# Corpus-écrits GT10 « Exploration de corpus »

Bilan 2012

Perspectives 2013

Animateurs GT10 : Céline Poudat , Marie-Paule Jacques, Serge Heiden

# OBJECTIF

## Documenter les pratiques existantes

Recenser chaque pratique en l'articulant à :

- l’objectif de recherche visé, qu’il soit descriptif ou applicatif ;
- le type de corpus d'étude ;
- les outils de traitement de corpus mobilisés pour s’acquitter de cette tâche, en distinguant entre outils disponibles et outils propriétaires, outils de traitement quantitatif vs qualitatif, etc., suivant les pratiques qui seront mises au jour grâce aux réflexions et aux échanges des participants.

# OBJECTIF
## Retombées escomptées

- Ce travail de synthèse et de recensement de l'existant permettra aux linguistes et plus généralement, aux chercheurs intéressés par un travail sur corpus d'orienter leurs pratiques et leurs choix techniques relativement à leurs objectifs de recherche ;

- Il aidera la communauté à s'orienter parmi l'offre disponible, ce qui n'est pas toujours simple à l'heure actuelle, tout en faisant découvrir d'autres pratiques et d'autres outils plus méconnus
  - Développement d'un catalogue d'outils, de formations, de pratiques...

# Participants

- 3 coordinateurs
- 26 unités de recherche (présentations sur Wiki)
- 33 inscrits à la liste de diffusion
- 27 sur le wiki du groupe
  - https://groupes.renater.fr/wiki/corpus-ecrits-exploration/
- 17 personnes différentes pour au moins une réunion

# Corpus-écrits GT10
# Bilan 2012

**Communication, information :**

- une liste de discussions (https://groupes.renater.fr/sympa/info/corpus-ecrits-exploration)
- un wiki (https://groupes.renater.fr/wiki/corpus-ecrits-exploration)

**Liste de discussion :**

- des nouveaux inscrits depuis l'été
- 15 messages
  - pour le moment un fonctionnement « descendant » : coordinateurs → membres du groupe
  - échange d'informations « pratiques » (réunions, wiki)

# Corpus-écrits GT10
# Bilan 2012

## Communication, information :

- **un wiki** (https://groupes.renater.fr/wiki/corpus-ecrits-exploration)
  - Une page par labo, chaque participant intègre les informations qui le concerne
  - Un système de smileys indique l'état de la page
  - A terme, chaque page synthétise des pratiques en décrivant :
    - Le type de corpus
    - Le type d'outil

mobilisés

## Présentation personnelle (une par personne de l'équipe)

Éditer

- Nom, prénom : **Poudat, Céline**
- Courriel : **cpoudat@gmail.com**
- Page personnelle : http://poudat.fr
- 3 publications (max) avec liens vers texte intégral:
- POUDAT, Céline et LOISEAU, Sylvain (2007). "Représentation et caractérisation lexicale des sciences dans Wikipédia" in Tutin, A. (coord.), « Lexique de la langue scientifique », Revue Française de Linguistique Appliquée, XII, 2007-2, pp. 29-44.
- LEGALLOIS, Dominique et POUDAT, Céline (2008). "Comment parler des livres que l'on a lus ? Discours et axiologie des avis des internautes" in Bertelli, D. et Chauvin-Vileno, A. (coord.), "De la médiacritique culturelle comme métadiscours. Objets, genres, dispositifs", Semen, n°26, pp.49-80. http://semen.revues.org/document8444.html
- AURAY, Nicolas, HURAULT-PLANTET, Martine, POUDAT, Céline et JACQUEMIN, Bernard (2009). "La négociation des points de vue. Une cartographie sociale des conflits et des querelles dans le Wikipédia francophone" in Cardon, D. (coord.), "Web 2.0.", Réseaux n°154, 2009-2, pp. 15-50.
- Thèmes de recherche (5 mots clefs max + descriptif court) : genres du discours, analyse du discours, linguistique de corpus, statistiques textuelles, typologies narratives

## Données / corpus à votre disposition (une section par corpus / ensemble de données)

Éditer

Disposez-vous de corpus ou de données ?

