CORPUS

Corpus

26 | 2025
Bruit de fond ou valeur ajoutée ? Gérer le bruit lors des traitements informatiques des corpus linguistiques

La question de la normalisation des écrits scolaires pour leur traitement automatique. Le cas de l'omission de mots

Martina Barletta and Claude Ponton

OpenEdition Journals

Electronic version
URL: https://journals.openedition.org/corpus/9322
DOI: 10.4000/1364v
ISSN: 1765-3126

Publisher
Bases ; corpus et langage - UMR 6039

Electronic reference
Martina Barletta and Claude Ponton, "La question de la normalisation des écrits scolaires pour leur traitement automatique. Le cas de l'omission de mots", *Corpus* [Online], 26 | 2025, Online since 27 January 2025, connection on 01 February 2025. URL: http://journals.openedition.org/corpus/9322 ; DOI: https://doi.org/10.4000/1364v

La question de la normalisation des écrits scolaires pour leur traitement automatique. Le cas de l'omission de mots

Martina Barletta and Claude Ponton

Merci à Olivier Kraif pour ses suggestions qui ont mené à la version actuelle de cette étude et à Catherine Brissaud pour sa contribution à la réalisation du corpus de référence.

1. Introduction

1 Les écrits scolaires présentent des caractéristiques qui peuvent rendre leur analyse automatique complexe. En effet, ils sont souvent éloignés de la norme (sur les plans de la ponctuation, de l'orthographe, de la cohérence, de la cohésion textuelle...) et présentent une certaine inventivité dans l'expression de leur contenu (sur le plan lexical comme sur le plan narratif) (Doquet & Ponton 2021). Cet éloignement à la norme de la langue de référence provoque des erreurs lorsqu'on conduit des analyses automatiques sur les différents niveaux linguistiques de ces écrits. Une solution pour limiter ces erreurs de traitement est de normaliser le texte puis d'aligner cette version normalisée avec la version d'origine. C'est cette approche qui a été adoptée dans le projet Scoledit (Wolfarth *et al*. 2017). Cette « normalisation » est vue comme la production d'un écrit orthographiquement normé restant au plus près de la production initiale (Wolfarth 2019), et elle est effectuée manuellement par les chercheurs-linguistes du projet. Elle permet à la fois des comparaisons avec la production de l'élève sur différents niveaux (lexical, orthographique, morphologique, etc.) et le recours aux outils TAL (Wolfarth *et al.* 2018).

2 Toutefois, les analyses menées sur ces comparaisons sont dépendantes des choix de normalisation effectués. C'est le cas notamment des mots omis par l'élève. Par exemple, dans l'approche Scoledit, la transcription du texte de la Figure 1 donne : « *Il étais une fois*

un robos étais seul au monde » alors que la normalisation donne : « *Il était une fois un robot* ***<omission type="pronom"/>*** *était seul au monde* ».

Figure 1. Scan d'un extrait du texte EC-CE1-2015-115-D1-S1350

3 Comme le montre cet exemple, les mots oubliés par l'élève sont simplement indiqués par une balise *<omission>* qui précise la catégorie grammaticale du mot omis. Il est à noter que le choix du type des catégories grammaticales est opéré manuellement par les chercheurs-linguistes du projet. Ce marquage des omissions entrave l'utilisation des outils de traitement conçus pour la langue standard, notamment dans des tâches de base comme l'étiquetage morphosyntaxique ou l'analyse en dépendance. Les erreurs de traitement peuvent ensuite avoir des répercussions sur les traitements de plus haut niveau comme l'annotation et l'analyse de la continuité référentielle qui sont les sujets de notre travail de recherche (Barletta 2024). Ces erreurs peuvent être de deux types : les erreurs de bruit (mention détectée qui n'en est pas) et les erreurs de silence (mention non détectée).

4 Le travail présenté dans cette contribution vise à limiter les erreurs de bruit de l'analyse morphosyntaxique et en dépendance liées aux omissions de mots. Ce travail est un préalable à nos recherches sur la continuité référentielle dans les écrits scolaires puisqu'il participera ensuite à la réduction des erreurs de silence sur l'annotation des mentions, comme dans le cas du pronom relatif « qui » omis par l'élève dans l'exemple précédent (cf. Figure 1), qui fait partie de la chaîne de référence du référent « robot ».

