Dix ans avec

Christophe Parisse et Céline Poudat

AG Cahier, 26 novembre 2021

Les consortiums de linguistique - 10 ans déjà

Les consortiums de linguistique ont 10 ans

- IRCE - IRCOM - 2011 - 2015
- CORLI - 2016 - 2021

Une réussite en tant que communauté

- partage, diffusion, mise en commun d’informations, de standards, d’outils
- finalisation de corpus
- ateliers, workshop, réflexion sur les corpus, leur création, leur utilisation

Bref historique

Deux consortiums de linguistique

IRCE (corpus écrits, coordination ILF)

Porteur: F. Neveu

- 11 groupes de travail
- une cinquantaine de finalisation de corpus en partenariat avec Ortolang
- organisation récurrente d'ateliers de formation
- projet national CoMéRé
- recensement et description des outils d'exploration de corpus et publication de l'ouvrage de référence "Explorer un corpus textuel" (Poudat & Landragin)

IRCOM (corpus oraux, coordination TUL)

Porteurs : S. Robert, D. Boutet

- 4 groupes de travail
- 27 projets de finalisation de corpus
- constitution d'un glossaire
- recensement des corpus
- outil de conversion TEICORPO

Consortium CORLI - 2016-2021

Porté par l'ILF de 2016 à 2018 (F. Neveu)

Porté par Modyco (2019) puis la MESHS (2020-2021) - C.Parisse et C. Poudat

Réalisations et actions

Quelques réalisations

Formations: une à deux séances par an - hors années COVID (2020/21)

- Transcriber, ELAN, CLAN, TXM, Unitex, Trameur, Hyperbase Web, Iramuteq, Dtm-Vic, Praat
- exploration de corpus, corpus multimodaux, réalisation MOOC, techniques audio/video

Finalisation de corpus - 6 années de 2016 à 2021

- 43 corpus de 2016 à 2019 (194 910€)
- 12 corpus en 2020 (35 000€)
- 6 corpus en 2021 (35 000€)

Groupes projets

- Interopérabilité / Pratique et outils d’exploration de corpus (GP1)
- Multimodalités et nouvelles formes de communication (GP2)
- Multilinguisme (GP3)
- QuECJ Questions juridiques (GP4)
- Annotation de haut niveau (GP5)
- Evaluation des corpus (GP6)

Quelques réalisations

Organisation d’ateliers de réflexion et de travail sur les pratiques communes et les besoins

De nombreuses participations à des conférences (CLARIN, CMC, TEI, utilisation des corpus, outils CORLI)

Publication sur le site de CORLI d’inventaires (outils, corpus écrits, oraux et multimodaux), du glossaire

Outils de conversion de corpus (TEICORPO) et d’édition des métadonnées (TEIMETA)

Situation en 2021

- Documentation, participation aux formations, outils de conversion ou d’édition, interface pour analyse grammaticale (STANZA)
- Participation à des congrès, présentations, actes de congrès, articles scientifiques, chapitres de livre
- Refonte du site, réorganisation des documents de formation, mise en place d’un système de ticket et d’une FAQ
- Réunions de concertation pour le projet à venir (AG, GPs annotation et MultiCOM)

Actions CLARIN

- Constitution de CORLI en tant centre K.
- Participation de CORLI au comité Clarin FR
- Participation au “Resource families” pour les langues des signes.
- Eva Soroli Ambassadrice CLARIN (pour représenter tous les pays de CLARIN).

Actions CLARIN

- Participation à la conférence CLARIN 2017, 2019, 2020.
- Participation à la réunion annuelle des centres K Clarin - prochaine réunion: le 14/12 de 13 à 17h
- Stage CORLI/U. Strasbourg sur les outils CLARIN (Switchboard) - Comment les utiliser et les étendre (Analyseur STANZA)
- Resserrement des liens avec les centres B et C de CLARIN (Cocoon, Ortolang)

Le nouveau projet CORLI

Un projet pour les prochaines années

Le conseil scientifique d'Huma-Num offre deux possibilités (soumises à évaluation)

1- continuer en tant que réseau avec un budget annuel d'environ 10 000€

2- déposer un projet nouveau intégrant les conditions du CS d'Huma-Num

- Projet scientifique avec des réalisations, un planning, des livrables
- Pas de création de corpus ou de données
- International et multidisciplinaire

Offrir une partie réseau et une partie recherche et développement

Réseau

- Maintenir les formations, la documentation, la FAQ, le centre K, les groupes projets
- Continuer de proposer et d’améliorer les services

Recherche et développement

- Projet Annotation
- Projet Citation

Changer de modèle de fonctionnement

Deux constats

- Plus de finalisation de corpus
- Un manque de temps pour les membres du copil et des groupes projets

