

CORLI, consortium et centre K Clarin



CORLI : CORpus, Langues et Interactions

- Consortium de l'infrastructure Huma-Num dédié à la linguistique de corpus
- Comité de pilotage d'un peu moins de 20 personnes
- Comité scientifiques: plus de 25 personnes
- Un réseau d'environ 180 chercheurs, 400 souscriptions à la liste CORLI
- Large couverture des différents domaines de la linguistique

- Une action "réseau"
- Des actions "développement"
 - voir <http://corli.huma-num.fr/>

CORLI : le réseau

- **Fédérer** les laboratoires travaillant sur corpus pour **recenser** les ressources, les pratiques et les besoins
- **Mutualiser** les ressources et les **diffuser** dans des modalités et des formats garantissant leur interopérabilité
- **Établir et partager** des bonnes pratiques
- Diffuser des **standards** européens et internationaux : contributions CLARIN , DARIAH et au Consortium TEI
- **Formation** tous chercheurs (débutants, confirmés, universitaires, EPST, etc.)
- **Veille** technologique, **information**, **discussion** des nouveaux standards, techniques, outils, méthodes, etc.
 - Par exemple, méthodes FAIR, évaluation des corpus, citation des corpus, données ouvertes

CORLI : les actions de développement

- **Annotation collaborative**
- **Citation de corpus ou d'extraits de corpus**
- **Corpus Ouvert du Français: données de corpus et outils pour la langue française (OFC: Open French Corpus)**

Annotation collaborative de corpus

- Mettre à disposition une plateforme de transcription et d'annotation des données langagières (**PALAMEDE**)
- Mettre à disposition une plateforme d'annotation de haut niveau avec des fonctionnalités d'active learning (**INCEPTION**)
- Annotations en classe de type **GUM**

Citation de corpus et d'extraits

Objectif général: lorsqu'un article scientifique fait une citation extraite d'un texte ou d'un corpus, avoir les moyens de:

- Pointer sur une ressource qui décrit et qui présente visuellement l'extrait ou les données
- Cette ressource peut pointer sur les données de corpus originales
- Avantages: pérenniser les citations de données scientifiques, permettre la mesure de l'usage des citations, permettre la réutilisation des citations

Open French Corpus & outils pour les linguistes

Objectif général: développer une ressource librement accessible, organisée, standardisée et de qualité pour le français à partir des corpus existants

- une ressource librement accessible
- une ressource organisée
- une ressource standardisée
- une ressource de qualité

Plan de travail

- Evaluation des corpus déposés sur Ortolang/Cocoon - évaluation des métadonnées
- Détermination de métadonnées minimales
- Harmonisation des métadonnées
 - collaboration Cahier/Ariane, utilisation de la TEI
- Déterminer liste (ouverte et évolutive) de corpus disponibles sur un site pérenne
- Conversion vers un ou des formats commun texte ou TEI (écrit, CMC, oral, multimodal)
- Utilisation de plateformes pour l'interrogation et autres traitements linguistiques
 - CLARIN - Content search
 - Ortolang/Frannot