Détecter une erreur dans les phrases coordonnées au sein des rédactions universitaires

Laura Noreskal[1] Iris Eshkol-Taravella[1] Marianne Desmets[2]
(1) MoDyCo UMR 7114, 200 avenue de la République, 92000 Nanterre, France
(2) LLF UMR 7110, 5, rue Thomas Mann, 75205 Paris cedex 13
Laura.noreskal@parisnanterre.fr, ieshkolt@parisnanterre.fr,
marianne.desmets@parisnanterre.fr

RÉSUMÉ

Beaucoup d'étudiants rencontrent des difficultés dans la maîtrise du français écrit. Sur la base d'une enquête linguistique préliminaire, il est apparu que les constructions syntaxiques comprenant des coordinations et des constructions elliptiques forment des contextes linguistiques sensibles aux erreurs ou aux maladresses dans les écrits des étudiants. Dans le cadre du projet écri+, notre recherche vise à développer un outil de détection automatique de phrases coordonnées erronées dans les rédactions des étudiants afin de leur permettre de s'auto-former en expression écrite. Après avoir constitué le corpus de phrases coordonnées extraites des différents écrits universitaires (exercices, examens, devoirs, rapports de stage et mémoires), nous avons établi une typologie des erreurs qui a servi de modèle pour l'annotation du corpus. Nous avons entrainé premièrement des classifieurs (Random Forest, SVM, CamemBERT et FlauBERT) afin de détecter deux étiquettes: erronée et correcte puis, dans un second temps, un classifieur multi-label pour diagnostiquer l'erreur.

ABSTRACT

Detecting error in coordinated sentences in students' writings.
Many students have difficulties in mastering written French. Based on a preliminary linguistic investigation, it appeared that syntactic constructions, including coordinations and elliptical constructions, are linguistic contexts sensitive to errors or awkwardness in students' writing. As part of écri+ project, the aim of our research is to develop a tool for automatic detection of erroneous coordinated sentences in students' essays in order to enable them to train themselves in written expression. After having constituted the corpus of coordinated sentences extracted from different academic writings (exercises, exams, homework, internship reports and dissertations), we established a typology of errors which served as a model for the annotation of the corpus. We first trained classifiers (Random Forest, SVM, CamemBERT and FlauBERT) to detect two labels: erroneous and correct, and then, in a second step, a multi-label classifier to diagnose the error.

MOTS-CLES : écri+, erreurs syntaxiques, phrases coordonnées, typologie d'erreurs, apprentissage automatique de surface, apprentissage profond, rédaction des étudiants.

KEYWORDS: écri+, syntactic errors, corpus, coordinated sentences, errors typology, machine learning, deep learning, student writings

1 Contexte de la recherche

Beaucoup d'étudiants rencontrent des difficultés dans la maîtrise du français écrit. Face à cette situation un réseau d'universités mobilisant une centaine d'enseignants et de chercheurs s'est associé au projet écri+[1] pour développer des méthodes pédagogiques innovantes. En se basant sur des recherches en informatique, en didactique et en linguistique, le projet vise à proposer une solution nationale pour accompagner, former, évaluer et certifier les étudiants du premier cycle universitaire. Sur la base d'une enquête linguistique préliminaire, il est apparu que les constructions syntaxiques comprenant des coordinations et des constructions elliptiques forment des contextes linguistiques sensibles aux erreurs ou aux maladresses dans les écrits des étudiants (section 3.1). Notre tâche dans ce projet est de développer un outil capable de détecter automatiquement les erreurs syntaxiques intra-phrastiques dans les constructions coordonnées afin de permettre aux étudiants de s'auto-former en expression écrite. Pour repérer les constructions coordonnées dans les rédactions des étudiants, nous nous sommes basées sur les travaux de Martinet (1980), de Goosse et al. (2008), de Riegel et al. (2009), d'Abeillé et Godard (2021) qui distinguent deux types de coordination : la coordination explicite qui requiert un coordonnant et la coordination implicite qui n'en attend pas. De plus, le rôle des éléments conjoints est important car il doit être respecté à chaque ajout. Nous avons décidé d'étudier les structures qui contiennent de la coordination explicite, soit des conjonctions de coordination (*mais, ou, et, or, ni, car, soit. . . soit*) ou des adverbes de liaison (*puis, ensuite, cependant, néanmoins*. . .).

