
Comment annoter dans un corpus des chats, des loups, de robots et des sorcières

Annotation de la coréférence dans le corpus Scolinter



Martina Barletta
Université Grenoble Alpes (France)
Thèse dirigée par C. Brissaud, C. Ponton, F. Da Milano



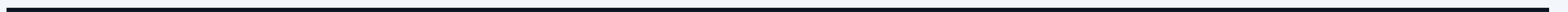
01. Introduction – coreference et Scolinter

02. Guide et outils pour l'annotation

03. TP – Annotation sur Inception

04. Conclusion

Plan



01.



Introduction

Tout d'abord

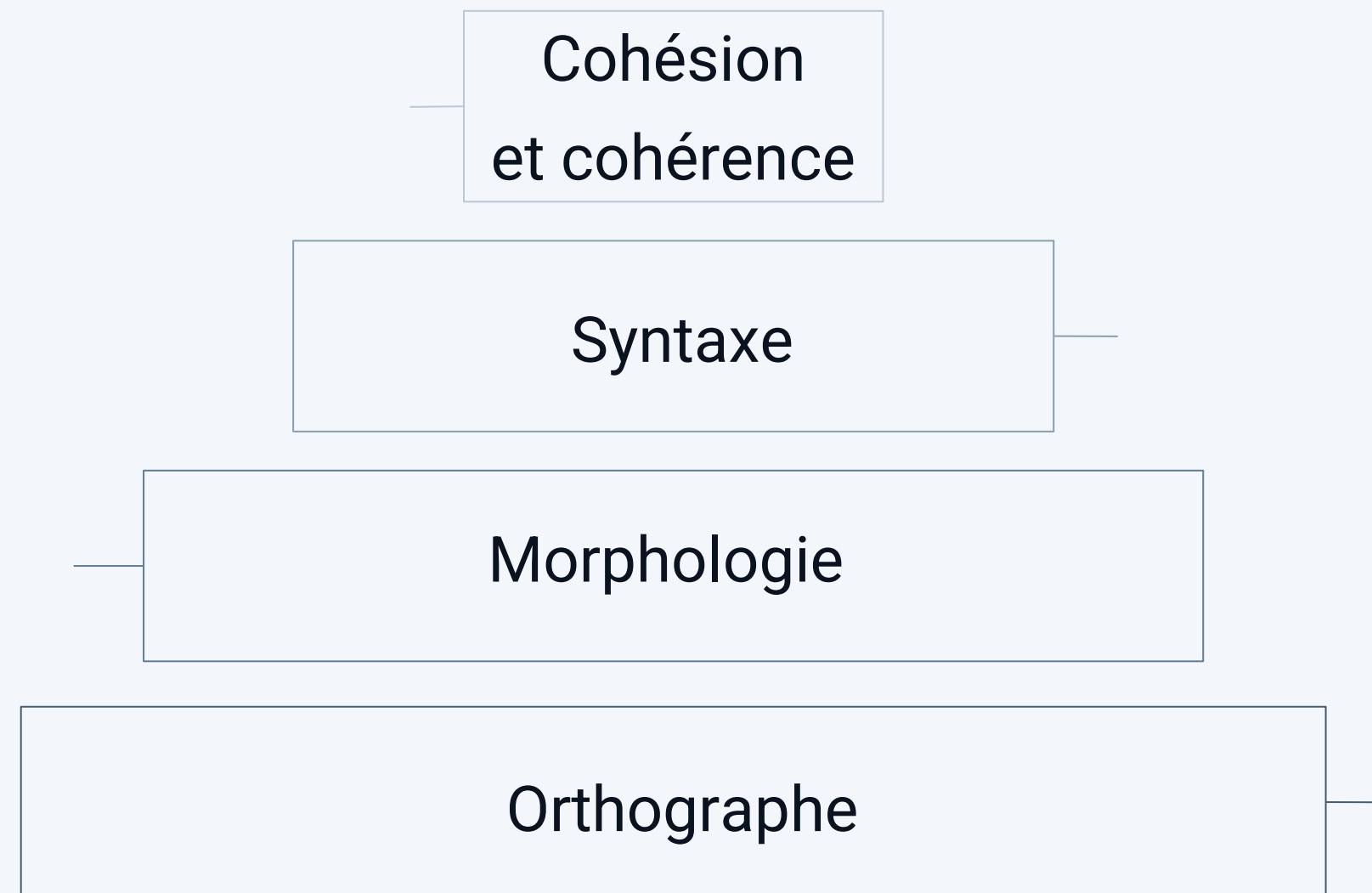
...

