Devoir à rendre pour le dernier cours de décembre (mercredi 20/12/2017) au format électronique clé USB ou email: pap@limsi.fr, avec sujet: "INALCO devoir L3"

Sujet:

Réaliser une analyse linguistique sur le corpus des cours en ligne de l'année dernière. http://perso.limsi.fr/pap/inalco/…

Pour cela il faudra télécharger les cours en pdf, les transcrire en texte (par ex. pdftotext), normaliser les documents, les mettre au format XML (création d'une DTD), calculer les différentes distributions, caractères, mots, etc. au moyen de programmes écrits en python. Vous pourrez aussi réaliser un étiquettage des mots avec les étiquettes associées aux mots d'un dictionnaire disponible à l'URL: ...à préciser...

Puis faire une analyse linguistique (si possible diachronique) à au moins 2 niveaux (par fichier et globalement).

Les programmes seront accompagné d'un rapport expliquant les différentes étapes du projet, les problèmes rencontrés et les solutions trouvées.

En fonction de vos compétences informatiques et linguistiques vous pouvez développer plus certains aspects et aller aussi loin que vous voulez/pouvez dans la réalistion (de qqs scripts à un environnement de traitement de corpus avec une interface graphique…).

Le devois sera noté sur la cohérence, l'homogénéité, la complétude, la qualité d'analyse, la qualité de programmation et la qualité de rédaction.