2 Constitution des corpus

Dans l'optique d'observer les constructions coordonnées et de relever les différents types d'erreurs récurrentes dans les rédactions des étudiants, il était nécessaire de constituer un corpus. Plusieurs méthodes étaient alors envisageables : (1) un *corpus contrôlé* (Tellier, 2012) nécessitant de préparer un protocole expérimental qui inciterait les étudiants à produire des phrases coordonnées ; (2) un *corpus semi-contrôlé*, c'est-à-dire un ensemble de données langagières produites lors d'une formation comme les mémoires ou les devoirs maison ; (3) le recueil d'un corpus écologique constitué de productions naturelles des étudiants telles que les notes de cours. Parmi ces trois types de corpus, la deuxième solution nous a semblé plus pertinente. En effet, la mise en place d'un protocole expérimental demande de proposer une consigne qui assure d'avoir des phrases coordonnées, une consigne pas assez claire ou trop exigeante pourrait donner des résultats inutilisables. Quant au corpus écologique, il est difficile de savoir si les étudiants utilisent beaucoup de structures coordonnées dans leurs prises de notes. Il nous a donc semblé plus pertinent de collecter un corpus semi-contrôlé constitué de productions réalisées dans le cadre de la formation des étudiants qui selon nous contiendraient davantage de structures coordonnées que les productions naturelles.

2.1 Corpus de rédactions

Nous avons collecté les productions dites « évaluatives », c'est-à-dire les productions réalisées dans le but d'être évaluées par un enseignant dans le cadre d'un enseignement supérieur. Parmi ces productions, nous avons retenu 4 types de rédactions : 139 devoirs maison, 167 exercices faits en classe, 47 rapports de stage et 27 mémoires, ce qui correspond au total à 380 rédactions. Les écrits proviennent de différents niveaux d'étude allant de la première année de licence à la deuxième année de master. Les domaines disciplinaires concernés sont les sciences du langage, l'histoire et le droit.

[1] anr-17-NCUN-00015

2.2 Corpus de phrases coordonnées

À partir du corpus de rédactions, nous avons pu extraire les structures coordonnées en utilisant des patrons morphosyntaxiques créés avec l'outil Unitex (Paumier, 2011) qui reconnaissent les conjonctions de coordination et les adverbes de liaison. Sur un total de 39 692 phrases, 6 645 phrases contenant de la coordination ont été extraites (voir tableau 1).

Types	**Nombre de rédactions**	**Total des phrases**	**Phrases avec coordination**
Devoirs maison	139	7 635	1 467
Exercices	167	2 502	593
Mémoire	27	13 244	1890
Rapports	47	16 311	2695
Total	380	39 692	6645

TABLEAU 1 : Composition du corpus de phrases coordonnées

3 Annotation manuelle

Le processus d'annotation a été mis en place par 3 annotateurs (une experte et deux linguistes non-experts). Cette annotation répond à deux objectifs : (1) proposer une analyse quantitative des erreurs dans les rédactions étudiantes ; (2) créer un corpus de référence pour l'apprentissage automatique de ces erreurs. L'annotation consistait à renseigner sur la présence ou non d'une erreur dans une structure coordonnées et sur son type. Nous avons annoté d'abord une partie des phrases coordonnées avec les étiquettes **correcte** et **erronée**. À partir de là, nous avons pu établir une typologie des erreurs et nous concentrer sur une annotation en types d'erreur.

3.1 Typologie des erreurs

Après avoir observé les phrases erronées collectées, nous avons pu distinguer 8 types d'erreurs que nous avons regroupés en 5 catégories : prépositions, syntagmes conjoints, accords sujet-verbe distant, ponctuation et autres.

3.1.1 Les prépositions

Les prépositions sont souvent sujettes aux erreurs dans les constructions coordonnées. Trois sous-types d'erreurs sont distingués : l'ajout d'une préposition non-attendue (PREP ADD) (1), le remplacement d'une préposition par une autre (PRED REMP) (2) et l'absence de préposition (PREP ABS) (3).

(1) *Avant de visionner la comédie musicale, il faudra étudier avec les élèves la période révolutionnaire pour comprendre les raisons de la Révolution et **de** rendre cette activité ludique mais pédagogique.

Dans cet exemple, le problème vient de la préposition non-attendue *de* qui forme un syntagme prépositionnel avec le syntagme verbal *rendre cette activité ludique mais pédagogique*.

(2) *Le fait de les aider à se construire et **à** les voir grandir, tout en leur apportant un savoir doit être gratifiant et réjouissant.

L'exemple (2) est considéré comme erroné car le syntagme prépositionnel *à les voir grandir* est introduit par la mauvaise préposition (*à* au lieu de *de*). En effet, le syntagme *à les voir grandir* est considéré comme le complément du verbe *aider* or il s'agit en réalité du complément de *fait* qui doit être introduit par *de*.

(3) *Ils illustrent leur propos en appliquant cette analyse aux appendices et les formules illocutoires, les actes indirects et les questions biaisées.

