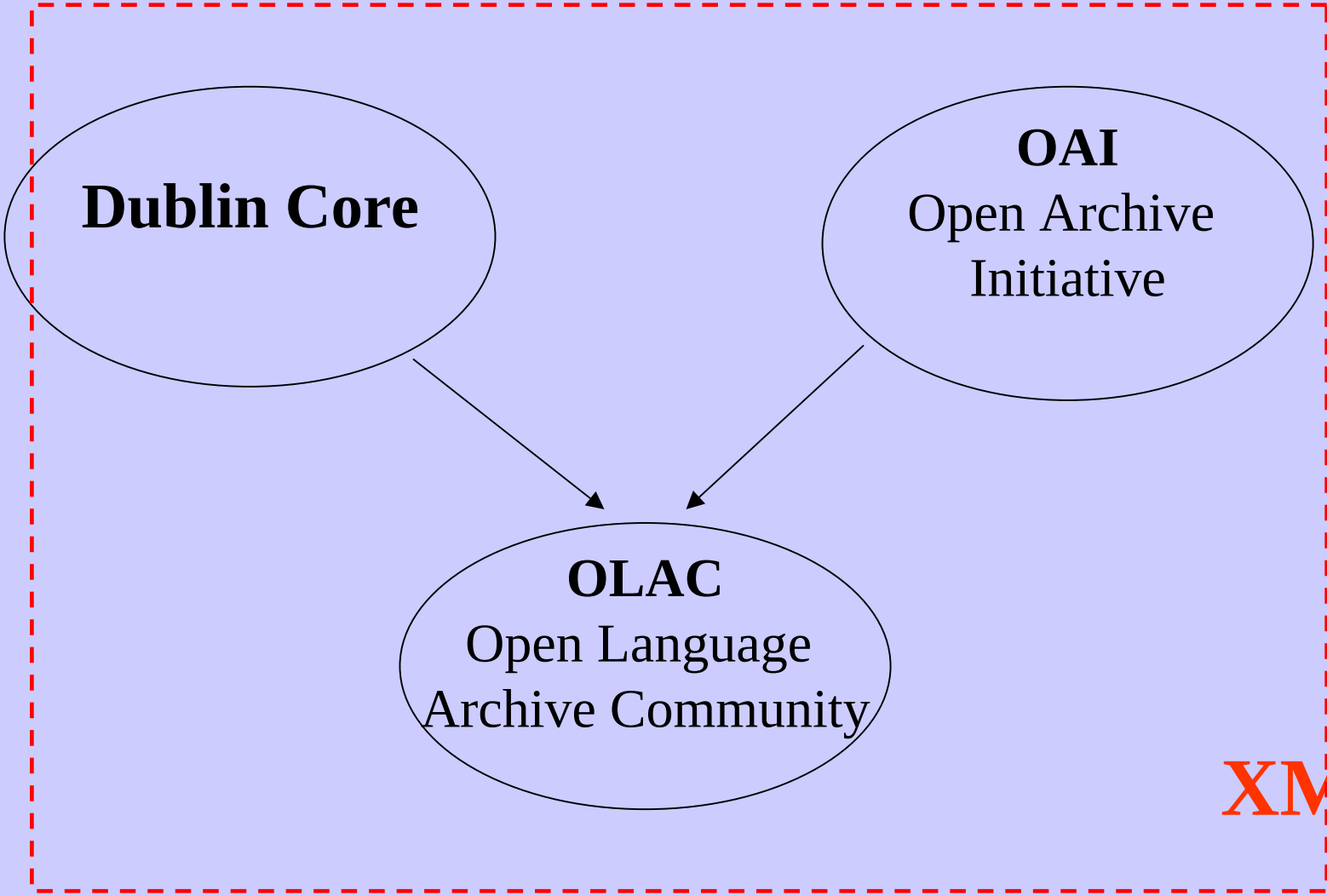
Extensions OLAC

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



MKM (Make Metadata) S. Fleury

- Comment insérer ces métadonnées de façon conviviale ?

