cqpweb:
applications pédagogiques et recherche en humanités numériques

Graham Ranger, Journée DARIAH. UAPV, 4 décembre 2017.

cqpweb

Introduction : méthodes et techniques

Travail linguistique sur du texte à la recherche de régularités généralement par le biais d'un corpus

Similitudes méthodologiques avec d'autres domaines (différences portant sur les problématiques et les finalités)

cqpweb

Introduction : méthodes et techniques

Travail mené :

- avec un corpus déjà constitué, souvent via un dispositif en ligne
- avec un corpus constitué ad hoc, préparé par le linguiste dans une finalité précise, exploité avec un dispositif local

cqpweb

Introduction : méthodes et techniques

Selon la première approche on pourra travailler à partir de sites comme :

https://corpus.byu.edu/

https://www.sketchengine.co.uk/

http://bncweb.lancs.ac.uk

Chaque site regroupe un ou plusieurs corpus, interrogeables via une application en ligne dédiée.

cqpweb

Introduction : méthodes et techniques

Selon la deuxième on constituera un corpus ad hoc, unique, éventuellement à l'aide d'outils comme BootCat http://bootcat.dipintra.it/

Puis on l'interrogera avec un concordancier tel que AntConc http://www.laurenceanthony.net/

ou en ligne de commande avec cwb http://cwb.sourceforge.net/ ou R https://www.r-project.org/, par exemple.

cqpweb

Introduction : méthodes et techniques

La première approche est :

- Abordable sans formation spécifique
- Gratuit, disponible pour tous et donc vérifiable

Mais :

- Le chercheur n’a pas de contrôle sur le corpus
- Ni sur les modalités d’interrogation

cqpweb

Introduction : méthodes et techniques

La deuxième approche permet de contrôler le corpus (structuration et exploitation).

Mais

- Elle ne se prête qu'aux corpus relativement petits.
- Les résultats sont difficiles à reproduire indépendamment.
- Une certaine connaissance informatique est parfois nécessaire.

cqpweb

Introduction : méthodes et techniques

L'outil cqpweb est une interface d'interrogation en ligne – avantages d'ergonomie, de disponibilité et de puissance de calcul – derrière laquelle on peut mettre son propre corpus – avec une préparation au préalable.

Il représente une interface graphique de cwb : corpus workbench.

Les deux sont des applications libre source : http://cwb.sourceforge.net/cqpweb.php

cqpweb

Plan :

Présentation et historique de cqpweb

Le projet Molière :

- Structuration
- Exploitation
- Installation
- Ouvertures pédagogiques et interdisciplinaires

cqpweb

Présentation et historique de cqpweb

- cqp : corpus query processor
- web : en ligne

Au départ : British National Corpus

L’un des premiers megacorpus, constitué de 100 m de mots d’anglais britannique dans les années 1990 avec plus de 4000 textes différents, catégorisés par type, genre, source, etc.

cqpweb

Présentation et historique de cqpweb

Le British National Corpus :

- D’abord un corpus privé, payant
- Puis interrogeable via différentes interfaces
- Puis gratuit et entièrement téléchargeable

cqpweb

Présentation et historique de cqpweb

Le British National Corpus :

L'une des interfaces du BNC s'appelle BNCweb : http://bncweb.lancs.ac.uk/ (Hoffmann et al 2008.)

Interface d'interrogation écrite en perl.

Avantages : souplesse et puissance : utilisation de cqp corpus query processor, notamment.