En (3), le problème est lié aux compléments du verbe *appliquer*, lequel attend généralement un complément direct mais peut prendre également un complément indirect introduit par la préposition *à*. Dans la phrase (3), le premier complément indirect *aux appendices* respecte la valence du verbe mais pas les autres compléments.

3.1.2 Les syntagmes conjoints

— Mauvaise cohérence des groupes syntaxiques (MCGS)

(4) *Par ailleurs, il appartient à tous les personnels de transmettre aux élèves les valeurs et doivent avoir un devoir de neutralité.

Dans l'exemple (4), les deux conjoints n'ont pas le même rôle syntaxique ni la même forme. De plus, il n'est pas possible de retrouver le sujet du syntagme verbal *doivent avoir un devoir de neutralité* qui représente le second conjoint.

— Grande distance entre conjoints (DIST CONJOINT)

(5) *S'associer avec des associations telles que celles du Téléthon qui permet de sensibiliser les élèves, de transmettre des valeurs républicaines, ainsi que Nettoyons la Nature.

Dans l'exemple (5), les deux conjoints (*Téléthon* et *Nettoyons la Nature*) reliés par *ainsi que* sont trop éloignés l'un de l'autre. Cela peut produire une incompréhension chez le lecteur.

3.1.3 Les accords entre sujet et verbe distant

(6) *Le personnage de droite est assis sur un tabouret et as une corpulence fine.

L'exemple (6) contient une erreurs d'accord entre le sujet et le verbe : le verbe avoir est conjugué à la deuxième personne du singulier alors que le sujet est à la troisième personne du singulier.

3.1.4 La ponctuation

Les erreurs de ponctuation sont également très fréquentes dans les phrases du corpus. Deux sous types d'erreurs sont proposés : les structures lourdes (7) et les problèmes d'absence de ponctuation (8).

— Structure Lourde (SL)

(7) *L'auteur dénonce les contrats sociaux très importants entre les pays du nord et les pays du sud plus précisément, la faim à l` échelle mondiale et il va donc accentuer ses disparités à travers une caricature qui à une visée humoristique mais également critique puisque elle met en relief les pays du Nord , riches et industrialisés, dominé par la Triade qui regroupe les grands pôles économiques du monde et les pays du Sud ,

pauvres et désigné comme le « Tiers-monde »qui réfléchissent à une solution pour lutter contre la famine autour d'un repas .

L'exemple (7) est considéré comme erroné car il comprend de nombreuses propositions imbriquées les unes dans les autres, ce qui rend la compréhension difficile.

— Absence de ponctuation (PONC ABS)

(8) *Chaque année des milliers de sacs à main tous différents les uns des autres rentrent sur le marché mondial et ces sacs à main répondent à une demande croissante de la part des femmes donc je me dit que c'est vraiment un effet de mode dont certaines femmes ne pourraient plus se passer.

Dans l'exemple (8), il n'y a aucune autre ponctuation dans le texte, à part le point final. Plusieurs segments s'enchainent avec les conjonctions de coordination comme seuls liens. Cela peut alors demander plus d'efforts cognitivement pour la lecture et la compréhension.

3.1.5 Autres

La classe Autres comprend les erreurs qui ne font pas partie des classes précédentes. Les phrases réunies dans Autres contiennent des erreurs différentes mais peu fréquentes comme la présence de *pas* avant *ni* (9).

(9) Pour ma part, je ne connaissais pas ni le master français langue étrangère (FLE) ni le métier du traitement automatique des langues.

3.2 Accord inter-annotateurs

Après avoir développé un guide d'annotation basé sur la typologie ci-dessus, nous avons lancé une campagne d'annotation et avons calculé un accord inter-annotateurs entre deux annotateurs non-experts. Pour vérifier la cohérence de la typologie, une experte a formé les non-experts sur la reconnaissance des différents types erreurs. À la suite de la formation, les deux annotateurs ont annoté la présence ou non d'erreurs et le type d'erreurs dans 200 phrases. L'accord de la présence ou non d'une erreur a été calculé avec le kappa de Cohen et le score a atteint 0,88, ce qui est un accord excellent selon Landis & Coch (1977). L'accord inter-annotateur pour le type d'erreurs s'est élevé à 0,72 avec le kappa de Cohen. Les types d'erreurs les mieux reconnus ont été *DIST CONJ* et *PRED REMP*. Concernant les erreurs *PRED REMP*, lors de la lecture d'un texte, il se peut que le non-respect de la sous-catégorisation des verbes, noms ou adjectifs interpelle le lecteur qui repère ainsi qu'il s'agit de la mauvaise préposition. Quant à l'erreur *DIST CONJ*, elle touche directement à la compréhension de la coordination. Le lecteur a du mal à identifier les conjoints ce qui rend l'erreur plus facilement repérable. Une fois la typologie validée, 3145 phrases ont été annotées dont 1153 phrases erronées. La répartition des erreurs dans les phrases annotées est présentée dans le TABLEAU 2.

