Annotation outillée de la continuité référentielle dans un corpus scolaire trilingue du primaire

Martina Barletta

Annotation outillée de la continuité référentielle dans un corpus scolaire trilingue du primaire

Martina Barletta[1,2]

[1] Laboratoire LIDILEM, Université Grenoble Alpes (France)

[2] Dipartimento di Scienze dell'Educazione Riccardo Massa, Université Milan-Bicocca (Italie)

martina.barletta@univ-grenoble-alpes.fr

Objectif

Dans une approche comparative, étudier le développement de la cohérence/cohésion textuelle à travers l'annotation de la continuité référentielle dans un corpus d'écrits scolaires du primaire en français, italien et espagnol pour nourrir la réflexion en didactique

Corpus français et coréférence

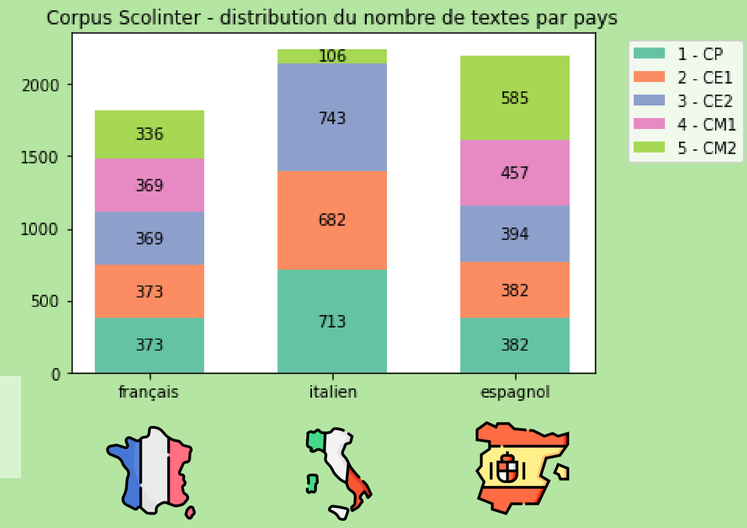
Corpus annotés 2000 – 2013
ARCADE (Tutin *et al.,* 2000)
DéDé (Gardent et Manuélian, 2005)
ANNODIS (Péry-Woodley *et al.,* 2011)
ANCOR (Muzerelle *et al.,* 2013)

Chaîne de référence
ensemble d'au moins trois expressions faisant référence à la même entité, appartenant à l'univers du texte (Schnedecker, 1997)

Corpus	Année	Taille	Genres	Phénomènes annotés
democrat	2019	58 textes, 560 000 tokens	écrits narratifs et autre genres variés	chaines de référence
<Corpus RésolCo/>	2021	385 textes, 72 873 tokens	écrits scolaires	continuité référentielle sur les entités provoquées par la consigne

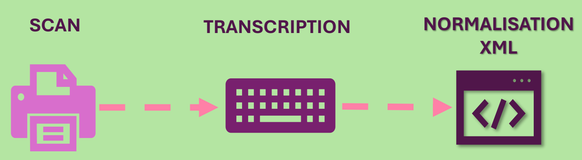
Corpus Scolinter

Production d'un récit à partir d'images, du CE1 au CM2



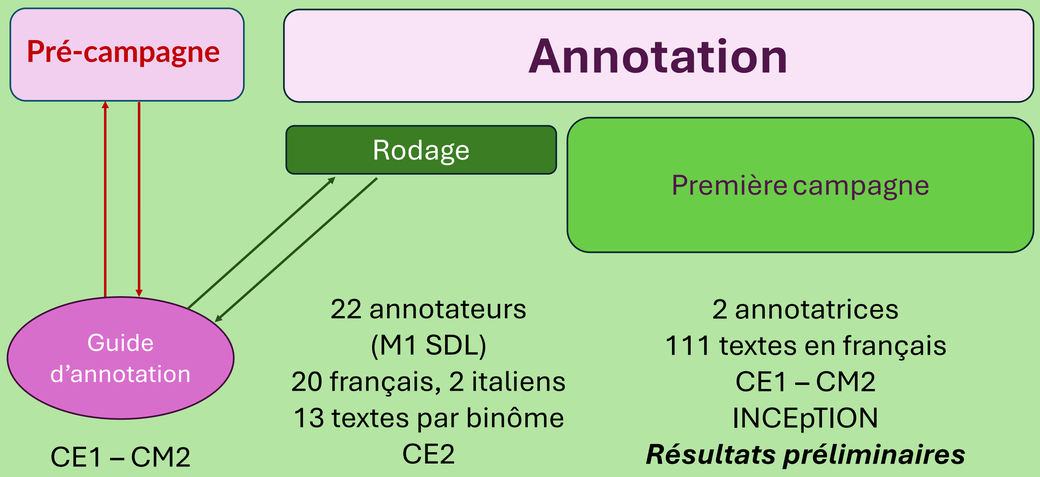
Lien vers le corpus Scolinter

Lien vers le corpus Scoledit (FR)



Campagne d'annotation

- Annotation des personnages de la consigne et des entités animés



CE1 – CM2
30 textes
(français, italien)
papier

22 annotateurs
(M1 SDL)
20 français, 2 italiens
13 textes par binôme
CE2
INCEpTION

2 annotatrices
111 textes en français
CE1 – CM2
INCEpTION
Résultats préliminaires

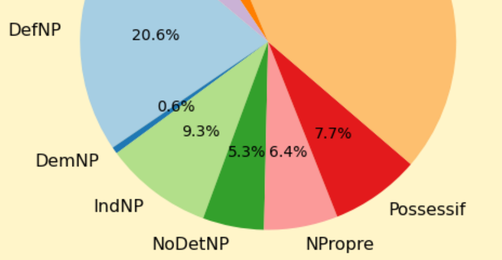
- Annotation sur INCEpTION (ATILF)
- Layer de la Coréférence par défaut
- Textes pré-annotés en CoNLL-U

Figure inspirée de Fort, 2013

Résultats préliminaires

Niveau	Nb textes	Nb tokens	Nb moyen tokens par texte	Maillons	Nb moyen de référents par texte	Démsité référentielle (maillons/tokens)	Longueur moyenne des chaines
CE1	32	2 388	74,63	461	2,21	19,3%	6,54
CE2	25	3 475	139	651	2,56	18,73%	10,68
CM1	27	4 541	168,19	868	3,11	19,11%	11,01
CM2	27	5 979	221,44	1 066	3,48	17,83%	11,60
Corpus	**111**	**16 383**	**150,82**	**3 046**	**2,84**	**18,59%**	**9,96**

Pronoms	1 248
Syntagmes nominaux définis	604
Syntagmes nominaux indéfinis	272
Déterminants possessifs	226
Noms propres	186
Syntagmes nominaux sans déterminant	154
Autre type	132
Anaphore zéro	89
Syntagmes nominaux démonstratifs	17
Total	2 928*



Conclusions

- Densité référentielle cohérente avec textes narratifs de scripteurs experts (Landragin *et al.,* 2024)
- Ambiguïtés du guide levées pour une nouvelle campagne d'annotation (juin-septembre 2024)
- Réaliser une campagne d'annotation sur des textes en français et en italien pour comparer les chaines dans les deux langues

Perspectives

- Troisième campagne d'annotation – ***nouveau guide***
  - 150 textes en italien : CE1 – CE2
  - 150 textes en français : CE1 – CE2
- Deux annotateurs pour le corpus français, deux annotateurs pour le corpus italien
- Comparaison des résultats entre français et italien