5 D'un point de vue méthodologique, nous examinons trois méthodes de traitement automatique afin de choisir la plus pertinente en termes de temps et de précision par rapport à un corpus de référence, élaboré manuellement par trois chercheurs linguistes du projet, dans lequel les mots omis ont été remplacés par le mot candidat le plus probable. La première méthode envisagée utilise un token « masque » identique pour remplacer chaque balise <omission>, la deuxième remplace chaque balise <omission> par un token choisi *a priori* pour chaque catégorie grammaticale indiquée dans la balise <omission>, alors que la troisième exploite le modèle Transformer FlauBERT (Le *et al.* 2020) pour remplacer automatiquement les balises <omission> par un mot le plus probable. Nous évaluons ensuite l'impact de ces trois méthodes sur l'analyse proposée par le parseur morphosyntaxique et en dépendance Spacy v.3 (Honnibal *et al.* 2020).

6 Avant de présenter les résultats de ces analyses, nous allons dans un premier temps décrire le corpus Scolinter (Ponton *et al.* 2021), objet de notre recherche, et plus précisément le corpus de référence utilisé pour nos expérimentations. Dans un deuxième temps, nous expliciterons la méthodologie suivie pour la réalisation de ces expérimentations. Après avoir présenté les métriques appliquées pour évaluer les résultats obtenus, inspirées des travaux en évaluation des parseurs, nous proposerons enfin une analyse quantitative et qualitative sur les trois méthodes présentées.

2. Le corpus Scolinter : un corpus scolaire trilingue longitudinal

7 Le corpus Scoledit représente à l'heure actuelle l'un des seuls corpus longitudinaux d'écrits scolaires français entièrement transcrit et accessible en ligne (Ponton *et al.* 2021)[1]. Ce corpus rassemble des productions manuscrites recueillies dans différentes écoles primaires françaises entre 2014 et 2018, en suivant les mêmes cohortes d'élèves du CP au CM2, sur la base de deux consignes composées d'images proposées par les chercheurs. Ce corpus a comme objectif initial le suivi du développement des compétences d'écriture au primaire.

8 S'appuyant sur cette même méthodologie et avec des objectifs similaires, le projet Scolinter[2] (Ponton *et al.* 2021) vise la constitution d'un corpus comparable de productions écrites d'élèves du primaire en France, Italie et Espagne avec un suivi longitudinal des mêmes cohortes d'élèves dans chaque pays. Le corpus Scolinter comprend à l'heure actuelle : 1 820 textes transcrits et normalisés en français ; 2 244 textes en italien ; 2 200 textes en espagnol. Notons que le corpus français est complet sur les cinq niveaux du primaire alors qu'une partie des textes des corpus italien et espagnol est encore en phase de transcription et de normalisation notamment pour les niveaux 2, 3 et 5 en Italie (CE1, CE2 et CM2) et pour les niveaux 3, 4 et 5 en Espagne (CE2, CM1 et CM2). L'un des objectifs de Scolinter est d'étudier et de comparer l'évolution des compétences à l'écrit des élèves dans ces trois pays. Actuellement, sur ce corpus, nous nous intéressons plus particulièrement aux aspects cohérence/cohésion des textes à travers l'étude des chaînes de continuité référentielle portant sur les personnages des histoires produites par les élèves. Ces personnages (le loup, la sorcière, le chat et le robot) sont induits par la consigne donnée qui est la même dans les trois langues du CE1 au CM2. Une première partie de ce travail consiste à annoter les chaînes de référence de ces textes (Barletta 2024).

9 Pour cette contribution, nous ne nous intéressons qu'aux niveaux CP-CM2 de la partie française du corpus Scolinter. Ce sous-corpus porte sur les 337 mêmes élèves et comporte 1 684 textes (un texte par élève et par niveau, dont un texte averbal qui n'a pas été pris en compte), totalisant 178 313 tokens. Sur cet ensemble, 228 textes (13,53 %) contiennent au moins une balise d'omission pour un total de 274 balises d'omission. Ce pourcentage, qui peut sembler faible par rapport à la taille de notre corpus, constitue toutefois un obstacle important lors de l'analyse morphosyntaxique de ces textes, une composante essentielle pour notre travail d'annotation des phénomènes tels que les chaînes de continuité référentielle.

3. Le corpus de référence

Tableau 1. Distribution des textes par nombre de tokens sur le corpus longitudinal et sur le corpus de référence