Types d'erreurs	Nombres
Préposition ajoutée (PREP ADD)	29
Préposition remplacée (PREP REMP)	44
Préposition absente (PREP ABS)	139
Mauvaise cohérence entre les groupes syntaxiques (MCGS)	35
Distance entre conjoint (DIST CONJ)	25
Mauvais accord sujet-verbe (MASV)	33

Structure lourde (SL)	244
Ponctuation absente (PONC ABS)	155
Autres	449
Total	1153

TABLEAU 2 : Constitution du corpus annoté

4 Détection automatique

De nombreuses recherches ont été faites sur la détection automatique des erreurs. Depuis quelques années la détection et la correction automatiques d'erreurs grammaticales (Grammatical Errors Detection, GEC en anglais) ainsi que les outils d'aide à la rédaction reçoivent beaucoup d'attention. Majoritairement développés dans le but d'aider les apprenants allophones (Garnier, 2014), les modèles de détection d'erreurs permettent souvent de reconnaître les erreurs d'accords (Fay-Varnier, 1990 ; Souque, 2014) ou les erreurs lexicales (Yuan *et al*., 2019). La détection d'erreurs orthographiques en français n'est pas en reste : Cordial de Synapse (1995), Antidote de Druide Informatique (1996) et ProLexis des Editions Diagonal (1997). Ces trois outils dominent le marché francophone qui, depuis quelques années, ne voit que très peu de nouvelles alternatives. De plus, les travaux se basant sur la syntaxe sont peu nombreux (Clément *et al.*, 2009). En anglais, trois types d'outils d'aide à l'écriture existent (Jourdan *et al*., 2023) : les outils de révision de phrases (Ouyang *et al*., 2022), les correcteurs grammaticaux (Tsai *et al.*, 2020) et les outils d'annotation de structures rhétoriques tels que AcaWriter (Knight *et al*., 2020). En nous concentrons sur le français, nous proposons de contribuer à la recherche dans ce domaine.

4.1 Détection binaire : erronée / correcte

La première étape de la détection d'erreurs syntaxiques dans les phrases coordonnées porte sur la détection des étiquettes **correcte** et **erronée** en utilisant les méthodes de l'apprentissage automatique de surface avec les deux classifieurs SVM et Random Forest et l'apprentissage profond fondé sur les modèles français CamemBERT (Martin *et al*., 2019) et FlauBERT (Le et al., 2019) en tant que classifieurs en utilisant Simple Transformers[2]. Pour ces expériences, nous avons utilisé deux corpus d'apprentissage : un corpus équilibré (600 phrases correctes et 600 phrases erronées) et un corpus déséquilibré en faveur des phrases erronées (300 phrases correctes et 900 phrases erronées) afin de vérifier si le déséquilibre peut aider à mieux détecter les erreurs. Le corpus de test est composé de 400 phrases et contient autant de phrases correctes que de phrases erronées. La répartition des erreurs dans les différents corpus est présentée dans le tableau suivant :

Types d'erreurs	Corpus équilibré	Corpus déséquilibré
Autres	167	320
Autres + PREP ADD	2	4
Autres + SL + PONC ABS	0	3
DIST CONJ	12	16
DIST CONJ + Autres	2	0
DIST CONJ + PONC ABS	0	4
MASV	16	23
MASV + Autres	2	2
MASV + SL	4	2

[2] https://simpletransformers.ai/

MCGS	18	23
MCGS + Autres	2	2
MCGS + PREP REMP	4	4
PONC ABS	100	119
PONC ABS + Autres	0	15
PONC ABS + PREP ADD	0	2
PONC ABS + PREP REMP	0	2
PREP ABS	56	73
PREP ABS + Autres	2	2
PREP ABS + PONC ABS	2	4
PREP ABS + PREP ADD + SL	3	3
PREP ABS + SL	6	0
PREP ADD	13	15
PREP ADD + SL + Autres	3	3
PREP REMP	24	25
PREP REMP + Autres	0	4
SL	150	172
SL + Autre	10	32
SL + PREP ABS	0	4
SL + PREP ADD	0	2
SL + PREP REMP	2	5
Total	600	900

TABLEAU 3 : Constitution des corpus d'entrainement

4.1.1 Apprentissage de surface

Nous avons utilisé une méthode de classification supervisée en utilisant deux classifieurs : Support Vector Machine (SVM) et Random Forest car ils ont obtenu de bons scores lors de la classification de documents lors des campagnes DEFT (DEFT 2021 et DEFT 2022).