- Outil « fait-maison »…

makeMETADATA@ED268-P3 v.1.0.6

Init | Import | Build | Exit

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

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.xsd

makeMETADATA@ED268-P3 v.1.0.6

Init | Import | Build | Exit

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

Dublin Core	Raffinement DC	Attribut OLAC	*Vocabulaire OLAC*	METADONNEES
Subject		linguistic-field	*anthropological_linguistics*	☐
Subject		linguistic-field	*applied_linguitics*	☐
Subject		linguistic-field	*cognitive_science*	☐
Subject		linguistic-field	*computational_linguistics*	☐
Subject		linguistic-field	*discourse_analysis*	☐
Subject		linguistic-field	*forensic_linguistics*	☐
Subject		linguistic-field	*general_linguistics*	☐
Subject		linguistic-field	*historical_linguistics*	☐
Subject		linguistic-field	*language_acquisition*	☐
Subject		linguistic-field	*language_documentation*	☐
Subject		linguistic-field	*lexicography*	☐
Subject		linguistic-field	*linguistics_and_literature*	☐
Subject		linguistic-field	*linguistic_theories*	☐
Subject		linguistic-field	*mathematical_linguistics*	☐
Subject		linguistic-field	*morphology*	☐
Subject		linguistic-field	*neurolinguistics*	☐
Subject		linguistic-field	*philosophy_of_language*	☐
Subject		linguistic-field	*phonetics*	☑

MkMe

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

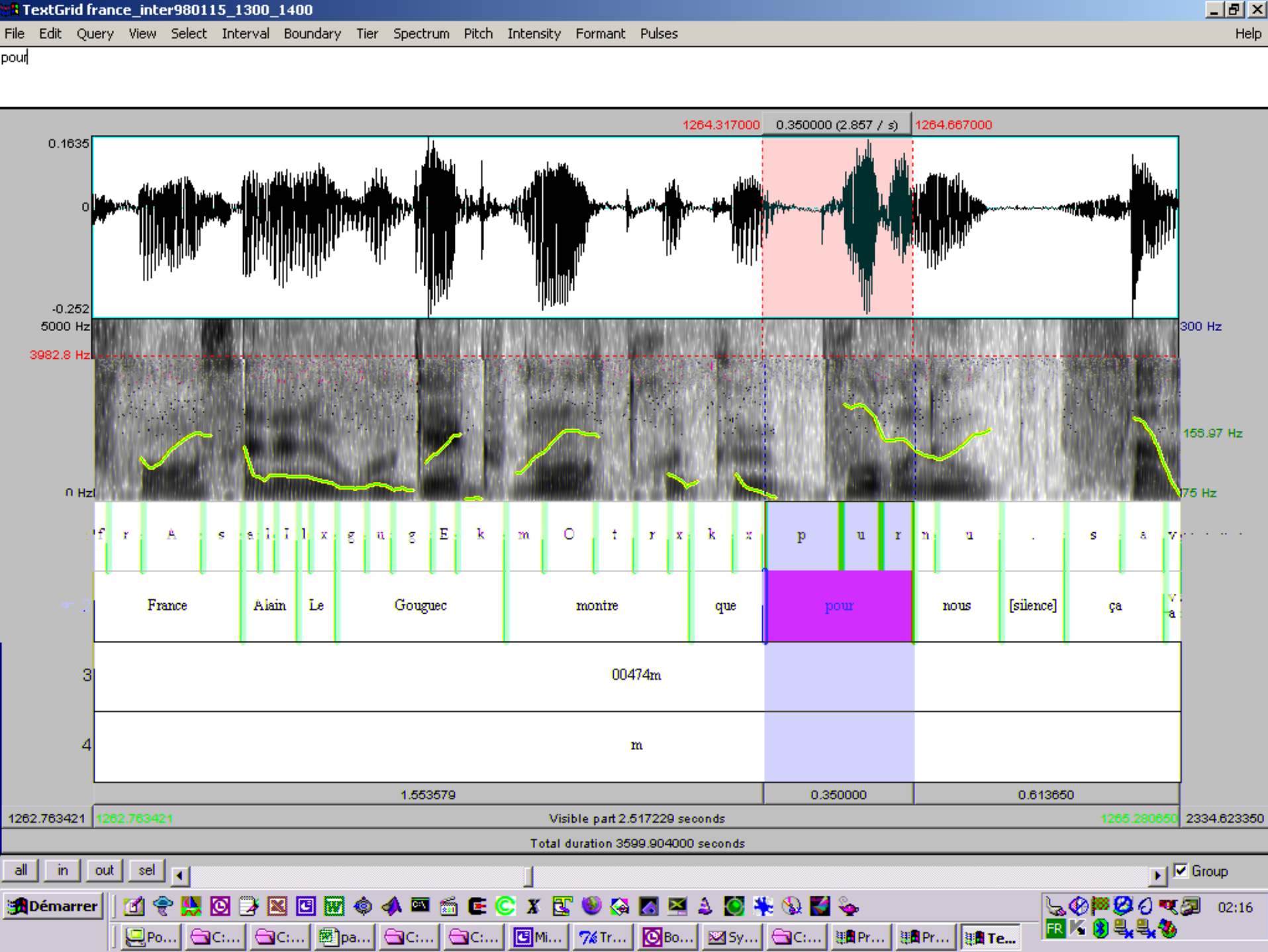
1. Aucun outil d'annotation du signal n'est évidemment parfait.
2. Tout dépend du type d'annotation que l'on souhaite effectuer (prosodique, phonétique, syntaxique, morphologique, orthographique, …)
3. Libre de droit !!

