Sens Texte Informatique Histoire

ceres Centre d'Expérimentation en méthodes numériques pour les REcherches en Sciences humaines et sociales

LETTRES SORBONNE UNIVERSITÉ

Europresse : Constitution et Exploitation de Corpus

Thibault Grison et Gaël Lejeune

30 septembre 2021

STIH, Sorbonne Université

La problématique

1. Constituer des corpus homogènes
2. satisfaisant des critères de qualité
3. utilisables avec des outils informatiques
4. (corollaire : une certaine taille)

La problématique

1. Constituer des corpus homogènes
2. satisfaisant des critères de qualité
3. utilisables avec des outils informatiques
4. (corollaire : une certaine taille)

Table 1 – Quelques solutions existantes (`fr`)

Corpus	Homog.	Qual.	Util.	Taille
Est républicain (Ortolang)	Limité temporellement	+++	+	++
Web As Corpus (Sketch Engine)	Tout venant	+	++	+++
Google News	OK	++	–	++
Europresse	++	+++	???	+++

Europresse So What ?

Au delà du côté Moteur de recherche, comment exploiter dans un concordancier par exemple ? ou conserver les corpus pour des exploitations futures ?

Limites rencontrées :

1. résultats par tranche de 100 documents jusqu'à une limite de 1 000
2. format peu adapté à la requête plein texte (PDF)
3. exploitation dans des outils de textométrie/fouille

Solutions :

1. Du bricolage : une recherche temporelle et un objet lourd
2. Du sioutage : la version classique d'Europresse
3. Du détourage : l'outil CERES

1 : obtenir ”tout” le corpus

Bricolage

Un objet lourd (mais non contondant) et la touche page down (mais on reste limité à 1 000 résultats)

2 : obtenir TOUT le corpus

Sioutage

La version classique d'Europresse qui permet d'obtenir des donnée siterrogeables en plein texte (plus rapide dans un html que dans un PDF)

3 : exploiter tout le corpus

Détourage (outil de S.Poder et F. Allié) Récupérer les textes, séparés et structurés pour pouvoir affiner l'analyse avec :

- Iramuteq
- TXM
- Antconc
- Voyant Tools...

Méta-données disponibles : titre, source, date, auteur

A venir : dé-doublonnage, rangement par dossier ...