4.1.1.1 Prétraitement

Notre chaîne de prétraitement comprend l'étiquetage morphosyntaxique, ~~l~~a lemmatisation, l'analyse syntaxique en dépendances, le chunking et d'autres traits résultant d'une observation du corpus. Chaque phrase est représentée par un ensemble de 20 traits linguistiques regroupés en trois classes :

— les traits généraux souvent utilisés lors du prétraitement des textes :
- les tokens
- les lemmes obtenus avec Treetagger (Schmid, 1994)
- les parties du discours obtenus avec *Treetagger*
- les chunks détectés grâce à *TreeTagger*,
- les relations de dépendances détectés grâce à la bibliothèque Python *spaCy*
- les trigrammes

— les traits binaires :
- la présence d'un verbe transitif grâce au dictionnaire électronique des mots (DEM) de Dubois (Dubois et al., 2010): en observant les phrases erronées, il est apparu que la sous-catégorisation des verbes transitifs n'était pas respectée. Nous cherchons donc à savoir si la seule présence d'un verbe transitif peut jouer un rôle dans l'apparition des erreurs.
- la présence d'une préposition : nous avons repéré plusieurs problèmes liés aux prépositions tels que *PREP ADD* ou *PRED REMP*. Face à cela, nous nous

sommes demandé si la présence d'une préposition pouvait être reliée à la présence d'une erreur.
- la présence de *que* : Lors de notre observation du corpus, nous avons pu remarquer que beaucoup de phrases erronées contenaient des propositions introduites par *que*. Nous avons donc ajouté ce trait afin d'observer si l'utilisation de *que* peut être liée à la présence d'une erreur.

— les traits numériques :
- le nombre de mots : en ajoutant le nombre de mots, nous espérons trouver une corrélation entre la longueur de la phrase et la présence d'une erreur. Le but de cette observation n'est pas de prescrire les phrases longues mais plutôt de savoir si ce type de phrases est plus sujet aux erreurs.
- le nombre de verbes transitifs : après avoir observé une tendance au non-respect de la sous-catégorisation des verbes, nous avons pensé qu'il serait intéressant de tester l'hypothèse selon laquelle le nombre de verbes transitifs aurait un impact sur la présence d'erreurs.
- le nombre de conjonctions de coordination : ce trait permet de vérifier s'il y a une corrélation entre le nombre de coordonnants et la présence d'erreurs dans une phrase.
- le nombre de *que* : la présence de *que* dans une phrase sous-entend que celle-ci est une phrase complexe avec une proposition subordonnée. De ce fait, chaque *que* présent dans une phrase la complexifie. Ainsi, nous cherchons à savoir si la présence de cette complexité dans une coordination peut être corrélée avec la présence d'une erreur.
- le nombre de prépositions : l'ajout de ce trait prend en compte le fait que les prépositions posent problème dans les phrases coordonnées. Nous cherchons alors à savoir si le nombre de prépositions dans une phrase peut être relié à la présence d'une erreur.
- le nombre de *à*, *de*, *sur*, *pour*, *dans* : nous avons ajouté ces traits reportant le nombre de chaque préposition dans une phrase afin d'observer si certaines prépositions sont plus sujettes aux erreurs que d'autres.
- le nombre de *ainsi que* : la locution conjonctive *ainsi que* joue un rôle de coordonnant dans certains de ses emplois. Nous avons pensé qu'il serait intéressant d'observer cette locution afin de savoir si ses propriétés peuvent poser problème.

Afin de sélectionner les traits les plus pertinents pour l'apprentissage, nous avons utilisé un algorithme de sélection de traits : RFE (Recursive Features Elimination). Lors de son application en validation croisée avec 5 échantillons, l'algorithme supprimait un ou plusieurs traits non pertinents pour l'apprentissage à chaque échantillonnage. Les 11 traits suivants ont été sélectionnés : l'analyse en dépendance, le chunking, les lemmes, le nombre de conjonctions de coordination, le nombre de *de,* le nombre de mots, le nombre de prépositions, le nombre de verbes transitifs, les parties du discours, les tokens et les trigrammes.

4.1.1.2 Expériences et résultats

Lors des expériences, nous avons testé deux aspects : le type de données (corpus équilibré/corpus déséquilibré) et les algorithmes de classification (Random Forest/SVM). De plus, afin d'optimiser les performances de SVM et Random Forest, nous avons utilisé, lors de tous les tests, GridsearchCV, permettant de tester les différents paramètres d'un algorithme d'apprentissage pour en sélectionner les meilleurs.

Les expériences sur les deux corpus ont montré que les données déséquilibrées rendent la tâche de détection plus compliquée pour le classifieur. En effet, les résultats pour les données déséquilibrées sont mauvais avec des exactitudes de 0,5425 pour Random Forest et de 0,585 pour SVM. Quant aux résultats pour les données équilibrées, ils sont de 0,665 pour Random Forest et de 0,6325 pour SVM. Afin de mieux comprendre les résultats, nous avons observé les mesures de précision, rappel et f-mesure de chaque classe lors du test avec le corpus équilibré.