- Praat
- Transcriber
- ELAN
- SoundIndex

=> Savoir-faire

Annotations (1)

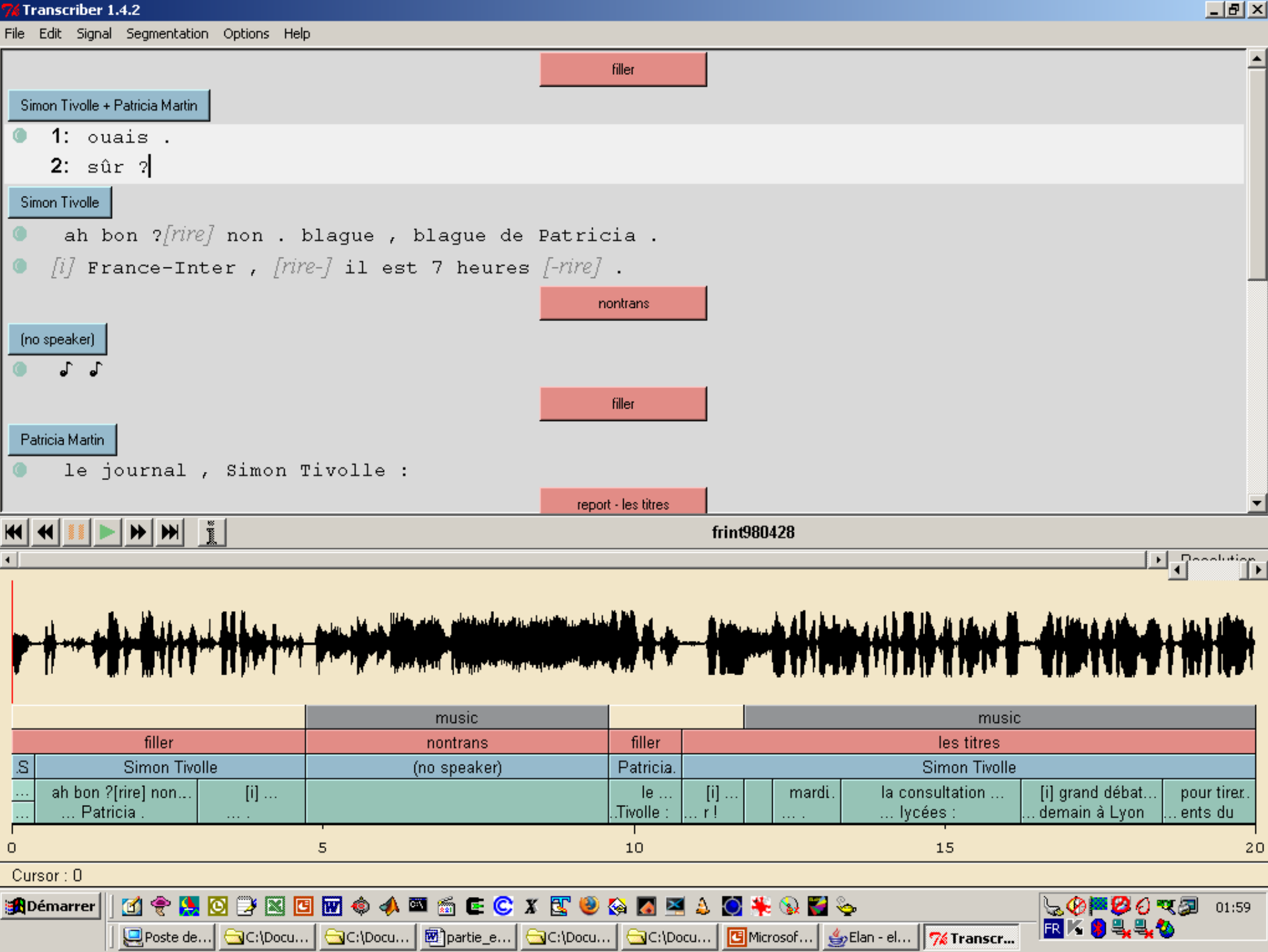
- Praat : http://www.fon.hum.uva.nl/praat
  - transcription fine
  - convertible dans une bonne mesure en un format XML
  - 8 niveaux d'analyse possibles (niveaux indépendants, ancrage temporel nécessaire)
  - analyse du signal