Comment peut-on définir un texte
“bien écrit”?

Quels critères objectifs pour le
déterminer?

...la coréférence en fait partie.

Evaluation de l'écriture



Coréférence



Terminologie

Coréférence
Chaîne de coréférence(s)
Clusters

Maillons
Mentions

Coréférence



Exemples

Joe Biden s'est exprimé aujourd'hui par rapport aux derniers événements. Le président américain (...)

Tom et Julie sont au cinéma. Ils ont acheté des popcorns.

Coréférence



Exemples (traduit depuis l'italien)

Il y avait un groupe de chats comme lui. Un jour un monsieur arrive et il en adopte un. Le jour suivant une dame en adopte deux autres. Peu à peu, il reste tout seul.

Corpus



ANR – 2018 – 2022

Constitution de corpus
d'écrits scolaires

(Doquet *et al.*, 2019;
Ponton, Jacques, *et al.*, 2022)



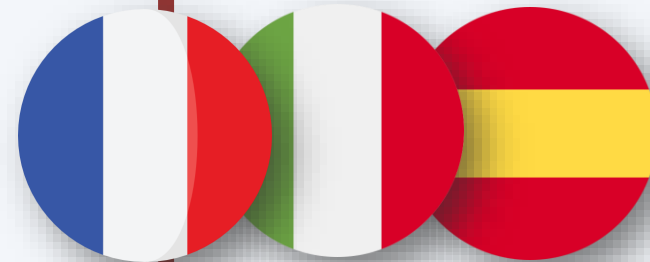
**Littéracie
avancée**

Scoledit

2014 - 2018

Corpus longitudinal d'écrits
scolaires à l'école primaire
en France...

(Wolfarth *et al.*, 2017;
Wolfarth, 2019)




2018 - en cours
... en Italie et en Espagne

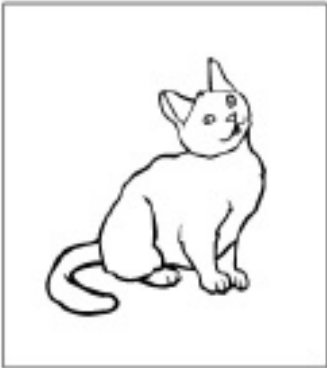
(Ponton *et al.*, 2021)