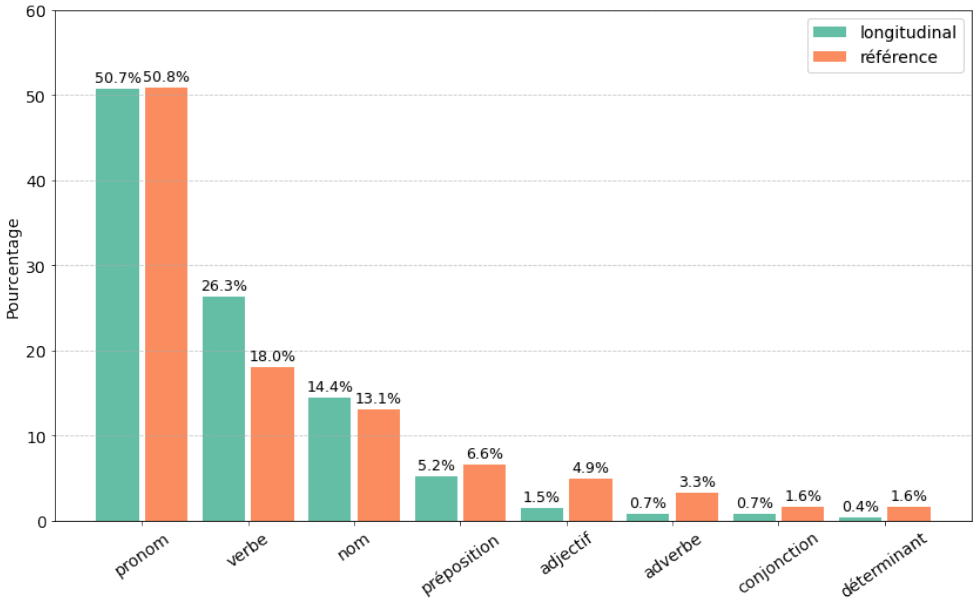
	Corpus longitudinal		Corpus de référence	
Nb tokens	Nb textes	Pourcentage sur le corpus longitudinal	Nb textes	Pourcentage sur le corpus de référence

0-20	126	7,48 %	0	0
21-50	372	22,09 %	6	11,11 %
51-100	414	24,58 %	17	31,48 %
101-150	343	20,37 %	12	22,22 %
151-200	237	14,07 %	6	11,11 %
201-250	111	6,60 %	5	9,26 %
251-plus	81	4,81 %	8	14,81 %
Tot.	1 684	100 %	54	100 %

10 Pour effectuer nos expérimentations, nous avons extrait parmi les 1 684 textes un sous-corpus composé de 54 textes des niveaux CE1 à CM2. Afin de conserver la même répartition en nombre de tokens que dans le corpus longitudinal, nous avons partitionné le corpus en sept tranches et conservé un pourcentage de textes comparable (cf. Tableau 1). Notons que la tranche 0-20 tokens n'a pas été prise en compte car elle propose des textes trop courts et souvent peu intéressants pour notre étude. Dans ce sous-corpus, une annotatrice experte (doctorante en TAL et sciences du langage) a remplacé manuellement les balises omission par le token le plus probable dans le contexte du texte, tout en respectant l'étiquette grammaticale proposée lors de l'étape de normalisation et présente en attribut de la balise d'omission. Cette annotation a été effectuée sur les textes normalisés. Par exemple, comme dans l'exemple de la Figure 1, la catégorie proposée était celle de pronom, et nous pouvons faire l'hypothèse que le token omis le plus probable pour reconstruire le sens de la phrase dans laquelle il est inséré soit le pronom relatif « qui ». L'annotation de ces omissions a été ensuite discutée et validée avec les chercheurs-linguistes du projet. Aucun de ces choix n'a suscité de désaccord mais, dans le cadre d'une expérimentation plus large, il serait certainement nécessaire de procéder à un calcul d'accord inter-annotateurs.

11 Après avoir effectué cette annotation, nous avons comparé la distribution des omissions par catégorie grammaticale (POS ou *part of speech*) entre les 274 omissions du corpus longitudinal et les 61 omissions du corpus de référence (cf. Figure 2). La catégorie grammaticale utilisée pour ces analyses est celle indiquée en attribut des balises. Cette distribution par POS est similaire entre les deux corpus sur les catégories les plus fréquentes, ce qui nous permet d'obtenir un échantillon suffisamment représentatif de notre corpus complet.

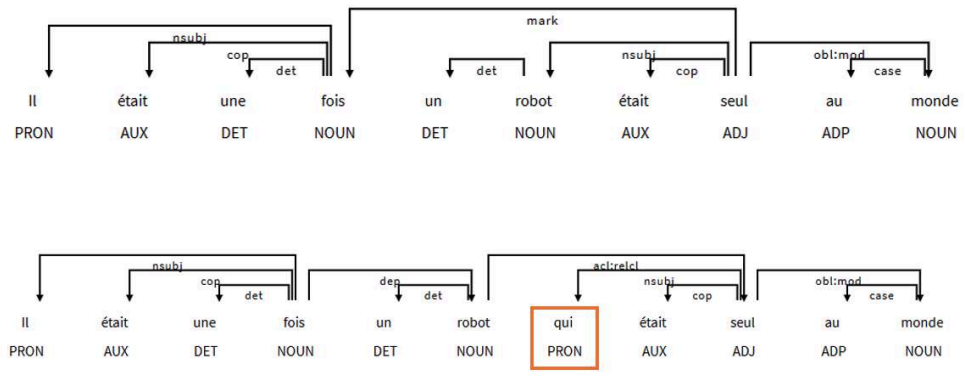
Figure 2. Distribution des omissions par étiquette morphosyntaxique dans les corpus longitudinal et de référence (catégorisation proposée manuellement lors de la normalisation)