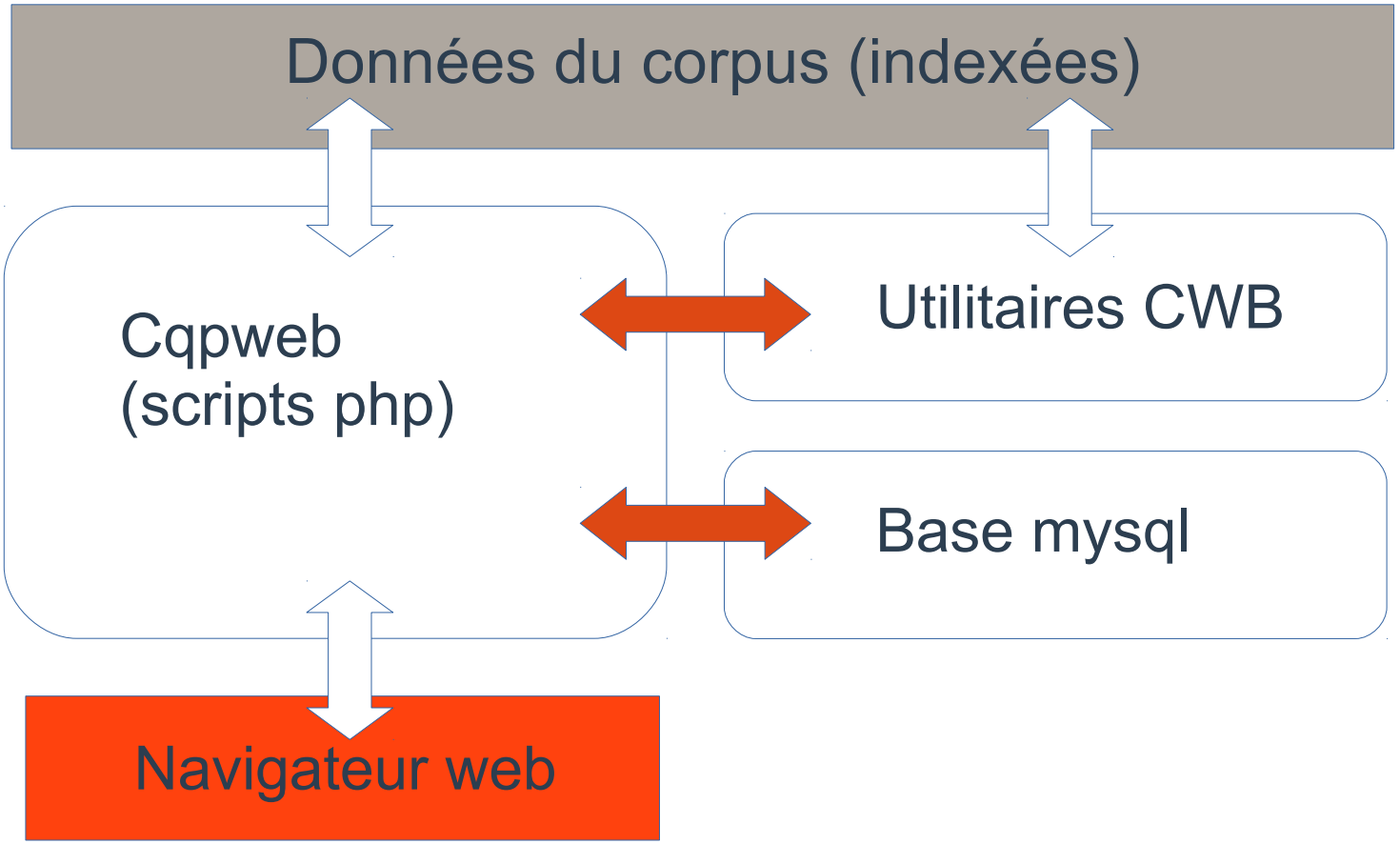
Inconvénients : lié par son architecture à un seul corpus.

cqpweb

Présentation et historique de cqpweb

Cqpweb représente une réécriture de BNCweb, en php, qui pourra servir d'interface pour n'importe quel corpus, à condition qu’il soit convenablement préparé (indexé).

Présentation et historique de cqpweb



cqpweb

Présentation et historique de cqpweb

Une fois installé, cqpweb permet des requêtes :

- simples : mots, chaînes de mots ou caractères.
- par POS ou par lemme, caractères génériques.
- portant sur la catégorisation des textes du corpus.
- portant sur le balisage xml du corpus.

cqpweb

Présentation et historique de cqpweb

Ces possibilités d’exploitation impliquent, en amont, un travail de structuration, très pertinent sur le plan pédagogique.

cqpweb

Structuration

Il est décidé de travailler sur un corpus théâtral :

- Pour des raisons de pertinence géographique
- Pour des raisons liées à la transférabilité des problématiques soulevées par la structuration des « textes de performance »

cqpweb

Structuration

Les textes au départ : corpus de pièces de théâtre de Molière, disponible en ligne sur wikisource :

https://fr.wikisource.org/wiki/Auteur:Moli%C3%A8re

Ces textes sont affichés en html qui balise en fonction d'un *affichage* recherché.

