



ASSISES DE LA RECHERCHE
Université Sorbonne Nouvelle (Paris 3)
2 et 3 mars 2006 - Paris

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

Serge Fleury, Cédric Gendrot, Michel Jacobson

Frédérique Bénard, Sonia Branca, Maria Candéa, Thierry Pagnier, Patrick Renaud, André Salem, Pollet Samvelian, Luigi Samsonetti, Jacqueline Vaissière.

Une base de données

par et pour

L'École Doctorale 268

- Une approche pluridisciplinaire, et dont l'objectif évident est une contribution à la **linguistique de corpus**.
- 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

Utilités d'une base de corpus

- Archivage de langues. (LACITO : <http://lacito.vjf.cnrs.fr/archivage/>)
- Analyse / synthèse de la parole. (ELDA / LDC)
- Favoriser les collaborations pluri-disciplinaires. (sur un même corpus)
- Aider les jeunes chercheurs pour la constitution de leurs corpus (but pédagogique – stratégie cumulative)

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 des bonnes pratiques ...

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

4- Exploitation, conservation, diffusion :

Notre travail !!

Le cœur du projet

- 1 L'élaboration d'une base de corpus (principalement oraux) regroupant des données de langues de différentes natures
→ Approche pluridisciplinaire
- 2 Des propositions de normalisation pour l'encodage de corpus de langue.

Exploitation, conservation, diffusion

- Nécessité de normaliser :

- pourquoi ?

Besoins de partager, diffuser et rechercher/retrouver

- Comment ? : Grâce à des métadonnées

XML → Dublin Core / OLAC

eXtensible
Markup Language

Open Language
Archive Community

Résultats visibles ...

Un Générateur de métadonnées : écrire de façon standardisée les données qui décrivent les données (ici les corpus)

il suffit de :

- cocher les cases
- compléter les espaces vides

Dublin Core	Raffinement DC	Attribut OLAC	Vocabulaire OLAC	METADONNEES
Title				<input checked="" type="checkbox"/>
Subject		discourse-type	drama	<input type="checkbox"/>
Subject		discourse-type	formulaic_discourse	<input type="checkbox"/>
Subject		discourse-type	language_play	<input type="checkbox"/>
Subject		discourse-type	interactive_discourse	<input type="checkbox"/>
Subject		discourse-type	oratory	<input type="checkbox"/>
Subject		discourse-type	narrative	<input type="checkbox"/>
Subject		discourse-type	procedural_discourse	<input type="checkbox"/>
Subject		discourse-type	report	<input type="checkbox"/>
Subject		discourse-type	singing	<input type="checkbox"/>
Subject		discourse-type	unintelligible_speech	<input type="checkbox"/>
Subject		language	2 letters -> ISO 639-1	<input type="checkbox"/>

Générateur de METADONNEES

ED268 [U. DE PARIS 3, Sorbonne nouvelle]

Projet Innovant 2004-2006 PIED268

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

un site Web avec :

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

- les travaux/outils
- les présentations officielles
- des liens

et bien sûr ...

les Données et le
MOTEUR de Recherche

PI-ED268 WEB PAGE : Constitution de ressources linguistiques normalisées - Mozilla Firefox

Echier Edition Affichage Aller à Marque-pages Outils ?

file:///C:/Documents%20and%20Settings/gendrot/Bureau/linguistique_de_corpus_gendrot/Linguistique_de_corp... OK

PI-ED268-0405 [U. Paris3, Sorbonne nouvelle]

Projet innovant ED268, "LANGAGE & LANGUES"
<http://ed268.univ-paris3.fr/>

- Accueil
- Catalogue
- BDD (état)
- BDD (Login)
- Bibliographie
- Contact
- Documents en cours (Login)
- Groupe
- Journées
- Liens
- Outils
- Réunions
- Travaux

Propositions de Normalisation pour une Base de Corpus Multimedia à l'ED268

Objectifs

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). Le projet vise à s'inscrire dans cette perspective de constitution de ressources électroniques normalisées dans le cadre des corpus de langue. Il s'attache à définir des perspectives de structuration de corpus en intégrant des marqueurs de strates dans les textes encodés pour donner à voir les textes sous ces différents facettes en parcourant en profondeur les strates définies.