12 En observant la distribution des omissions par POS sur les corpus longitudinal et de référence, nous pouvons observer que dans ces deux corpus la catégorie la plus fréquemment omise est celle du pronom. Les pronoms représentent ici plus de 50 % des tokens omis. Cette catégorie est suivie par les verbes (26,3 % dans le corpus longitudinal, qui descend à 18 % dans celui de référence) et les noms (14,4 % dans le longitudinal et 13,1 % dans le corpus de référence), alors que d'autres catégories comme prépositions, adjectifs, adverbes, conjonctions et déterminants représentent au total 8,5 % des omissions dans le corpus longitudinal et 18 % dans le corpus de référence. À noter que lors de la normalisation les chercheurs ont décidé de regrouper les deux catégories d'auxiliaire et de verbe sous la même étiquette.

13 Si des éléments tels que les prépositions peuvent être peu pertinents pour notre recherche, les pronoms et les noms sont susceptibles d'être annotés comme partie des chaînes de continuité référentielle[3], comme dans le cas de l'exemple suivant (Figure 3). Dans cette phrase, l'absence du pronom relatif « qui » impacte la structure de la phrase qui le contient. Les verbes font aussi partie des tokens qui peuvent porter une marque de présence du référent dans le texte et pourraient donc être susceptibles d'être annotés. Par exemple dans le cas de verbes qui se trouvent dans des phrases coordonnées où le sujet n'est pas répété[4]. Ces observations préliminaires confirment la nécessité de traiter au mieux ces tokens absents pour assurer la réussite de notre travail d'analyse du corpus Scolinter à travers l'annotation et l'analyse des chaînes de continuité référentielle (Barletta 2024).

Figure 3. Analyse en dépendance d'une phrase tirée du texte EC-CE1-2015-115-D1-S1350, générée par Spacy (v.3) en utilisant le modèle *fr_core_news_lg*. En haut, la phrase sans le mot omis, en bas la phrase avec le token réintroduit (surligné dans la représentation de l'analyse)



4. Méthodologie

14 Sur le sous-corpus de référence, nous avons effectué le traitement des tokens omis selon trois méthodes distinctes dont nous avons ensuite évalué l'efficacité en utilisant les métriques couramment employées pour l'évaluation des analyseurs en dépendance.

15 La première méthode dite « xxx » consiste à remplacer chaque balise <omission> par le token de forme *xxx*. Ce token est habituellement utilisé en TAL en tant que token « masque » dans d'autres applications[5].

16 La deuxième méthode, dite « prototypique », consiste à remplacer chaque catégorie morphosyntaxique proposée dans les balises d'omission par les chercheurs chargés de la normalisation par le même token. Ici nous avons à faire à une méthode semi-automatisée : après une phase de réflexion entre plusieurs annotateurs experts destinée à choisir des tokens non-ambigus au niveau de la catégorie grammaticale, nous avons établi une liste de correspondance (catégorie grammaticale–token), utilisée pour le remplacement automatique des balises (cf. Tableau 2).

Tableau 2. Tokens choisis pour remplacer chaque catégorie morphosyntaxique proposée lors de la normalisation

Catégorie	**Token**
conjonction	et
adverbe	vraiment
adjectif	méchant
préposition	avec
nom	maison
verbe	mange
pronom	qui

17 La troisième méthode utilise le modèle Transformer FlauBERT[6] (Le *et al.* 2020) pour substituer la balise <omission> par un token prédit par le modèle (un token par balise). Nous avons utilisé le modèle FlauBERT pré-entraîné (FlaubertLMHeadModel) et son tokeniseur sur les textes du corpus. Chaque balise <omission> est remplacée par un mot masque propre au tokeniseur, qui permet de prévoir la meilleure substitution possible, option qui est ensuite intégrée dans le texte à la place de la balise <omission>.

18 Du point de vue technique, nos textes ont fait l'objet de plusieurs traitements, dont une transformation du format XML d'origine au format txt, avec les balises d'omission remplacées selon le scénario concerné. Tous les textes ont été ensuite parsés au format CoNLL-U en utilisant le parseur Spacy v. 3 (Honnibal *et al.* 2020) et le modèle pour le français *fr_core_news_lg*.