Il faut désormais les baliser en fonction de leur *structuration*.

cqpweb

Structuration

Choix de balises : un jeu de balises xml tiré de celles du Text Encoding Initiative http://www.tei-c.org/ mais limité dans un souci d'efficacité selon les recommandations du développeur de cqpweb :

"Modest XML for corpora : not a standard, but a suggestion." Hardie, Andrew. 2014.

cqpweb

Structuration

<text id=text0001> … </text>

<title> … </title>

<div1 type="act" n="1"> … </div1>

<set> … </set>

<sp who="SGANARELLE"> … </sp>

<stage> … </stage>, etc.

cqpweb

Structuration

Les textes sont récupérés sur le site wikisource, puis sauvegardés en plain text, UTF-8 (html inexploitable).

Structuration

ACTE PREMIER.
Scène I
CHRYSALDE, ARNOLPHE

Chrysalde.
Vous venez, dites-vous, pour lui donner la main ?

Arnolphe.
Oui, je veux terminer la chose dans demain.

Chrysalde.
Nous sommes ici seuls ; et l'on peut, ce me semble,
Sans craindre d'être ouïs, y discourir ensemble.

Arnolphe.
Il est vrai, notre ami. Peut-être que chez vous
Vous trouvez des sujets de craindre pour chez nous ;
Et votre front, je crois, veut que du mariage
Les cornes soient partout l'infaillible apanage.

cqpweb

Structuration

Les balises sont insérées de manière semi-automatique, selon les régularités du texte de départ, par rechercher / remplacer avec expressions régulières.

cqpweb

Structuration

Exemples :

Rechercher : (\nScène (\d?)\n)

Remplacer : </div2><div2 type="scene" n="\2">\1

Rechercher : \n(Arnolphe)\n

Remplacer par : \n<sp who="ARNOLPHE">\n<speaker>\n\1\n</speaker>\n

cqpweb

Structuration

Puis relecture et vérification de l'intégrité de la structuration à l'aide d'une application telle que XML Copy Editor : http://xml-copy-editor.sourceforge.net/

Structuration

<div1 type="act" n="1">ACTE I
<div2 type="scene" n="1">Scène 1
<set>CHRYSALDE, ARNOLPHE</set>
<sp who="CHRYSALDE"><speaker>Chrysalde</speaker>
Vous venez, dites-vous, pour lui donner la main ? </sp>
<sp who="ARNOLPHE"><speaker>Arnolphe</speaker>
Oui, je veux terminer la chose dans demain. </sp>
<sp who="CHRYSALDE"> <speaker>Chrysalde</speaker>
Nous sommes ici seuls ; et l'on peut, ce me semble,
Sans craindre d'être ouïs, y discourir ensemble.
Voulez-vous qu'en ami je vous ouvre mon cœur ?
Votre dessein pour vous me fait trembler de peur ;
Et de quelque façon que vous tourniez l'affaire,
Prendre femme est à vous un coup bien téméraire. </sp>
<sp who="ARNOLPHE"> <speaker>Arnolphe</speaker>
Il est vrai, notre ami. Peut-être que chez vous
Vous trouvez des sujets de craindre pour chez nous ;
Et votre front, je crois, veut que du mariage
Les cornes soient partout l'infaillible apanage.</sp>

cqpweb

Structuration

Ensuite, annotation à l'aide de TreeTagger : étiquetage des catégories grammaticales, lemmatisation et verticalisation.

http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

Ceci se fait en ligne de commande, ou éventuellement via une interface web.

Structuration

Nous	PRO:PER	nous
sommes	VER:pres	sommer|être
ici	ADV	ici
seuls	ADJ	seul
,	PUN	,
et	KON	et
l'	PRO:PER	la|le
on	PRO:PER	on
peut	VER:pres	pouvoir
,	PUN	,
ce	PRO:DEM	ce
me	PRO:PER	me

cqpweb

Structuration

Et, dans la mesure du possible, nettoyage et vérification :

seroit VER:futu <unknown> →

seroit VER:cond être

Ouais VER:impf <unknown> →

ouais INT oui

Structuration

Choix de catégories textuelles (déterminant pour définir un sous-corpus) et préparation d'un fichier de métadonnées en format .csv :

Text id	Titre	Auteur	Date
l_ecole_des_femmes	L_Ecole_des_Femmes	Moliere	1662
les_facheux	Les_facheux	Moliere	1661
les_precieux	Les_Precieuses_ridicules	Moliere	1659

cqpweb

Installation sur http://cqpweb.univ-avignon.fr/

Phase technique qui ne sera pas développée aujourd'hui.