Random Forest	Précision	Rappel	F-mesure	Exactitude
Correcte	0,64	0,76	0,70	0,665
Erronée	0,70	0,57	0,63	

TABLEAU 4 : Random Forest : mesures de précision, rappel et f-mesure pour les deux classes

SVM	Précision	Rappel	F-mesure	Exactitude
Correcte	0,61	0,73	0,67	0,6325
Erronée	0,66	0,54	0,59	

TABLEAU 5 : SVM : mesures de précision, rappel et f-mesure pour les deux classes

Les résultats obtenus grâce à Random Forest montrent que la différence de f-mesure entre la classe correcte et la classe erronée est de 6%. Cependant, on remarque également que la précision pour la classe erronée (0.70) est plus élevée que celle de la classe Correcte (0.64). Cela signifie que le modèle est plus précis pour détecter les erreurs que pour détecter les phrases correctes. Néanmoins, le modèle rapporte moins de phrases erronées (0.57) que de phrases correctes (0.76).

4.1.2 Apprentissage profond

Pour l'apprentissage profond, nous avons utilisé deux modèles français pour une tâche de classification des phrases en **correcte**/**erronée** : CamemBERT et FlauBERT, avec *Simple Transformers*, une librairie d'*HuggingFace*. Les expériences ont été réalisées en variant le nombre d'époques.

	FlauBERT Corpus équilibré	FlauBERT Corpus déséquilibré	CamemBERT Corpus équilibré	CamemBERT Corpus déséquilibré
5 époques	0,5	**0,7**	**0,7125**	**0,77**
8 époques	0,71	0,5	0,695	0,73
10 époques	0,7	0,5825	0,685	0,72
15 époques	**0,715**	0,6275	0,6875	0,7375

TABLEAU 6 : Exactitudes des expériences menées en apprentissage profond

Nous observons que les résultats sont meilleurs que ceux que nous avons obtenus lors des expériences avec l'apprentissage de surface. Cela peut être dû au fait que les modèles utilisés pour l'apprentissage profond sont entraînés sur de gros corpus composés de phrases correctes, il pourrait être ainsi plus simple pour les modèles de détecter les phrases qui s'éloignent des phrases correctes apprises auparavant. Le meilleur résultat provient de l'expérience avec CamemBERT et le corpus déséquilibré lors des 5 époques (0.77).

	Précision	Rappel	F-mesure	Exactitude
Correcte	0,99	0,54	0,69	0,77

Erronée	0,68	0,99	0,81	

TABLEAU 7 : CamemBERT avec le corpus déséquilibré: mesures de précision, rappel et f-mesure pour les deux classes

Le modèle CamemBERT avec le corpus déséquilibré montre une très bonne précision (0,99) pour la classe correcte et un bon rappel (0,99) pour la classe erronée. La F-mesure de la classe erronée est aussi très élevée puisqu'elle atteint 0,81.

4.2 Détection multi-label

Un autre objectif de cette recherche est de réussir à diagnostiquer l'erreur en lui attribuant une étiquette qui détermine le type d'erreur mis en cause. Pour ce faire, nous avons utilisé la classification multi-label. Pour débuter la classification multi-label, nous avons gardé les données utilisées pour la classification binaire.

Lors de la classification multi-label, nous avons fait 4 expériences soit deux expériences avec CamemBERT et deux autres avec FlauBERT. Pour chaque modèle de langue, nous avons testé l'apprentissage avec un corpus déséquilibré et l'apprentissage avec un corpus équilibré. Le nombre d'époque a été figé à 5 afin de ne pas avoir de dégradation des résultats comme pour la classification binaire.

Modèles	Corpus	Bien classées	Mal classées	Exactitude
CamemBERT	Equilibré	18	182	4,5
	Déséquilibré	218	182	54,5
FlauBERT	Equilibré	83	317	20,75
	Déséquilibré	176	224	44

TABLEAU 8 : Exactitudes des expériences menées pour la détection multi-label

En observant les résultats nous remarquons que seul le modèle CamemBERT avec le corpus déséquilibré a une exactitude supérieure à 50%. En somme, les résultats sont assez mauvais puisque les modèles peinent à classer correctement les phrases. Pour mieux comprendre ces erreurs, nous avons observé les résultats des différentes classifications.