TextGrid france_inter980115_1300_1400
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses Help
pour
1264.317000
0.350000 (2.857 / s)
1264.667000
0.1635
0
-0.252
5000 Hz
3982.8 Hz
0 Hz
300 Hz
156.97 Hz
75 Hz
France
Alain
Le
Gouguec
montre
que
pour
nous
[silence]
ça
00474m
m
1.553579
0.350000
0.613650
1262.763421
1262.763421
Visible part 2.517229 seconds
1265.280650
2334.623350
Total duration 3599.904000 seconds
all
in
out
sel
Group
Démarrer
02:16

Annotations (2)

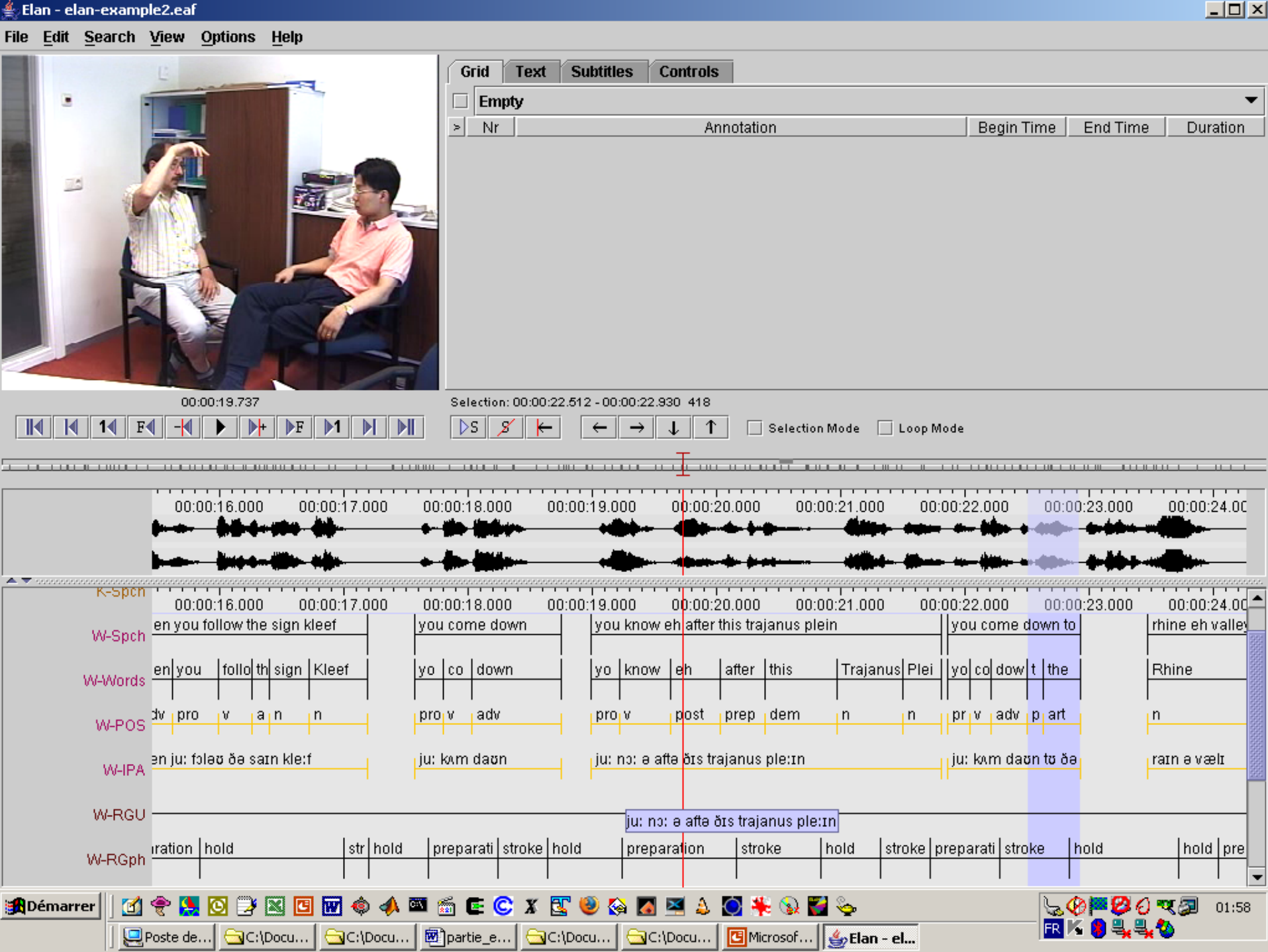
Transcriber : http://www.etca.fr/CTA/gip/Projets/Transcriber/IndexFr.html

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


Transcriber 1.4.2
File Edit Signal Segmentation Options Help
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...
... Patricia .
[i] ...
... .
le ...
..Tivolle :
[i] ...
... r !
mardi.
... .
la consultation ...
... lycées :
[i] grand débat...
... demain à Lyon
pour tirer..
... ents du
0
5
10
15
20
Cursor : 0
Démarrer
Poste de...
C:\Docu...
C:\Docu...
partie_e...
C:\Docu...
C:\Docu...
Microsof...
Elan - el...
Transcr...
01:59

Annotations (3)