Si oui, renseignez les sections suivantes (en les dupliquant pour chaque corpus) :

### Descriptif

Éditer

- Nom du corpus (s'il en a un)
- Genres de discours et types de documents représentés

### Mode de constitution

Éditer

- Votre corpus a-t-il été réalisé manuellement (encodage précis) ou (semi-)automatiquement (aspirateur Toile ...) ?

### Couverture

Éditer

- Volumétrie (nbre de mots ou de textes, Ko):
- Circonstances de collecte, nbre de participants (bref et approximativement) :
- Empan temporel (quelle fenêtre diachronique prise en compte ?):
- Langues :

### Format des données

Éditer

- formats / mode des données : s'agit-il de textes, transcriptions, audio, vidéo, etc. :
- données en texte simple (Ascii...) / formats propriétaires (Word...) ?
- encodage des caractères en Unicode ?
- données structurées en XML ?
- si oui, suivant standards (TEI ou autres , précisez) :
- organisation des corpus / textes / données (regroupement des fichiers avec empaquetage pour un corpus/texte ? fichiers séparés les uns des autres ? combien de fichiers (environ) ? combien d'ensembles de données ? de textes ? de corpus ?) :

## Annotations

Quelles segmentations (lexicale, phrase, paragraphe, section…) ?

Annotation linguistique automatique

- Quel type d'annotation linguistique (morphosyntaxique, lemmatisation, syntaxique, entités nommées…) ?
- Quel outil d'annotation utilisé (TreeTagger, Cordial…) ?

Annotation linguistique manuelle

- Quel type d'annotation (sémantique, syntaxique, etc.) ?
- Quel outil d'annotation utilisé (Oxygen, Analec, Glozz, Sato, Excel…) ?

## Disponibilité

Votre corpus est-il accessible

- modalités d'accès - téléchargement (transmissible ?), interface d'accès en ligne… ?
- restrictions éventuelles (propriété intellectuelle, déontologique…) ?
- quelle(s) licence(s) (ouverte, sous conditions…) ?
- coûts d'accès (achat, abonnement, barrière mobile, gratuit…) ?

## Documentation disponible

- vos corpus ont-ils des métadonnées ? Si oui, format (entête TEI, Dublin Core, OLAC, tableau Excel, base de données, système documentaire…) ?
- votre corpus est-il déjà documenté dans une infrastructure (labex EFL, Adonis, CLARIN, etc.) ? Si oui, donner un lien vers cette documentation.

## Serveurs

- vos données sont-elles sur un serveur en ligne (URL) ?
- ce serveur a-t-il un protocole OAI (est-il moissonnable par d'autres serveurs) : pour les métadonnées ?
- ce serveur a-t-il un protocole FTP : pour le corps des textes ?
- ce serveur est il accessible par services web (SOAP, REST…) ?

## Autres informations qui vous semblent pertinentes pour décrire vos corpus

---

# Outils

Êtes-vous producteur et / ou utilisateur d'outils?

## Descriptif

Décrire les grands principes d'utilisation de chaque outil.

## Nom / maintenance de l'outil

Indiquer le nom de chaque outil.

Si vous ou votre laboratoire en est le producteur, préciser s'il est maintenu et par qui.

## Objectifs méthodologiques poursuivis

Méthodologies employées (par ex. contraster deux à n ensembles, dégager des spécificités, caractériser un usage, caractériser une / des unité(s) linguistique(s), cooccurrences de phénomènes, etc.) ?

## Fonctionnalités exploitées

Quelles fonctionnalités des outils susnommés utilisez-vous pour réaliser ces objectifs?

## Disponibilité

Les outils sont-ils accessibles ?

