



Les dynamiques des communautés de linguistes et la création du centre K

Christophe Parisse et Céline Poudat, coordinateurs de CORLI



CORLI

- Consortium de l'infrastructure Huma-Num dédié à la linguistique de corpus
- Fusion des deux précédents consortiums de linguistiques
 - IRCOM (corpus oraux)
 - IRCE (corpus écrits)



Dynamiques de CORLI

- Une dynamique entre laboratoires:
 - un comité de pilotage de plus d'une vingtaine de personnes, organisé en laboratoires + un CS
- Une dynamique structurée en thématiques et en points d'intérêt
 - Inter-Explo : Interopérabilité /Pratique et outils d'exploration de corpus
 - Multicom : CMC / LSF et gestualité
 - Corpus multilingues et plurilingues
 - Questions Éthiques & Cadre Juridique
 - Annotation
 - Évaluation des corpus



Standards et objectifs communs

- **Fédérer** les laboratoires travaillant sur corpus pour **recenser** les ressources, les pratiques et les besoins
 - recension des corpus, métadonnées, outils...
- **Mutualiser** les ressources et les **diffuser** dans des modalités et des formats garantissant leur interopérabilité
 - Ortolang, Cocoon
 - XML-TEI, formats outils (CLAN, ELAN...)
- **Développer** ou **compléter** les formats, outils, ressources utilisés par toute la communauté



Standards et objectifs communs

- **Établir et partager** des bonnes pratiques
 - bonnes pratiques juridiques, méthodologiques
 - pour résoudre les problèmes des chercheurs
 - pour améliorer la qualité des travaux
- Diffuser des **standards** européens et internationaux : contributions CLARIN , DARIAH et au Consortium TEI
- Établir des **critères d'évaluation des corpus** en tant que production scientifique



Leviers: formations, bonnes pratiques et finalisation de corpus

Christophe Parisse et Céline Poudat, coordinateurs de CORLI



Les formations

Les retours du réseau des consortiums IRCE et IRCOM ont montré l'importance de l'organisation de formations pour compléter les cursus universitaires et former les chercheurs.

CORLI a donc prolongé et amplifié l'organisation de formations dans divers domaines et surtout:

- outils de création et édition de corpus
- outils pour l'exploration et l'annotation des corpus
- dépôts de données, édition de métadonnées

Année	Type d'Activités	Nombre total sur l'année	Total des participants sur l'année
2016	<i>3 sessions de formation CORLI sur 5 jours</i>	<i>8 ateliers de formation d'une demi-journée</i>	<i>Une centaine</i>
2017	<i>2 sessions de formation CORLI sur 5 jours</i>	<i>12 ateliers de formation d'une demi-journée</i>	<i>Une centaine</i>
	<i>3 ateliers de formation CORLI aux JLC 2017</i>	<i>3 ateliers de formation d'une demi-journée</i>	
2018	<i>2 sessions de formation CORLI sur 6 jours</i>	<i>10 ateliers de formation d'une demi-journée</i>	<i>Une centaine</i>
	<i>2 ateliers de formation ELAN</i>	<i>2 ateliers de formation</i>	<i>40 participants</i>
2019	<i>2 sessions de formation CORLI sur 5 jours</i>	<i>8 ateliers de formation d'une demi-journée</i>	<i>70 participants</i>
	<i>5 ateliers de formation CORLI aux JLC 2019</i>	<i>5 ateliers de formation</i>	<i>-</i>



Définir des bonnes pratiques

CORLI agit en vue de favoriser le dépôt de données corpus et de recherche et de documenter ces actions.

Les pratiques défendues par CORLI depuis de nombreuses années correspondent aux principes FAIR:

- utilisation de formats pérennes, de métadonnées
- diffusion sans restriction des données
- formats permettant la consultation et la réutilisabilité des données



Finalisation de corpus

- Levier actif depuis 2016
 - et même avant dans les deux premiers consortiums de linguistique
 - <https://corli.huma-num.fr/node/210>
- 195 000 euros ont été injectés pour aider des laboratoires et des projets à finaliser et diffuser leurs travaux.
 - ex: en 2019, nous avons reçu 29 demandes, la quasi-totalité étant sérieuses et argumentées. Nous n'avons pu financer que 10 demandes.
- Action qui répond aux manques existants dans les projets de recherche et dans les laboratoires.
 - Le corpus est une denrée aujourd'hui incontournable de la recherche en linguistique
 - Mais c'est une denrée coûteuse que les appels à projet et les laboratoires ont du mal à financer, étant une denrée peu productive en publications directes.
- D'où l'importance pour beaucoup de chercheurs des appels à finalisation de corpus et l'importance du travail sur l'évaluation des ressources.



Pistes pour le futur du consortium en fonction des expériences passées

- Maintenir le réseau : nécessaire pour la veille, la diffusion, la validation des techniques et des données
- Maintenir la formation et les bonnes pratiques
- Formats, données, outils pour faciliter le travail des chercheurs
 - dans le cadre de la science ouverte
 - des données partagées et disponibles à tous (libres au moins pour la recherche)
 - sécuriser les données des chercheurs
 - fournir ou compléter les outils que les laboratoires ne peuvent créer ou obtenir seuls
 - s'ouvrir aux projets européens et au partage de données au delà de la France