Besoin de temps de travail et donc de main d'oeuvre

- Pour maintenir le site, les informations, les données, les formations
- Pour développer de nouveaux outils qui viendront mieux utiliser les ressources existantes
- Un poste à mi-temps CDD pour nous aider dans la partie réseau et participer aux projets de R&D
- Des prestations externes (contrainte CNRS) pour les développements

Ouvrir CORLI en tant que réseau: collaborations, liens avec les laboratoires

Ouvrir le comité de pilotage aux infrastructures: ORTOLANG, COCOON, HUMA-NUM, CLARIN

Ouvrir à tous les laboratoires sans participation nécessaire au comité de pilotage

La participation au comité de pilotage, comme la participation aux groupes projets, est fonction des motivations, de la disponibilité, et de l’investissement des personnes dans l’utilisation de corpus de langage

Ouvrir CORLI pour les projets: MSH, CLARIN, TAL

Projet Annotation:

- Participation de la MSH Lorraine, d'Huma-Num, ATILF, Litt&Art (Grenoble, Littérature, Sociologie, Arts)

Liens à développer avec les GDR Lift et le GDR TAL

Liens avec CLARIN à renforcer: mieux utiliser les ressources de CLARIN, mettre à disposition les ressources françaises

Projet Annotation

- Réunions de travail MSH Lorraine / Huma-Num / CORLI
- Développement en cours d'un corpus échantillonné
- Trois axes principaux
  - mettre à disposition une plateforme de transcription et d'annotation des données langagières
    - projet Palamède, coopération avec les concepteurs de TACT (Grenoble)
  - mettre à disposition une plateforme d'annotation de haut niveau avec des fonctionnalités d'active learning
    - coopération avec INCEPTION (TU Darmstadt)
  - développer une ressource d'annotations en classe de type GUM

Projet Citation

Collaboration de CORLI, ATILF/ORTOLANG, HUMANUM/COCOON, Prismes, Modyco (20 personnes dont 8 pour CORLI)

But:

- gérer les citations d’extraits de textes ou de corpus en générant un identifiant pérenne pour chaque citation,
- lier finement écrits scientifiques et données de langage (écrits, sons, vidéo, images) présentées dans leur contexte
- faciliter la réflexion scientifique et la réutilisation des données.

Liens avec le consortium CAHIER

Comme c’est le cas pour CLARIN, le consortium CORLI s’intéresse et propose des données de langage, indépendamment de la discipline qui étudie ces textes.

- Partage d’outils (TXM)
- Partage de corpus (données écrites, scolaires)
- Partage de format (TEI)

Mieux faire connaître les outils des deux consortiums et mieux les partager.

- Outils de CLARIN (Switchboard) ou de CORLI (analyse syntaxique, exploration de corpus, annotation)
- Outils de CAHIER (présentation/visualisation des textes)

Exemple: Le fonds BOISSY

http://www.licorn-research.fr/Boissy.html

http://www.licorn-research.fr/Boissy/Admete_et_Alceste/HTML/Admete_et_Alceste.html

http://www.licorn-research.fr/Boissy/Admete_et_Alceste/TEI/Admete_et_Alceste.xml

Analyseur STANZA: https://corliapi.ortolang.fr/stanza/

Analyse en format vertical CONLL

```
#sent_id = 5
#text = POLIDECTE, grand Prêtre , frere d'Admete HERCULE CLÉONE, confidente d'Alceste.
1	POLIDECTE	POLIDECTE	PROPN	_	_	0	root	_	start_char=90|end_char=99
2	,	,	PUNCT	_	_	4	punct	_	start_char=99|end_char=100
3	grand	grand	ADJ	_	Gender=Masc|Number=Sing	4	amod	_	start_char=101|end_char=106
4	Prêtre	prêtre	NOUN	_	Gender=Masc|Number=Sing	1	appos	_	start_char=107|end_char=113
5	,	,	PUNCT	_	_	6	punct	_	start_char=114|end_char=115
6	frere	frere	NOUN	_	Gender=Masc|Number=Sing	1	appos	_	start_char=116|end_char=121
7	d'	de	ADP	_	_	8	case	_	start_char=122|end_char=124
8	Admete	Admete	PROPN	_	_	6	nmod	_	start_char=124|end_char=130
9	HERCULE	HERCULE	PROPN	_	_	8	flat:name	_	start_char=131|end_char=138
10	CLÉONE	CLÉONE	PROPN	_	_	9	flat:name	_	start_char=139|end_char=145
11	,	,	PUNCT	_	_	12	punct	_	start_char=145|end_char=146
12	confidente	confidente	NOUN	_	Gender=Fem|Number=Sing	1	appos	_	start_char=147|end_char=157
13	d'	de	ADP	_	_	14	case	_	start_char=158|end_char=160
14	Alceste	Alceste	PROPN	_	_	12	nmod	_	start_char=160|end_char=167
15	.	.	PUNCT	_	_	1	punct	_	start_char=167|end_char=168
```

Visualisation avec BRAT