Tableau 3. Extrait du texte EC-CE1-2015-120-D1-S2449 dans sa version de référence et les trois méthodes testées

Référence	XXX	Prototypique	FlauBERT
(...) Toma va montrer à l'entraîneur de l'équipe de badminton. Il lui **<omission type="verbe" prop="dit"/>** « tu as 3 min pour me montrer, tu vas jouer contre Tom » à la fin du match Toma met le dernier point et remporte le match. (...)	(...) Toma va montrer à l'entraîneur de l'équipe de badminton. Il lui **xxx** « tu as 3 min pour me montrer, tu vas jouer contre Tom » à la fin du match Toma met le dernier point et remporte le match. (...)	(...) Toma va montrer à l'entraîneur de l'équipe de badminton. Il lui **mange** « tu as 3 min pour me montrer, tu vas jouer contre Tom » à la fin du match Toma met le dernier point et remporte le match. (...)	(...) Toma va montrer à l'entraîneur de l'équipe de badminton. Il lui **dit** « tu as 3 min pour me montrer, tu vas jouer contre Tom » à la fin du match Toma met le dernier point et remporte le match. (...)

5. Métriques utilisées

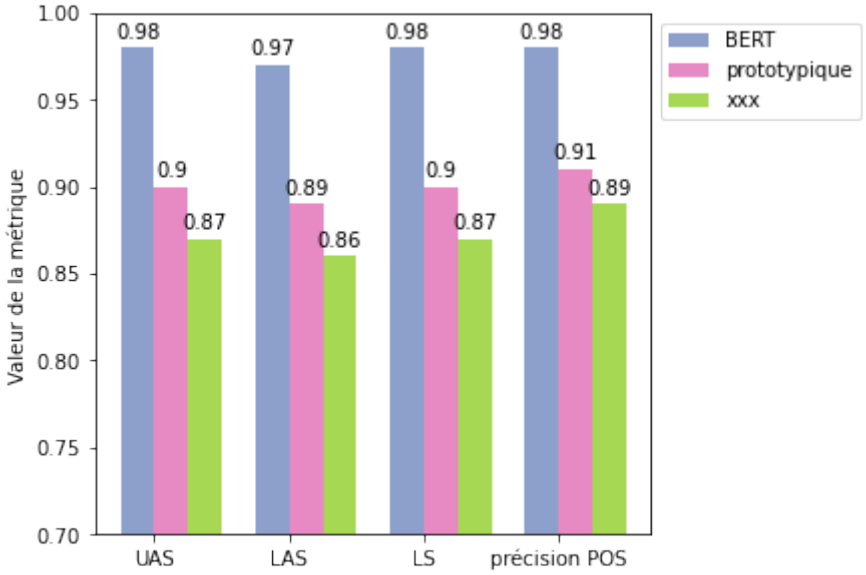
19 Afin d'évaluer la capacité des différentes méthodes à choisir un token afin d'obtenir une analyse en dépendance aussi proche que possible de celle d'une référence qui inclut un token remplaçant l'omission choisi selon son contexte, nous avons opté pour l'utilisation des métriques couramment employées dans la littérature pour évaluer les analyseurs en dépendance (Jurafsky & Martin 2024 : 431-432). Nous avons ainsi choisi les métriques suivantes pour évaluer nos résultats : UAS, LAS, LS et la précision sur la partie du discours. Ces métriques ont été ensuite appliquées uniquement aux phrases contenant des tokens omis.

20 Le LAS (Labeled Attachment Accuracy) fait référence à l'assignation correcte d'un mot à sa tête ainsi qu'avec la relation de dépendance correcte, alors que UAS considère seulement le rattachement à la tête (Jurafsky & Martin 2024 : 431-432). Le LS (Label accuracy Score) considère le nombre de tokens auxquels a été assigné l'étiquette en dépendance correcte. La précision correspond ici au nombre de tokens omis dont le POS est identique à celui prévu dans l'annotation de référence.

6. Résultats obtenus

21 Tout en ayant exploré l'application de méthodes semi-automatisées comme celle dite « prototypique » ou totalement automatisées comme la méthode « xxx », celles-là ont obtenu des métriques moins satisfaisantes par rapport à la méthode qui exploite le transformer. Cette méthode a atteint des résultats de 0.98 pour UAS, le LS et la précision sur POS, ainsi que de 0.97 pour le LAS.