Composition


1




2



3

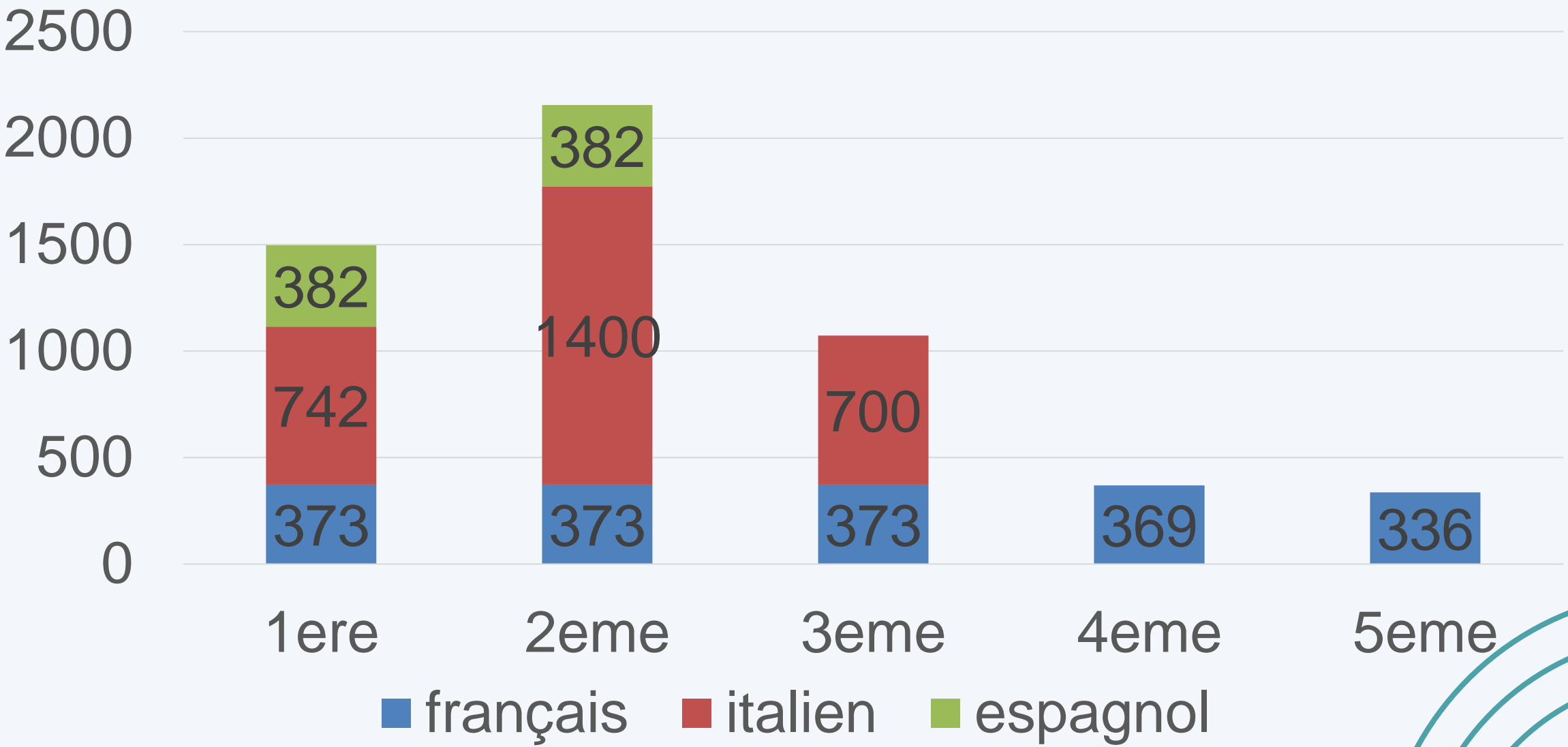


4



Consigne du CE1, CE2, CM1 et CM2

Distribution des textes sur Scolinter
- en cours



Récolte



LIDILEM

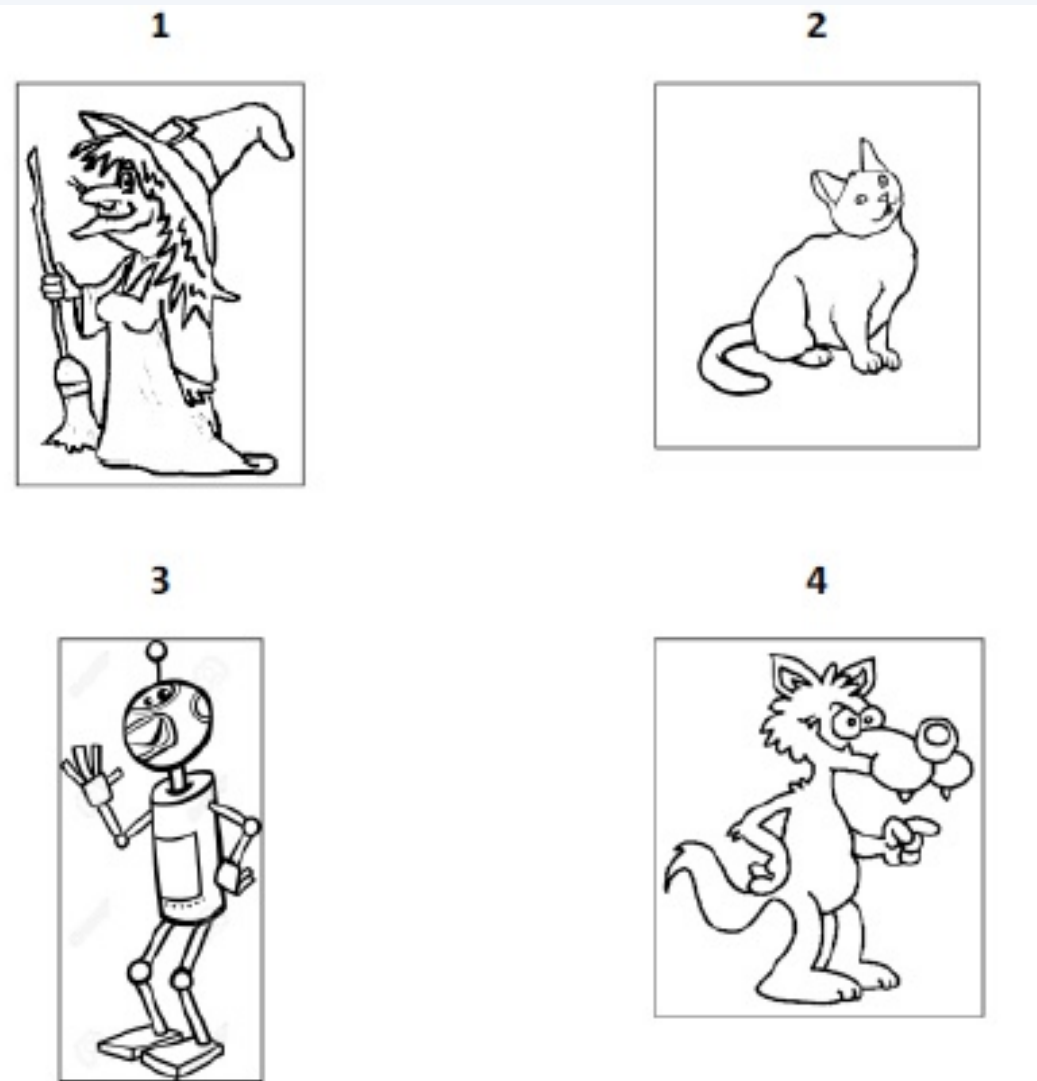
Université
Grenoble Alpes



UNIVERSIDAD
DE ALMERÍA



UNIVERSITÀ DEGLI STUDI
DI MILANO
BICOCCA



Voici 4 personnages. Choisis un ou deux personnages et raconte une histoire.
Entoure le ou les personnages que tu as choisis.

Consigne du CE1, CE2, CM1 et CM2

Constitution du corpus

La sorcière et le loup

S

Il éte tunc fois une sorcière voult transformé
un loup en chat mes le loup ~~se~~ ne
voulé par se lésé per il courra jusca sa
meute et leur dis scile lui tes arrivé mes la
sorcièr la suivi juscas sa meute et le loup
sauté sur la sorcière et la mange.

<text>

<body>

<p>

<head>La sorcière et le loup </head>

Il était une fois une sorcière <omission
type="pronom"/> voulait transformer un loup en chat
mais le loup ne voulait pas se laisser faire <seg
type="segm. forte"/> il courut jusqu'à sa meute et leur
dit ce qu'il lui était arrivé mais la sorcière l'a suivi
jusqu'à sa meute et le loup saute sur la sorcière et la
mange.