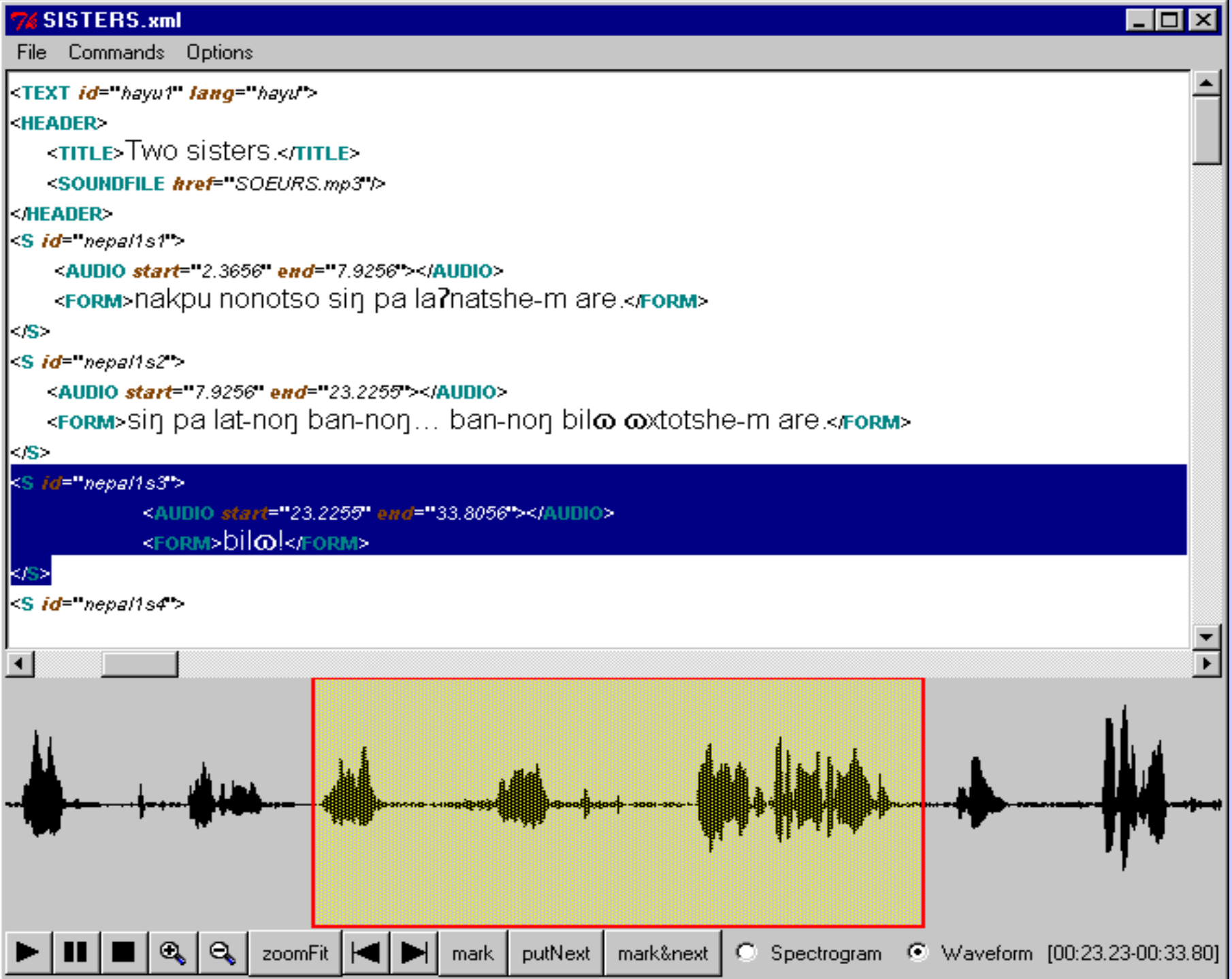
- ELAN : http://www.mpi.nl/tools/elan.html
  - sortie XML (dtd spécifique)
  - très complet (vidéo)
  - Plusieurs niveaux d'analyse possibles (niveaux dépendants/indépendants, nécessité de spécifier l'ancrage temporel pour chaque niveau).


Elan - elan-example2.eaf
File Edit Search View Options Help
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
Selection Mode
Loop Mode
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 00:00:24.00
W-Spch
W-Words
W-POS
W-IPA
W-RGU
W-RGph
en you follow the sign kleef
you come down
you know eh after this trajanus plein
you come down to
rhine eh valley
en you follo th sign Kleef
yo co down
yo know eh after this Trajanus Plei
yo co dow t the
Rhine
pro v adv
pro v post prep dem n n
pr v adv p art
n
ju: kʌm daʊn
ju: nɔ: ə aftə ðɪs trajanus pleːɪn
ju: kʌm daʊn tʊ ðə
raɪn ə vælɪ
preparation stroke hold stroke preparation stroke hold stroke hold
Démarrer
01:58

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…


SISTERS.xml
File Commands Options
<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 bilɷ ɷxtotshe-m are.</FORM>
</S>
<S id="nepal1s3">
<AUDIO start="23.2255" end="33.8056"></AUDIO>
<FORM>bilɷ!</FORM>
</S>
<S id="nepal1s4">
zoomFit
mark
putNext
mark&next
Spectrogram
Waveform
[00:23.23-00:33.80]

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
- Cédric Gendrot :
  cgendrot@univ-paris3.fr