Documentation ici :

http://cwb.sourceforge.net/files/CQPwebAdminManual.pdf

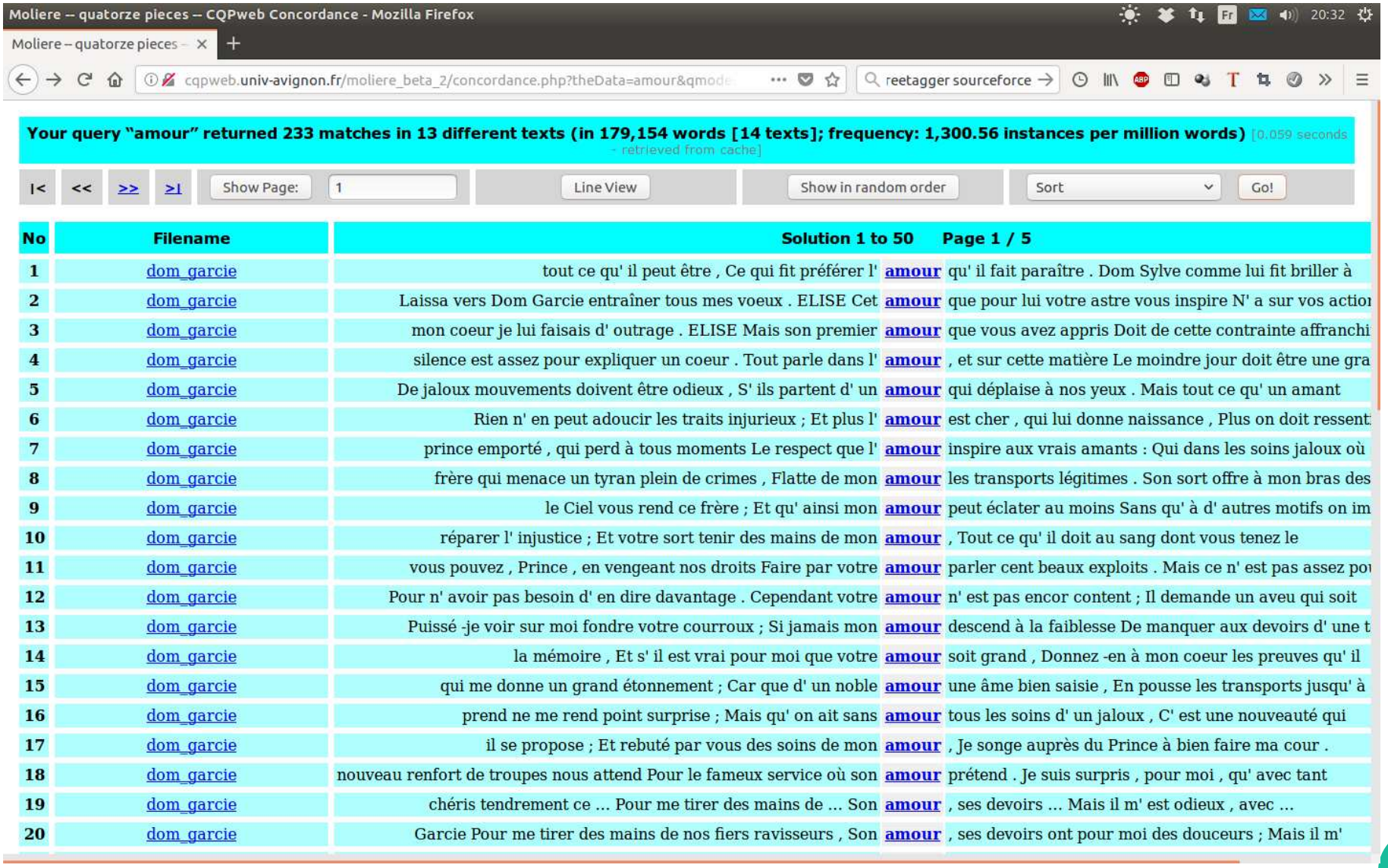
cqpweb

Exploitation

Requêtes simples :