</p>

</body>

</text>

Corpus



Corpus complet

1820 textes

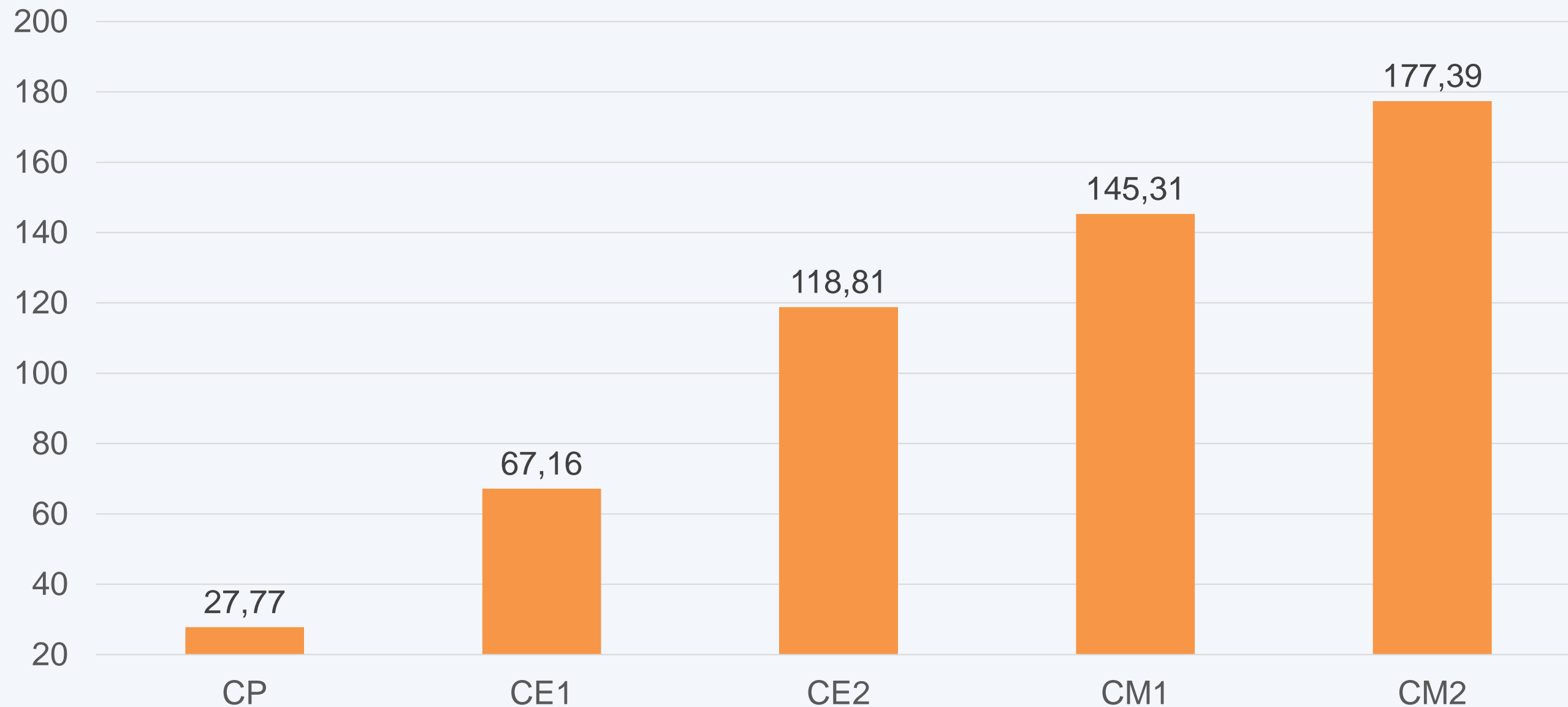
Corpus longitudinal de référence

1685 textes

337 par niveau

180 782 tokens

Nombre moyen de mots par niveau
sur le corpus longitudinal de référence





Le corpus RésolCo

Résolution de la Continuité Référentielle

Raconte une histoire dans laquelle tu insèreras séparément et dans l'ordre donné les trois phrases suivantes :

- P1 - Elle habitait dans cette maison depuis longtemps.
- P2 - Il se retourna en entendant ce grand bruit.
- P3 - Depuis cette aventure, les enfants ne sortent plus la nuit.

Garcia-Debanc et al., 2021

02.



Guide et outils d'annotation

TP



1. Annoter les textes

Chacun annote un texte (le même que son binôme) et annote ses doutes, interrogations, remarques...

2. Export des textes annotés

Exporter ses résultats au format Conll-2012 → format qui contient l'annotation en coréférence

3. Evaluation qualitative des résultats obtenus

Et comparaison avec l'annotation de son binôme



Le guide d'annotation

Annotation de la continuité référentielle

- Sur ***toutes*** les expressions langagières selon la description du guide
- En utilisant le niveau « coreference » de Inception

Demo annotation → comment utiliser l'annotation et définir le jeu d'étiquettes



TP

- Lecture du guide (et éventuelles remarques initiales)
 - Choix d'un texte à annoter en parallèle à son binôme pendant le cours
 - Annotation (chacun de son côté)
- Pendant l'annotation, annoter ses doutes, remarques, interrogations, qui peuvent faire partie du rapport final
- Export du fichier du texte annoté en format Conll-12
 - Comparaison des résultats obtenus et évaluation qualitative (quantitative aussi si l'on souhaite)
 - Continuer l'annotation sur les textes restants pour enrichir la rédaction du rapport final

Rendu : rapport, fichiers conll-12 des textes annotés
AVANT LES VACANCES DE NOEL à M. Ponton et à Martina

TP



- Télécharger le guide et la liste des textes à annoter depuis le cloud uga : <https://cloud.univ-grenoble-alpes.fr/s/fqY8fzNLeQ35rW6>
- Se connecter à <https://inception-demo.atilf.fr/p/corpus-ecrits>
- Trouver les textes à annoter par son binôme
- Commencer l'annotation aujourd'hui en commençant par le même texte que son binôme
- Sélectionner le niveau de coréférence pour réaliser son annotation



03.



