Procédures numériques pour la sémantique historique

Plan

1. Préalables
2. Construction d'un corpus indexé
3. Explorations sémantiques

1. A. Objectif

- Aspects pratiques seulement
- Public visé : débutants avancés (!?)
- Résumer en 2h la matière d'un cours annuel (au moins) → un survol...

1.B. Un nouveau système technique

- « Système technique » au sens de Bertrand Gille : nombreux outils et techniques formant un ensemble
- Système qui s'est développé en un temps éclair : démarrage réel début des années 80 = moins de 30 ans
- énormes écarts générationnels → blocages
- Matériels
- Internet
- Procédures > logiciels
- Un potentiel + **inédit** + **gigantesque**

1.C. Domaine public et copyfraud

- Principe universel ultra-simple : 70 ans après la mort de l'auteur. Auteur = créateur
- Les philologues n'ont aucun droit sur les textes qu'ils éditent
- Les éditeurs commerciaux n'ont aucun droit sur les textes anciens qu'ils publient
- L'ignorance, l'irréflexion, la lâcheté des universitaires facilitent le copyfraud
- Il faut rappeler sans relâche que **tous les textes anciens sont dans le domaine public, sans exception.**

1.D. Priorité absolue à l'open-source

- Science = échange, discussion réglée, coopération ; la science est l'opposé du privilège, de la concurrence et de l'affrontement
- Utilisateurs potentiels : désargentés
- Possibilité de vérification des résultats
- Amélioration bien plus rapide des outils
- Aucun verrouillage hiérarchique

1.E. Ça ne marche pas tout seul...

- Une exploration ne peut pas être fondée sur une routine, si bien conçue soit-elle : il existe de bons outils, mais toute recherche réelle implique d'imaginer des procédures ad hoc : on ne peut pas faire l'économie d'un minimum de programmation
- Les nouvelles procédures ne viennent pas à la place des anciennes, elles viennent **en plus** : pas de sémantique historique sans de solides connaissances philologiques et historiques

1.F. adresses

- www.glossaria.eu
- guerreau@msh-paris.fr
bruno.bon@irht.cnrs.fr
renaud.alexandre@irht.cnrs.fr
krzysztofn@ijp-pan.krakow.pl

2.A. Corpus indexé : définition

- Un ensemble de textes ayant « quelque chose » en commun
- Formalisés
- Intégrés dans une base de données gérée par un logiciel d'indexation et d'interrogation

2.B. Corpus indexé : contraintes

- Pas de corpus, pas d'analyse statistique formalisée
- Corpus mal conçu : résultats insignifiants et/ou biaisés
- Corpus de taille insuffisante : restriction des possibilités de recherche
- Corpus insuffisamment formalisé : résultats pauvres
- Construire un corpus historique : beaucoup de travail : bien calculer son affaire avant de commencer !

2.C. Récupérer des fichiers existants

- Sites nombreux, en essor, mais très mal repérés par les moteurs de recherche : chercher avec ténacité !!
- Conditions de récupération/téléchargement varient du tout au tout ; en général pas très commode (scripts ad hoc souvent utiles)
- Formats et encodages très variés, souvent exotiques : important travail de mise en ordre de base (tout recoder en utf-8, sans caractères bizarres)

2.D. Numériser des textes imprimés

- Possibilité d'OCR dépend de la qualité des images : type d'impression, précision du scan
- Logiciels libres aujourd'hui au meilleur niveau : tesseract + gImagereader + fichier de paramètres par langue (latin depuis juin 2015)
- Correction nécessaire : utiliser myspell/hunspell dans un éditeur (geany) ; récupérer les fichiers .dic+.aff appropriés (pour le latin : http://extensions.libreoffice.org/extension-center/r/latin-spelling-and-hyphenation-dictionaries pas idéal, mais rapide et efficace ; c'est un fichier zip, récupérer les fichiers .dic et .aff)

2.E. Prétraitement 1 : nettoyage

- Tout ramener en utf-8
- Supprimer tous les éléments adventices
- Normalisation au moins partielles des graphies, si nécessaire

2.F. Prétraitement 2 : tokenisation

- Découpage en « token » = unité minimale
- Peut correspondre à plusieurs « mots »
- Problème des enclitiques
- Problème des abréviations
- Problème des mots composés (allemand)
- Scripts passe-partout très insuffisants

2.G. Prétraitement : étiquetage

- Alias : pos-tagging
- Combien de POS ?
- Outil automatique : **tree-tagger** le plus efficace aujourd'hui, même si ne dépasse pas 97,5% de succès
- Fichiers de paramètres pour beaucoup de langues, pas toutes (moyen français, Mittelhochdeutsch)
- Latin : paramètres **OMNIA**

2.H. Prétraitement : les textes

- Catégoriser les textes d'un corpus selon plusieurs critères : auteur, date/période, type de texte, localisation……
- L'efficacité des analyses historiques dépend largement de la pertinence des catégories choisies
- Nombreuses questions non-résolues : granularité des catégories de date, de lieu, de type (problème de la hiérarchie ou des chevauchements)

2.I. Construction

- Choix d'un moteur d'indexation ; relativement limité : mysql, lucene, open-cwb
- Efficacité d'open-cwb : puissant, très rapide, langage de questionnement très riche (spécialement conçu pour interrogations lexicales et linguistiques)
- Implique un formatage spécifique des fichiers en entrée

2.J. Logiciels d'indexation-lecture

- Reposant sur open-cwb :
  > **cqp** : en ligne de commande, efficace et rapide
  > **TXM** : logiciel lyonnais conçu pour une extrême facilité d'installation et d'utilisation ; très commode ; en plein essor, forum et liste de discussion très actifs
  > **cqpweb** : fonctionnalités voisines de TXM, pas de limite de taille de corpus, utilisable pour mise en ligne de corpus ; installation beaucoup moins simple ; liste de discussion très réactive

2.H. Construire un corpus : remarques générales

- Opération préalable
- Opération en générale longue et laborieuse
- Si chercheur en début de recherche : bien calculer en termes de temps disponible / résultats souhaités, attention cependant à la taille
- Si groupe institutionnel : bien définir au départ l'ensemble du processus, prévoir au moins une modalité de téléchargement simple (serveur ftp si nécessaire!)

QUESTIONS ?