Figure 4. Métriques UAS, LAS, LS et précision sur le POS calculé en comparaison avec le scénario manuel ou de référence sur les 54 textes du corpus de référence



22 Notre analyse qualitative se focalisera principalement sur la méthode utilisant FlauBERT, étant donné que les autres alternatives présentent des résultats plus faibles. En analysant plus en détail les données obtenues, nous avons pu observer des résultats très encourageants surtout sur le remplacement des pronoms. Parmi les 35 textes qui contiennent des balises de pronoms omis, 49 % ont obtenu le score de 1 aux quatre métriques. Cela peut être expliqué par le fait que les pronoms relatifs représentent 51 % des pronoms omis. Ces pronoms relatifs ont été correctement reconstruit par rapport à leur POS et leur rôle dans le texte à 100 % des cas par le modèle FlauBERT[7].

Tableau 4. Texte NORM-EC-CM1-2017-117-D1-S3012 dans ses versions de référence et de la méthode FlauBERT

Référence	FlauBERT
Il était une fois une sorcière **qui** était chez elle et elle s'est dit qu'elle n'avait rien à faire donc elle est partie dans sa salle magique (...)	Il était une fois une sorcière **qui** était chez elle et elle s'est dit qu'elle n'avait rien à faire donc elle est partie dans sa salle magique (...)

23 Dans certains cas, le modèle n'a pas proposé des tokens cohérents avec le contexte de la phrase analysée, donnant lieu à des phrases « incomplètes » dont la structure reste

déviante à la norme. C'est le cas de l'exemple ci-dessous, où à la place du verbe qui avait été proposé, le modèle FlauBERT a proposé le token « tout » en fonction d'adverbe (erronément catégorisé comme adjectif dans l'analyse en dépendance obtenue en sortie).

Tableau 5. Texte NORM-EC-CE1-2015-108-D1-S2984 dans ses versions de référence et de la méthode FlauBERT

Référence	FlauBERT
Un chat **était** seul mais un jour il rencontra un loup alors il courut puis se reperdit.	Un chat **tout** seul mais un jour il rencontra un loup alors il courut puis se reperdit.

Figure 5. Analyse en dépendance des deux versions du texte NORM-EC-CE1-2015-108-D1-S298 effectuée automatiquement à l'aide de Spacy v. 3 et du modèle *fr_core_news_lg*

manual					predicted			
#id NORM-EC-CE1-2015-108-D1-S2984-V					#id NORM-EC-CE1-2015-108-D1-S2984-V			
1	Un	DET	2	det	1	Un	DET	2	det
2	chat	NOUN	4	nsubj	2	chat	NOUN	9	nsubj
3	était	AUX	4	cop	3	tout	ADJ	4	advmod
4	seul	ADJ	14	advmod	4	seul	ADJ	2	amod
5	mais	CCONJ	9	cc	5	mais	CCONJ	9	cc
6	un	DET	7	det	6	un	DET	7	det
7	jour	NOUN	9	nsubj	7	jour	NOUN	9	nsubj
8	il	PRON	9	nsubj	8	il	PRON	9	nsubj
9	rencontra	VERB	4	conj	9	rencontra	VERB	0	ROOT
10	un	DET	11	det	10	un	DET	11	det
11	loup	NOUN	9	obj	11	loup	NOUN	9	obj
12	alors	ADV	14	advmod	12	alors	ADV	14	advmod
13	il	PRON	14	nsubj	13	il	PRON	14	nsubj
14	courut	VERB	0	ROOT	14	courut	VERB	9	advcl
15	puis	CCONJ	17	cc	15	puis	CCONJ	17	cc
16	se	PRON	17	expl:pass	16	se	PRON	17	expl:pass
17	reperdit	PROPN	14	conj	17	reperdit	PROPN	9	conj
18	.	PUNCT	14	punct	18	.	PUNCT	9	punct

24 Certains cas restent cependant difficiles à traiter, de manière manuelle tout comme à l'aide d'un modèle de langage. Il s'agit surtout des cas où le choix du terme à insérer est sujet à diverses interprétations pour les chercheurs, à cause de la nature fortement ambiguë du texte traité ou du contexte. Par exemple, dans le texte suivant, la proposition faite par les chercheurs ne correspond pas à celle proposée par le modèle. Le choix de token effectué dans la référence peut être questionnable et cela est dû à la lacunosité de la phrase, qui contient seulement un syntagme nominal avant d'être interrompue par une conjonction de coordination. La relative pauvreté du contexte ne permet donc pas de reconstruire un token omis assez probable, et cela est confirmé dans la proposition de token aussi redoutable faite par FlauBERT dans ce cas (le token non-mot « dor »).