De façon générale, les modèles ont tendance à combiner l'étiquette *None,* qui correspond à l'absence d'erreur, à une autre étiquette d'erreur telles que *SL* et autres. Seul le modèle CamemBERT avec le corpus déséquilibré ne reproduit pas cette erreur. Quant au modèle CamemBERT avec le corpus équilibré, il est celui qui est le plus touché par cette erreur d'étiquetage puisqu'il comptabilise 200 phrases annotées *None* et *SL*. Les modèles FlauBERT sont aussi touchés par cette erreur. FlauBERT avec le corpus équilibré produit cette erreur sur 148 phrases alors que FlauBERT avec le corpus déséquilibré la produit sur 56 phrases.

Concernant les phrases bien classées, elles concernent en général les types d'erreurs *PREP ADD*, *PREP ABS* et MASV (*mauvais accord Sujet-Verbe*). Aussi, les phrases correctes sont bien classées par les modèles entrainés avec les corpus déséquilibrés alors que les autres modèles arrivent surtout à annoter les phrases avec les erreurs d'*absence de préposition*.

	VP	VN	FP	FN	Précision	Rappel	F-mesure
Autres	0	270	3	127	0	0	0

DIST CONJ	0	396	1	3	0	0	0
MASV	3	391	6	0	0,33	1	0,5
MCGS	0	396	1	3	0	0	0
PONC ABS	20	380	0	0	1	1	1
PREP ABS	11	389	0	0	1	1	1
PREP ADD	4	396	0	0	1	1	1
PREP REMP	3	396	0	1	0,75	1	0,86
SL	127	249	0	24	1	0,84	0,91

TABLEAU 9 : Mesures de la détection multilabel avec CamemBERT et le corpus déséquilibré

En observant les mesures sur ce modèle, on réalise que certains types d'erreurs sont mieux reconnus que d'autres. Les résultats montrent que ce modèle reconnaît très bien les erreurs PONC ABS, PREP ABS et PREP ADD mais qu'il est incapable de détecter les erreurs DIST CONJ, Autres et MCGS. Les modèles d'apprentissage profond n'indiquant pas réellement ce qui est pris en compte pour la détection, il est difficile de préciser la raison pour laquelle certaines erreurs seront mieux reconnues que d'autres. Cependant, il semblerait que les phrases correctes soient mieux reconnues lors des entrainements avec les corpus déséquilibrés car les erreurs étant plus nombreuses dans le corpus, il devient plus simple pour l'algorithme de différencier une erreur d'une phrase correcte. Pour les *PREP ABS*, la raison pour laquelle ils sont bien repérés par tous les modèles pourrait être due au fait qu'il y a plus de données que pour d'autres erreurs. Une augmentation du corpus serait alors bénéfique pour tous les autres types d'erreurs.

En résumé, la détection multi-label n'a pas donné de résultats probants puisque nous avons obtenu des exactitudes très faibles et des erreurs d'étiquetage combinant l'étiquette *None* à d'autres étiquettes d'erreurs. En utilisant l'apprentissage profond, il est difficile de savoir ce qui permet de classer une phrase dans un type d'erreur plutôt qu'un autre, ainsi il serait intéressant d'envisager une approche symbolique pour cibler et mieux localiser les erreurs dans les phrases. En testant ce type de classification automatique, nous avons cependant pu observer que certains types d'erreurs sont mieux reconnus que d'autres.

5 Conclusion et perspectives

Cette recherche vise à détecter automatiquement les phrases coordonnées erronées dans les rédactions des étudiants. Après avoir constitué le corpus de phrases coordonnées extraites des différents écrits universitaires (exercices, examens, devoirs, rapports de stage et mémoires), nous avons établi une typologie des erreurs que nous avons validée par un accord inter-annotateurs. Par la suite, nous avons procédé à plusieurs expériences portant sur l'apprentissage profond et l'apprentissage de surface. Les premières expériences concernaient la détection des étiquettes **erronée** et **correcte**. L'apprentissage profond a donné de meilleurs résultats puisque nous avons obtenu une exactitude de 0,77 avec le modèle CamemBERT. Puis, nous avons testé la détection multi-label en utilisant l'apprentissage profond. Nos résultats n'ont pas été probants mais ils laissent à penser qu'il faudrait sûrement observer plus précisément les résultats et peut-être aussi envisager d'autres approches comme des méthodes symboliques qui permettraient de mieux cibler les erreurs. Il serait également utile de revoir la typologie des erreurs afin de voir si les types d'erreurs les moins bien reconnus comme structures lourdes (SL) et absence de ponctuation (PREP ABS) sont vraiment pertinents pour notre recherche. En effet, lors des apprentissages, ce sont les types d'erreurs qui ont été les moins bien reconnus. Ces deux types se fondent sur l'utilisation de la ponctuation, ainsi, il se pourrait que les règles adoptées pour catégoriser les erreurs de ponctuation ne soient pas assez distinctives pour que ces erreurs soient repérables par les algorithmes.

