



HAL
open science

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

Martina Barletta

► To cite this version:

Martina Barletta. Annotation outillée de la continuité référentielle dans un corpus scolaire trilingue du primaire. 35èmes Journées d'Études sur la Parole (JEP) 31ème Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 26ème Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL), Jul 2024, Toulouse, France. hal-04646867

HAL Id: hal-04646867

<https://hal.science/hal-04646867>

Submitted on 12 Jul 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

Martina Barletta^{1,2}

¹Laboratoire LIDILEM, Université Grenoble Alpes (France)

²Dipartimento di Scienze dell'Educazione Riccardo Massa, Università 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)

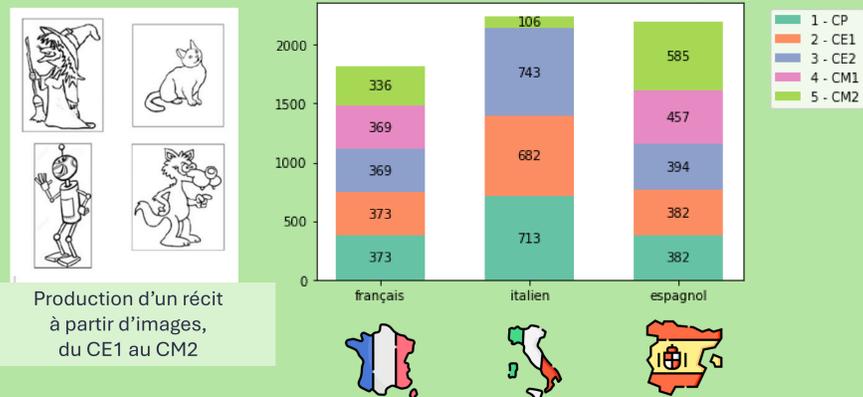
Chaîne de référence

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

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

Corpus Scolinter

Corpus Scolinter - distribution du nombre de textes par pays

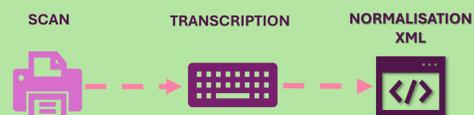
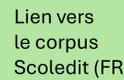


Production d'un récit à partir d'images, du CE1 au CM2

Lien vers le corpus Scolinter



Lien vers le corpus Scoledit (FR)



Campagne d'annotation

- Annotation des personnages de la consigne et des entités animés

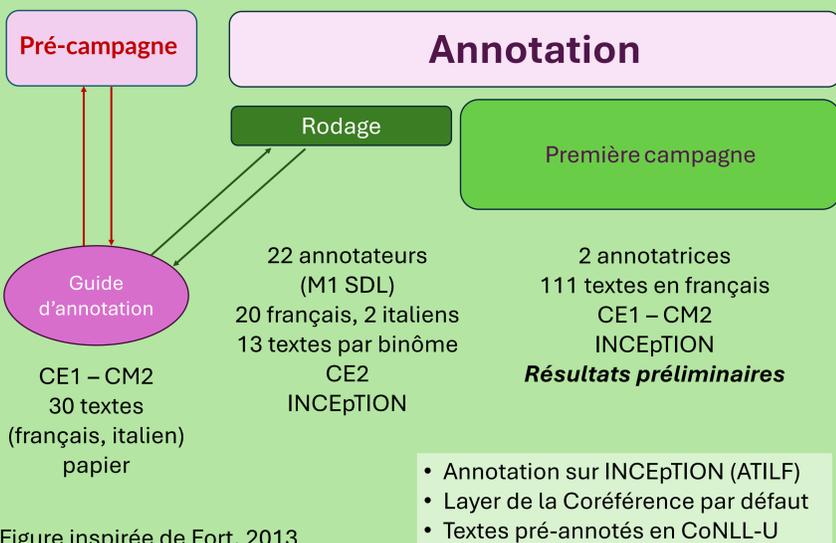


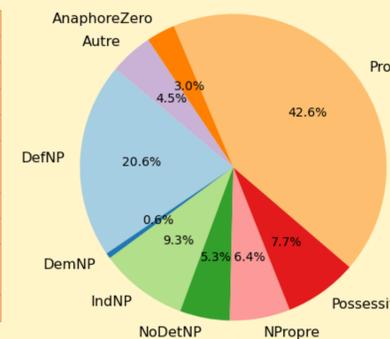
Figure inspirée de Fort, 2013

Résultats préliminaires

Niveau	Nb textes	Nb tokens	Nb moyen tokens par texte	Maillons	Nb moyen de référents par texte	Densité référentielle (maillons/tokens)	Longueur moyenne des chaînes
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

Distribution des mentions par type

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 chaînes 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