Présentation du projet Septembre 2005, Bénard Frédérique

Normalisation de corpus oraux : des métadonnées à l'annotation de transcriptions, Maîtrise Sciences du Langage mention "Industries de la Langue", ILPGA, Université Paris 3, Sorbonne Nouvelle, Bénard Frédérique, soutenue le 21/09/2005.

Présentation du projet à mi-parcours, faite au cours des RJC-ED268 (Rencontre Jeunes Chercheurs ED268), faite le 21.05.2004

Résumé Soumission RJC-ED268, "Propositions de Normalisation pour une Base de Corpus Multimédia à l'ED268", Bénard Frédérique, Gendrot Cédric, 2005.

Article complet RJC-ED268, "Propositions de Normalisation pour une Base de Corpus Multimédia à l'ED268", Bénard Frédérique, Gendrot Cédric, 2005.

Transparents de Présentation du projet, 10.09.2004

Actualités

Dans le cadre de la Journée d'étude : "Constitution, exploitation, diffusion et conservation des corpus oraux", Mardi 17 mai 2005, BnF
URL : http://www.culture.gouv.fr/culture/dglf/corpus_oraux.htm
Publication d'un "Guide des bonnes pratiques pour la constitution, l'exploitation, la diffusion et la

Terminé

et bien sûr ...

les Données et le MOTEUR de Recherche

Taper un mot-clé pour rechercher des corpus

XPath search - Moteur pour le catalogue de METADONNES du projet INNOVANT PIED268 - Mozilla Firefox

Fichier Edition Affichage Aller à Marque-pages Outils ?

file:///C:/Documents%20and%20Settings/gendrot/Bureau/assises_de_la_rechercl OK

PI-ED268 [ED268 & U. DE PARIS 3, Sorbonne nouvelle]

Moteur pour le catalogue de métadonnées du Projet Innovant PI-ED268

(compatible IE5+/Mozilla/Firefox, non compatible avec Safari/IE5-mac)

Chargez **le (CATALOGUE)** avant de lancer des requêtes...

Choisissez une métadonnée à afficher (requête XPath) ou bien rechercher toutes les métadonnées contenant :

choisissez la métadonnée à afficher...

Requête XPath utilisée :

Ce moteur de requête reprend une application développée par John Udell (<http://udell.roninhouse.com>) permettant de sélectionner et de récupérer les rubriques de son weblog classées thématiquement via l'utilisation d'une interface web utilisant des requêtes XPath ([http://udell.infoworld.com:8000/?//blockquote\[@cite='infoWorld'\]](http://udell.infoworld.com:8000/?//blockquote[@cite='infoWorld'])). Dans la version proposée ici, on peut avoir accès au contenu des métadonnées de l'ensemble des ressources du catalogue du corpus construit par le projet PIED268.

Mode d'emploi : (1) Il faut commencer par charger le catalogue (*cf* lien *supra*), (2) on peut ensuite lancer des requêtes : **deux types de requêtes sont possibles.** Soit on sélectionne une métadonnées dans la liste de choix proposée et dans ce cas, on obtiendra comme résultat la description de l'ensemble des métadonnées visées, soit on entre une chaîne de caractère à rechercher (dans la zone de saisie à droite) dans les différents contenus des métadonnées disponibles, on obtiendra en sortie l'affichage des métadonnées contenant la

Terminé

En résumé :

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

en cours ...

- Amélioration du moteur de recherche ...
- Candidature acceptée aux centres de ressources numériques proposés par le C.N.R.S.

dans un futur proche ...

- **extraire et/ou afficher des occurrences prononcées ...**

cela nécessite :

- un moteur de recherche qui travaille sur l'annotation et non les métadonnées (facile)
- que les annotations soient complètes et compatibles (+ difficile)

* cours Linguistique de corpus créé pour le LMD

Merci de votre attention

Contacts et Liens ...

- Site Web :

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

- Serge Fleury : serge.fleury@univ-paris3.fr
- Cédric Gendrot : cgendrot@univ-paris3.fr