amour : 233 occurrences / 13 textes / 179.154 mots

cqpweb



No	Filename	Solution 1 to 50 Page 1 / 5		
1	dom_garcie	tout ce qu' il peut être , Ce qui fit préférer l'	amour	qu' il fait paraître . Dom Sylve comme lui fit briller à
2	dom_garcie	Laissa vers Dom Garcie entraîner tous mes voeux . ELISE Cet	amour	que pour lui votre astre vous inspire N' a sur vos actio
3	dom_garcie	mon coeur je lui faisais d' outrage . ELISE Mais son premier	amour	que vous avez appris Doit de cette contrainte affranchi
4	dom_garcie	silence est assez pour expliquer un coeur . Tout parle dans l'	amour	, et sur cette matière Le moindre jour doit être une gra
5	dom_garcie	De jaloux mouvements doivent être odieux , S' ils partent d' un	amour	qui déplaise à nos yeux . Mais tout ce qu' un amant
6	dom_garcie	Rien n' en peut adoucir les traits injurieux ; Et plus l'	amour	est cher , qui lui donne naissance , Plus on doit ressent
7	dom_garcie	prince emporté , qui perd à tous moments Le respect que l'	amour	inspire aux vrais amants : Qui dans les soins jaloux où
8	dom_garcie	frère qui menace un tyran plein de crimes , Flatte de mon	amour	les transports légitimes . Son sort offre à mon bras des
9	dom_garcie	le Ciel vous rend ce frère ; Et qu' ainsi mon	amour	peut éclater au moins Sans qu' à d' autres motifs on im
10	dom_garcie	réparer l' injustice ; Et votre sort tenir des mains de mon	amour	, Tout ce qu' il doit au sang dont vous tenez le
11	dom_garcie	vous pouvez , Prince , en vengeant nos droits Faire par votre	amour	parler cent beaux exploits . Mais ce n' est pas assez po
12	dom_garcie	Pour n' avoir pas besoin d' en dire davantage . Cependant votre	amour	n' est pas encor content ; Il demande un aveu qui soit
13	dom_garcie	Puissé -je voir sur moi fondre votre courroux ; Si jamais mon	amour	descend à la faiblesse De manquer aux devoirs d' une t
14	dom_garcie	la mémoire , Et s' il est vrai pour moi que votre	amour	soit grand , Donnez -en à mon coeur les preuves qu' il
15	dom_garcie	qui me donne un grand étonnement ; Car que d' un noble	amour	une âme bien saisie , En pousse les transports jusqu' à
16	dom_garcie	prend ne me rend point surprise ; Mais qu' on ait sans	amour	tous les soins d' un jaloux , C' est une nouveauté qui
17	dom_garcie	il se propose ; Et rebuté par vous des soins de mon	amour	, Je songe auprès du Prince à bien faire ma cour .
18	dom_garcie	nouveau renfort de troupes nous attend Pour le fameux service où son	amour	prétend . Je suis surpris , pour moi , qu' avec tant
19	dom_garcie	chéris tendrement ce ... Pour me tirer des mains de ... Son	amour	, ses devoirs ... Mais il m' est odieux , avec ...
20	dom_garcie	Garcie Pour me tirer des mains de nos fiers ravisseurs , Son	amour	, ses devoirs ont pour moi des douceurs ; Mais il m'

cqpweb

Exploitation

Requêtes simples : amour

- KWIC view : "mot clé en contexte"
- Noter l’étiquetage du texte
- Line view…
- Show in random order…
- Distribution…
- Categorise…

cqpweb

Exploitation

Requêtes par lemme :

{aimer} > toutes les formes du lemme

Requêtes avec métacaractères :

amour+ > amour, amours, amoureux, amoureuse…

ser?it, ser[a,o]it > serait ou seroit

ser*it > serait, seroit mais aussi servait, seraient…

ser[a,o]i[,en]t > serait, seroit, seraient, seroient

cqpweb

Exploitation

Requêtes par catégorie grammaticale :

sort_NOM > *sort* en tant que nom

_VER:impf > toutes les formes étiquetées *imparfait*

*ment_ADV > tous les adverbes se terminant en *ment*

cqpweb

Exploitation

Délimitation du corpus à certains textes seulement, par l'option *Restricted query* qui fait appel aux catégories décidées auparavant.

cqpweb

Exploitation

Utilisation plus poussée de requêtes formulées avec la syntaxe cqp :

<sp_who="ARNOLPHE"> []+ </sp_who>

Toutes les répliques d'*Arnolphe* (choisir Line view)

n:[word="coeur"] :: n.text_id="l_ecole_des_maris" & n.sp_who="SGANARELLE" & n.div1_type="act" & n.div1_n="2";