Résultats

CONSIGNE TP

POUR RAPPEL :

Le rapport est à rendre le 20 décembre 2023 avant minuit à Martina Barletta et à Claude Ponton en tant que TP pour le cours de Corpus écrits.

Chaque binôme a un numéro de référence qui indique les textes qu'il doit annoter dans le fichier csv présent dans le cloud. Merci d'indiquer dans votre rendu quel est votre numéro de groupe.

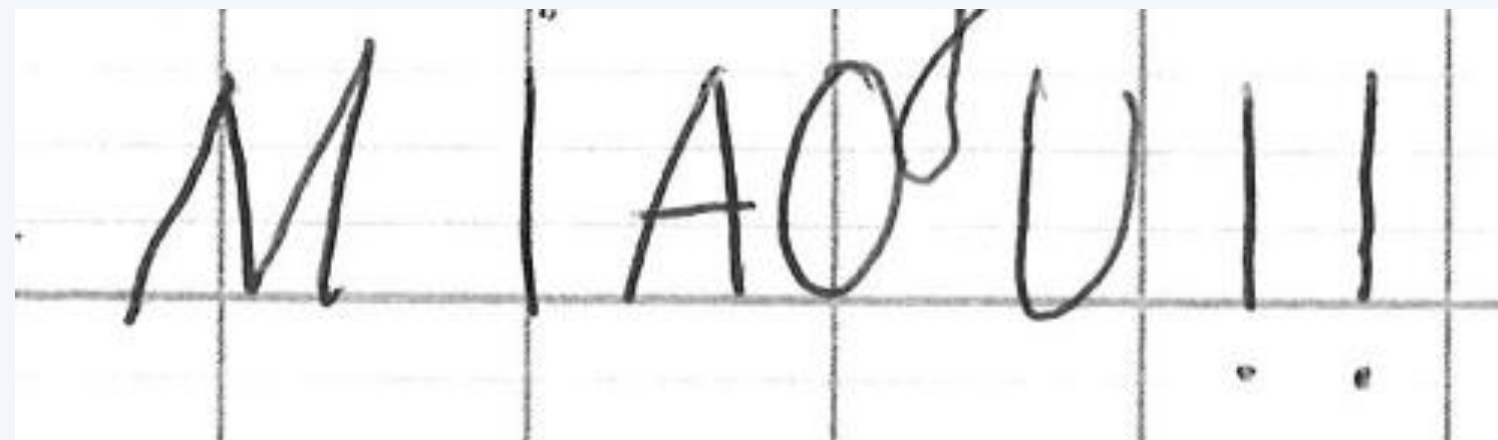
CONSIGNE

Les deux personnes du binôme doivent annoter TOUS les textes assignés au binôme et effectuer une comparaison qualitative lors que leurs annotations soient divergentes. Le but est d'annoter les textes chacun de son côté pour ensuite évaluer de manière qualitative l'accord interannotateur, notion abordée lors du corpus de Corpus écrits.

Le rapport doit porter donc sur l'analyse du processus d'annotation en binôme des textes de CE2 du corpus Scoledit ainsi que sur analyse plutôt qualitative de l'accord inter-annotateur rencontré dans le binôme lors du processus d'annotation.

Ce rapport vise à décrire le processus d'annotation de la continuité référentielle effectué sur Inception en suivant le guide d'annotation fourni. Le but principal de ce TP est analyser les difficultés rencontrées lors de l'annotation, et d'avoir une prise de recul sur les principes d'annotation à suivre décrits dans le guide. Ex. Vous avez rencontré dans vos annotations des cas d'éléments linguistiques qui indiquent un référent qui n'est pas exactement pris en compte dans les exemples du guide et comment vous avez décidé d'aborder ces éléments, de quelle manière vous avez traité des « cas compliqués » lors de votre annotation et motivation du choix de traitement effectué etc... L'annotation doit suivre au plus près possible les principes décrits dans le guide. La compréhension du guide et sa mise en pratique font partie des critères d'évaluation du TP.

Évaluer la continuité référentielle dans l'écriture des enfants ou évaluer la qualité de l'écriture des textes annotés N'EST PAS L'OBJECTIF DE CE TP.



MIAOUII

Merci à Léa Wichroff pour son travail d'évaluation du corpus
et à Olivier Kraif pour les pistes de réflexion
Thèse sous la direction de Catherine Brissaud,
Claude Ponton (LIDILEM) et Federica Da Milano (UNIMIB)