Tableau 6. Texte NORM-EC-CE2-2016-108-D1-S2983 dans ses versions de référence et méthode FlauBERT

Référence	FlauBERT
C'est l'histoire d'un chat qui se faisait toujours embêter par un loup. Ce méchant loup n'arrêtait pas d'aller chez le chat et de mettre le bazar. Ce pauvre animal **fatiguait** mais un jour Rouqui le chat décida de mettre des pièges de partout.	C'est l'histoire d'un chat qui se faisait toujours embêter par un loup. Ce méchant loup n'arrêtait pas d'aller chez le chat et de mettre le bazar. Ce pauvre animal **dor** mais un jour Rouqui le chat décida de mettre des pièges de partout.

7. Conclusion

25 Cette étude nous a permis d'explorer différents scénarios envisageables pour le traitement des tokens omis dans le corpus Scolinter. Cela nous a permis de déterminer quelle méthode peut être la plus efficace pour reconstruire les tokens omis afin de minimiser le bruit sur les analyses morphosyntaxiques et en dépendance. Ces résultats sont prometteurs, avec des performances atteignant entre 97 % et 98 % selon les différentes métriques utilisées pour évaluer les résultats de l'étude, c'est-à-dire UAS, LAS, LS et précision sur le POS limitées aux phrases contenant des omissions. Néanmoins, cette méthode présente également des limites, notamment lorsque les substitutions suggérées ne semblent pas cohérentes avec le contexte du texte analysé ou lorsque l'éventuel token à remplacer n'est pas évident à prévoir même pour les chercheurs humains. Le scénario qui exploite un modèle de langage de type transformer a été retenu pour la suite de notre travail, et sera utilisé pour compléter les versions normalisées des textes du corpus Scolinter. Ces textes seront ensuite annotés en chaînes de référence, avec le but d'étudier à terme le développement des phénomènes de cohésion/cohérence textuelles dans ce corpus.

26 La suite du travail présenté dans cette contribution consistera à développer ultérieurement la méthode transformer FlauBERT pour générer une liste de tokens prédits, et de sélectionner au sein de la liste le premier token dont le POS correspond à la catégorie grammaticale préconisée par les chercheurs lors de la normalisation du texte, catégorie indiquée dans la balise d'omission en attribut (ex. si la balise d'omission contient l'attribut pronom, le programme sélectionne le premier pronom disponible dans la liste).

27 Cette combinaison entre l'automatisation et le processus de réflexion déjà entrepris par les chercheurs nous semble être la solution la plus efficace à la fois d'un point de vue technique et temporel pour répondre à la problématique du bruit causé par ces omissions de mots dans les textes du corpus Scolinter.

BIBLIOGRAPHY

Barletta M. (2024). « Annotation de la continuité référentielle dans un corpus scolaire – premiers résultats », *in* Balaguer M., Bendahman N., Ho-Dac L.-M. *et al.* (éd.) *35es Journées d'Études sur la Parole (JEP) 31e Conférence sur le Traitement Automatique des Langues Naturelles (TALN) 26e Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RECITAL)*. ATALA & AFPC, 28-41. https://inria.hal.science/hal-04622985.

Barletta M. & Ponton C. (2023, avril). « La question de la normalisation des écrits scolaires pour leur traitement automatique. Le cas de l'omission de mots », *Bruit de fond ou valeur ajoutée? Gérer le bruit lors des traitements informatiques des corpus linguistiques*. https://hal.science/hal-04090669.

Doquet C. & Ponton C. (2021). « Écrire de l'école à l'université : Corpus, traitements, analyses outillées. Présentation », *Langue francaise* 211(3) : 11-20. https://doi.org/10.3917/lf.211.0011.

Garcia-Debanc C., Rebeyrolle J. & Ho-Dac L.-M. (2021). « La continuité référentielle dans le corpus RÉSOLCO : Méthode d'annotation et premières analyses », *Langue francaise* 211(3) : 99-114.

Honnibal M., Montani I., Van Landeghem S. & Boyd A. (2020). *spaCy : Industrial-strength Natural Language Processing in Python*. https://doi.org/10.5281/zenodo.1212303.

Jurafsky D. & Martin J. H. (2024). *Speech and Language Processing : An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd éd.). https://web.stanford.edu/~jurafsky/slp3/.

Kodandaram S. R., Honnappa K. & Soni K. (2021). « Masking private user information using Natural Language Processing », *International Journal of Advance Research in Computer Science and Management* 7 : 1753-1763.

Le H., Vial L., Frej J., Segonne V., Coavoux M., Lecouteux B., Allauzen A., Crabbé B., Besacier L. & Schwab D. (2020). « FlauBERT : Unsupervised Language Model Pre-training for French », *in* Calzolari N., Béchet F., Blache P. *et al.* (éd.) *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, 2479-2490. https://aclanthology.org/2020.lrec-1.302.