Occurrences de *coeur* dans *L'école des maris* prononcées par Sganarelle dans le deuxième acte…

cqpweb

Exploitation

Les résultats peuvent être téléchargés en format texte, en vue d'une exploitation ultérieure (graphiques), catégorisés avec des catégories *ad hoc* déterminées par l'utilisateur, etc.

cqpweb

Ouvertures pédagogiques et interdisciplinaires

Sur le plan pédagogique cet exercice favorise chez l'étudiant :

- Une réflexion sur le formatage de texte, et une connaissance des bases d'html, de xml qui peuvent favoriser une construction rigoureuse par la suite ;
- Une réflexion sur les balises les plus pertinentes selon les textes et selon les objectifs et donc sur la structuration de textes de nature hétérogène ;

cqpweb

Ouvertures pédagogiques et interdisciplinaires

- Une réflexion sur la nature des catégories grammaticales, lexicales, via l'utilisation de l'étiquetage et les questions qu'il pose ;
- Une réflexion sur les catégories extratextuelles – auteur, date, etc. pertinentes pour la recherche ;
- Un vivier d'exemples ou une base textuelle pour la recherche ultérieure : ressource pérenne.

cqpweb

Ouvertures pédagogiques et interdisciplinaires

Sur le plan interdisciplinaire on voit que le même type d'interrogation pourra se poser dès lors qu'on a affaire à des objets textuels hétérogènes.

L'exemple du texte de théâtre est particulièrement intéressant, puisque ce type de structuration se retrouve dès lors qu'on a affaire à de la représentation, au sens large.

cqpweb

Ouvertures pédagogiques et interdisciplinaires

Un travail de master déjà entamé consiste à structurer des discours politiques de la même manière, en tenant compte des dates, des intervenants, des événements, des éléments extralinguistiques (applaudissements, huées), etc.

cqpweb

Ouvertures pédagogiques et interdisciplinaires

Ainsi, en substituant *Sarkozy* ou *Trump* à de *Sganarelle* ou *Tartuffe*, par exemple, dans les requêtes mentionnées plus haut, on pourra rechercher toutes les interventions (ou toutes les occurrences de tel lexème, tous les noms, adjectifs, verbes…) de tel personnage politique pendant telle période, prononcées devant tel type de public, etc.

cqpweb

Ouvertures pédagogiques et interdisciplinaires

Enfin, un corpus peut être déclaré *caché* ou *visible* et l'accès au corpus peut être finement paramétré pour le restreindre à une classe d'utilisateurs, à un groupe de recherche, à des collègues collaborant sur un projet en commun, à un seul chercheur…

cqpweb

Quelques références :

Evert, Stefan & Andrew Hardie. 2011. "Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium." Presentation at *Corpus Linguistics 2011*, University of Birmingham, UK.

Hardie, Andrew. 2012. "CQPweb — combining power, flexibility and usability in a corpus analysis tool." International Journal of Corpus Linguistics 17(3). 380–409. doi:10.1075/ijcl.17.3.04har.

Hardie, Andrew. 2014. "Modest XML for Corpora: Not a standard, but a suggestion." ICAME Journal 38(1). doi:10.2478/icame-2014-0004.

Hoffmann, Sebastian, Stefan Evert, Nicholas Smith & David Lee. 2008. *Corpus linguistics with BNCweb: a practical guide*. (English Corpus Linguistics v. 6). Frankfurt am Main: Peter Lang.

Helmut Schmid. 1995. "Improvements in Part-of-Speech Tagging with an Application to German." Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Helmut Schmid. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." Proceedings of International Conference on New Methods in Language Processing, Manchester, UK.

cqpweb

D’autres installations de cqpweb (liste non exhaustive) :

http://cqpweb.lancs.ac.uk/

http://corpora.clarin-d.uni-saarland.de/cqpweb/

http://cwb-test.eila.univ-paris-diderot.fr/ims/index.php

Et l’ancêtre :

http://bncweb.lancs.ac.uk/

cqpweb

Ressources logicielles utilisées pour le projet :

Textes de départ :

https://fr.wikisource.org/

Mise en forme :

https://notepad-plus-plus.org/fr/

https://wiki.gnome.org/Apps/Gedit

XML Copy Editor :

http://xml-copy-editor.sourceforge.net/

TreeTagger :

http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/

http://corpora.lancs.ac.uk/tree-tagger/

Cqpweb et corpus workbench :

https://sourceforge.net/projects/cwb