- modalités d'accès - téléchargement (transmissible ?), interface d'accès… ?
- gratuit, payant, abonnement ?
- restrictions éventuelles
- les sources sont elles disponibles (licence(s) open-source,…) ?
- type d'installation (système d'exploitation, installateur/manuel…)
- paramétrage de l'outil (fichiers paramètres, lexiques utilisés…), relation avec d'autres outils,…

## Technologies composant l'outil

- langage(s) de programmation (C, Python, Java…)
- standard d'architecture (OSGi, J2EE, .NET, …)

## Formats de corpus gérés

- TXT, XML, HTML, TEI…
- Annotations (lexicales, syntaxiques…)
- Langues traitées…
- Métadonnées traitées : entête TEI…

## Type d'interface pour l'utilisateur

- application pour poste avec interface graphique utilisateur, en ligne de commande, scripts, portail web en ligne, …
- langue(s) de l'interface

## Documentation

- pour l'utilisateur : formats, accès
- pour les développeurs : formats, accès
- site web
- listes de diffusion
- fiches de description : fiche Plume…
- votre logiciel est-il déjà documenté dans une infrastructure (labex EFL, Adonis, CLARIN, Bamboo, etc.) ?

## Autres informations qui vous semblent pertinentes pour décrire vos outils

# Bilan provisoire Wiki

- Difficultés de connexion et d'édition
  - Mise en ligne d'une notice explicative (à venir, mutualiser d'ailleurs)

- Documentation des corpus
  - Recrutement de Linda Hriba + expérience de Sarra El Ayari → formulaire de description des corpus sur le site du consortium

# Bilan 2012
# 2 réunions

- En présentiel: le 16 mai 2012 (16 participants)
  - Tour de table : objectifs / thématiques de recherche des participants
  - Clarification des nœuds notionnels à dénouer (qu'est-ce qu'une pratique, comment veut-on explorer les corpus...)
  - Présentation des recherches et des outils des participants
  - Discussion finale (validation des objectifs, moyens, logistique, programmation)
- En ligne: le 2 juillet 2012 (9 participants)
  - Travail sur le format d'une synthèse descriptive des usages d'exploitation de corpus les plus répandus dans la communauté francophone (format base de données)
    - Travail collectif sur gdoc disponible sur disponible sur https://docs.google.com/spreadsheet/ccc?key=0AnyqPxDqTJA7dFRTeGY4LW1GSkFzMFhOTFdmdEh4V3c#gid=0
    - Clarification de la structure du document (usages et outils)
  - Réflexion sur les perspectives 2012-2013

# Synthèse descriptive

- Pas seulement inventaire mais outil de réflexion
- Construction collective et collaborative des critères pertinents
  - Qu'est-ce qu'une pratique ?
  - Par quoi décrire un outil ? (fonctionnalités prévues, scénarios effectifs d'utilisation ?)
- *« Chacune des pratiques recensées sera précisément documentée et articulée à deux entrées :*
  - *l'objectif de recherche escompté, qu'il soit descriptif ou applicatif;*
  - *les outils de traitement de corpus mobilisés pour s'acquitter de cette tâche, en distinguant entre outils disponibles et outils propriétaires, outils de traitement quantitatif vs qualitatif, etc., suivant les pratiques qui seront mises au jour grâce aux réflexions et aux échanges des participants.* »

# Mise en œuvre

Travail collectif – tableur gdoc

## USAGES (orientation utilisateurs)

- Objectif de recherche
  - L'objectif poursuivi par la recherche : ce qui motive l'utilisation de cette fonctionnalité sur ce corpus
- Pratique
  - Catégorie générique à discuter, qui pourra être remplie ultérieurement - scénarios types d'analyse
- Outil (nom, visée, scénarios)
- Corpus (type, annotation, format, volume)

## OUTILS (orientation développeurs)

- Visée
- Fonctionnalités
- Accessibilité
- SE
- Format des corpus
- Plateforme, modules
- Langage de programmation
- Encodage
- Ergonomie, interface
- Etc.

# Perspectives

- Base de données :
  - Finaliser (définir très précisément chaque champ)
  - « Peupler » (ajouter de nouvelles données)
  - Publier (hébergement, maintenance, maintien ?)
- Pistes pour alimenter et motiver les travaux du groupe :
  - Projet de publication
  - Journées de formation
  - Journée d'étude
  - Montage de projets communs (appels à projets de l'ANR)