LESCLaP
Laboratoires d'Etudes Sociolinguistiques sur les Contacts de Langues et la Politique Linguistique

J-M. Eloy, C. Rey

LESCLAP (CERCLL)

Université de Picardie Jules Verne

jean-michel.eloy@u-picardie.fr

christophe.rey@u-picardie.fr

Journée d'étude du LESCLAP

07 décembre 2012

Amiens

Une base de données lexicale en picard : la base PICARTEXT

UNIVERSITÉ de Picardie Jules Verne

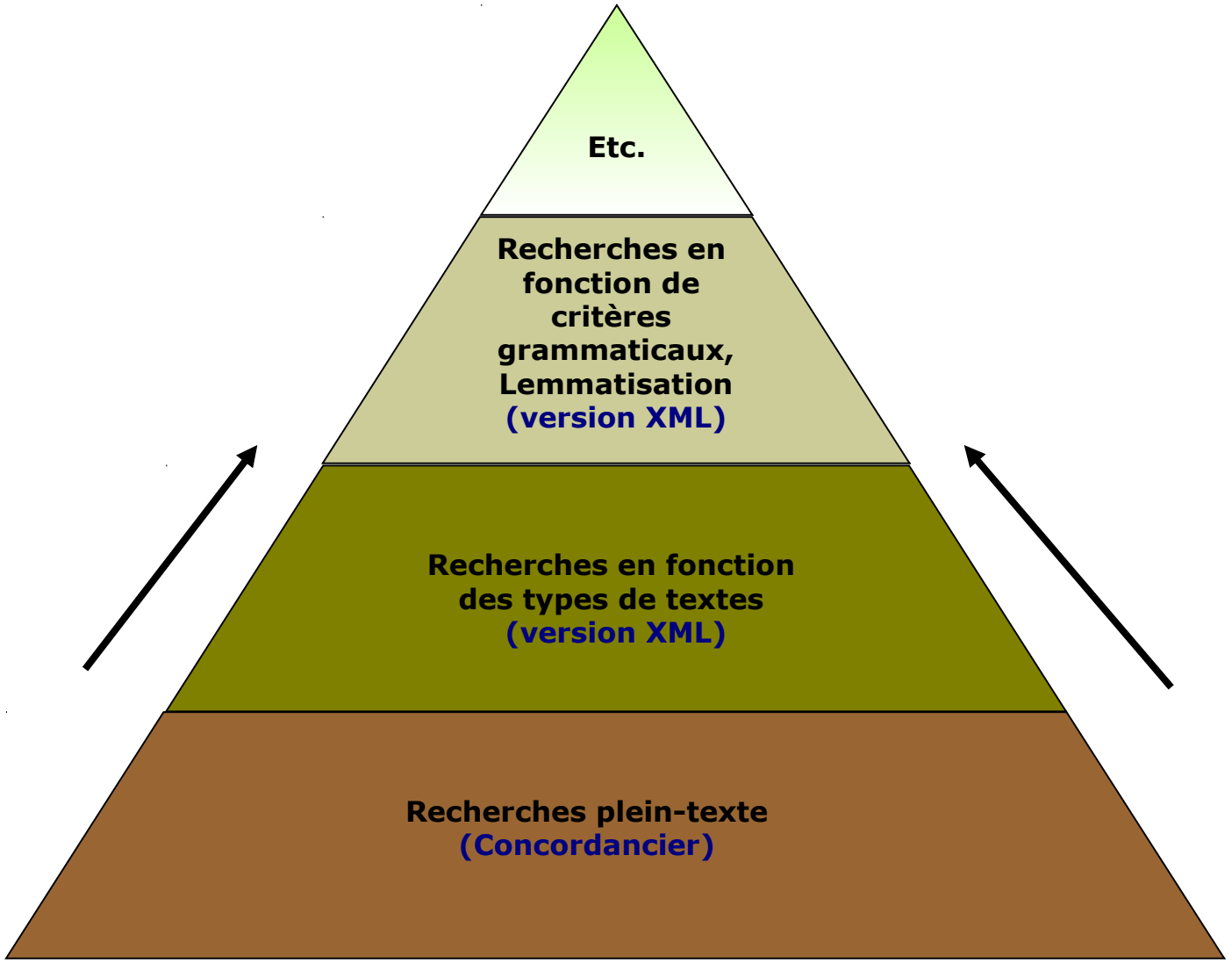
PICARDIE LA RÉGION

Exploitation de la base Picartext

L'avenir

Disponible

Disponible



Quelques informations techniques

2 profils de post-doctorant

/

2 phases techniques différentes

Une première base de données MySQL

Accessible (avec identification) à l'adresse : http ://www.u-picardie.fr/LESCLaP/PICARTEXT.

Les deux tables principales de la base dont :

'registre' : contient les informations de type bibliographique sur chaque document

'auteur' : contient les informations de type biographique sur chaque auteur

Ces deux tables sont mises en relation à travers la table 'titre_auteur', qui permet d'associer à chaque document un ou plusieurs auteurs.

Deux autres tables sont importantes :

'inventaire' : liée à 'registre', sert à décrire les traitements reçus par chaque document et son étape actuelle

'auteur_lieux' : permet d'associer à chaque auteur des lieux de vie successifs au cours de sa vie

Les lieux dans 'auteur' et 'auteur_lieux' sont codés selon les nomenclatures de l'Insee (complétées pour la Belgique) et organisés selon la hiérarchie commune

< canton < arrondissement < département < région 6.

Balisage XML

Le balisage des textes a été prévu selon le standard TEI (The Text Encoding Initiative) en deux étapes :

Etape 1 : **ajout de l'entête**,

Etape 2 : **balisage du corps des texte**.

Une entête Picartext a été conçue, conforme au standard TEI. Le système de gestion comporte un générateur automatique d'entête qui extrait les données de la base MySQL et les formate automatiquement afin de pouvoir les insérer à la demande dans les textes.

L'étape 2 reste à concevoir. La base Picartext pourra être considérée comme entièrement opérationnelle lorsque l'ensemble des textes aura passé cette étape.

De nouveaux traitements

A- Mise en place d'un traitement automatique des apostrophes dans les textes du corpus (nouvelle segmentation en « mot ».

B- Indexation d'un sous-corpus (tables MySQL)

C- Développement d'un module expérimental de recherche par correspondances dialectales

* indexation phonétique du corpus (système TTS-French développé par David Haubensack). A nécessité une adaptation des règles de base aux règles spécifiques aux picard :

Ex : neutralisation des voyelles moyennes : réinterprétation systématique en /e/, /o/, /ø/ en toutes positions, y compris en syllabe fermée (notation phonologique plutôt que phonétique)

- interprétation de tous les e muets comme [zéro]

- interprétation de o + voyelle en /u/ (=[w]) dans la notation des diphtongues ascendantes picardes : oé , oai , etc.

- ajout de l'apostrophe à la liste des voyelles (dont elle partage le comportement combinatoire en picard)

- prise en compte de la possibilité de l'omission de l'accent sur é , è et réinterprétation en [e], en début de mot et devant consonne (mere pour mère)

- correction de divers cas où l'interprétation selon les règles du français est impropre en picard. Ex. : la graphie tieu à interpréter [tjø] en picard (catieu), et non [sjø] comme en français (cf. ambitieux).