Ponton C., Gutiérrez-Caceres R., Teruggi L., Farina E., Brissaud C. & Wolfarth C. (2021). « Scolinter : Un corpus trilingue. L'exemple de la segmentation en mots », *Langue française* 211(3) : 37-50. https://doi.org/10.3917/lf.211.0037.

Wolfarth C. (2019). *Apport du TAL à l'exploitation linguistique d'un corpus scolaire longitudinal*, Thèse de doctorat, Université Grenoble Alpes. https://tel.archives-ouvertes.fr/tel-02517320.

Wolfarth C., Brissaud C. & Claude P. (2018). « Transcrire et normer un corpus scolaire : pour quelles analyses? », *in* Brissaud C., Dreyfus M. & Kervyn B. (éd.) *Repenser l'écriture et son évaluation au primaire et au secondaire*. Presses universitaires de Namur, 121-145. https://books.openedition.org/pun/5283?lang=en.

Wolfarth C., Ponton C. & Totereau C. (2017). « Apports du TAL à la constitution et à l'exploitation d'un corpus scolaire au travers du développement d'un outil d'annotation orthographique », *Corpus, Spécificités et contraintes des grands corpus de textes scolaires : problèmes de transcription, d'annotation et de traitement* 16. https://doi.org/10.4000/corpus.2796.

NOTES

1. Le corpus Scoledit est disponible au lien suivant : https://scoledit.org/scoledition/.
2. Le corpus Scolinter est disponible au lien suivant : http://scoledit.org/scolinter/.
3. En particulier, dans notre travail nous allons annoter les maillons des chaînes faisant référence aux personnages animés dans les écrits des élèves, ce qui délimite la tâche d'annotation, mais qui n'exclue pas le fait que ces omissions concernent très probablement des personnages à annoter.
4. Ou de manière plus fréquente, dans les textes des langues où le sujet explicite peut être omis comme l'italien et l'espagnol, langues présentes dans notre corpus.
5. Par exemple, dans l'anonymisation des données personnelles dans les corpus (Kodandaram *et al.* 2021).
6. Le modèle utilisé pour cette tâche est le modèle pré-entraîné FlaubertWithLMHeadModel, dont la documentation est disponible au lien suivant https://huggingface.co/docs/transformers/model_doc/flaubert#transformers.FlaubertWithLMHeadModel.
7. Le seul cas de texte qui n'a pas obtenu des métriques à 1 sur la totalité du texte tout en contenant une omission de type pronom relatif est dû à la présence d'une deuxième balise d'omission dans le texte, c'est-à-dire d'une deuxième omission présente dans une autre phrase du texte.

ABSTRACTS

This paper addresses the treatment of noise caused by word omissions in a corpus of school writings, in order to facilitate their subsequent automatic processing. While a normalization step may facilitate the processing of these texts, certain linguistic expressions remain challenging to comprehend, particularly in instances where the writer omits words from the text. The present contribution proposes three automatic and semi-automatic potential solutions to this problem. The first method employs a "mask" token in the form of xxx. The second is a semi-automatic approach whereby each morpho-syntactic category proposed during normalization is replaced by the corresponding "prototypical word." The third involves a FlauBERT method, using this language model to "reconstruct" the most probable token in the text. The three methods are evaluated quantitatively, and the results obtained using method 3, which proved to be the most effective in the context of our research, are also presented qualitatively.

Cette contribution porte sur le traitement du bruit provoqué par les omissions de mots dans un corpus d'écrits scolaires en vue de leur traitement automatique. Même si une étape de normalisation rend possible le traitement de ces textes très fautifs, certains éléments de langage demeurent difficiles à appréhender notamment dans le cas où le scripteur omet des mots dans le texte. Nous proposons dans cette contribution trois méthodes possibles de traitement automatique ou semi-automatique pour résoudre cette problématique : (1) méthode exploitant un token « masque » de forme xxx ; (2) méthode semi-automatisée où chaque catégorie morpho-syntaxique proposée lors de la normalisation est remplacée par le même « mot prototypique » ; (3) méthode FlauBERT où le modèle de langage est utilisé pour « reconstruire » le token le plus probable dans le texte. Nous évaluons de manière quantitative ces trois méthodes pour ensuite présenter qualitativement les résultats obtenus à travers la méthode (3), la plus efficace dans le contexte de notre recherche.

INDEX

Mots-clés: écrits scolaires, TAL, normalisation, analyse morphosyntaxique
Keywords: children corpus, NLP, normalization, morphosyntactic analysis

AUTHORS

MARTINA BARLETTA

Laboratoire LIDILEM, Université Grenoble Alpes
Dottorato in Educazione nell'età contemporanea, Università Milano-Bicocca

CLAUDE PONTON

Laboratoire LIDILEM, Université Grenoble Alpes