Références

Abeillé, A., Godard, D., en collab. Avec Delaveau A. & Gautier A. (2021). *La Grande Grammaire du français*. Actes Sud / Imprimerie Nationale.

Clément, L., Gerdes, K. et Marlet, R. (2009). Grammaires d'erreur–correction grammaticale avec analyse profonde et proposition de corrections minimales In Nazarenko, A. & Poibeau, T. Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles),* p. 158–167, Senlis, France.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20.1, p. 37-46.

Dubois, J. et Dubois-Charlier, F. (2010). La combinatoire lexico-syntaxique dans le Dictionnaire électronique des mots. Les termes du domaine de la musique à titre d'illustration. *Langages* 3, p. 31-56.

Fay-Varnier, C. (1990). *Aide à la détection de fautes grammaticales par une analyse progressive des phrases*. Thèse de doctorat. Institut National Polytechnique de Lorraine.

Garnier, M. (2014). *Utilisation de méthodes linguistiques pour la détection et la correction automatisées d'erreurs produites par des francophones écrivant en anglais*. Thèse de doctorat. Université Toulouse le Mirail-Toulouse II.

Goosse, A. et Grevisse, M. (2008). *Le bon usage*. De Boeck Superieur. Louvain-la-Neuve.

Grouin, C. et Illouz, G. (2022). Notation automatique de réponses courtes d'étudiants : présentation de la campagne DEFT 2022 (Automatic grading of students' short answers : presentation of the DEFT 2022 challenge). In Y. Estève, T. Jiménez, T. Parcollet, M. Zanon Boito, Éds., *Actes de TALN 2022 (Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, p. 1-10, Avignon, France.

Grouin, C., Grabar, N. et Illouz, G. (2021). Classification de cas cliniques et évaluation automatique de réponses d'étudiants : présentation de la campagne DEFT 2021. In *Actes de TALN 2021 (Traitement Automatique des Langues Naturelles. Atelier DÉfi Fouille de Textes (DEFT)*, p. 1-13, Lille, France.

Jourdan, L., Boudin, F., Dufour, R., Hernandez, N. (2023). Text revision in Scientific Writing Assistance: An Overview. In *13th International Workshop on Bibliometric-enhanced Information Retrieval (BIR 2023)*, Dublin (IE), Ireland. ⟨hal-04053934⟩.

Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P. (2020) Acawriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research.*

Landis, J. R. et Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, p. 363-374.

Le, H., Vial, L., Frej, F., Segonne, V., Coavoux, M., Lecouteux, B., Allauzen, A., Crabbé, B., Besacier, L. et Schwab, D. (2020). FlauBERT: Unsupervised Language Model Pre-training for

French. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 2479–2490, Marseille, France. European Language Resources Association.

Li, X., Michel, P., Anastasopoulos, A., Belinkov, Y., Durrani, N., Firat, O., Koehn, P., Neubig, G., Pino, J. et Sajjad, H. (2019). Findings of the First Shared Task on Machine Translation Robustness. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1),* p. 91–102, Florence, Italy. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. et Stoyanov, V. (2019). RoBERTa : A Robustly Optimized BERT Pretraining Approach. arXiv abs/1907.11692. Retrieved from https://arxiv.org/abs/1907.11692.

Noreskal, L., Eshkol-Taravella, I. & Desmets, M. (2021). Erroneous Coordinated Sentences Detection in French Students' Writings. *Communications in Computer and Information Science 1463*, p. 586-596.

Martin, L., Müller, B., Suárez, P. J. O., Dupont, Y., Romary, L., Villemonte de la Clergerie, E. et Seddah, D. et Sagot, B. (2019). « CamemBERT: a Tasty French Language Model». arXiv: 1911.03894.

Martinet, André. (1980). *Éléments de linguistique générale*. Collection U Prisme. Albin Michel.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022) Training language models to follow instructions with human feedback, arXiv preprint arXiv:2203.02155.

Suárez, O., Javier, P., Romary, L. et Sagot, B. (2020). A Monolingual Approach to Contextualized Word Embeddings for Mid-Resource Languages. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. p. 1703-1714. Association for Computational Linguistics.

Riegel, M., Pellat, J-C. et Rioul, R. (2008). *Grammaire méthodique du français*. Édition PUF. Paris.

Schmid, H. (1994). TreeTagger-a language independent part-of-speech tagger. Institut Für Maschinelle Sprachverarbeitung: Universität Stuttgart.

Souque, A. (2014). *Modèle de vérification grammaticale automatique gauche-droite*. Thèse de doctorat. Université de Grenoble.

Tellier, M. (2012). De l'usage du corpus semi-contrôlé dans la recherche en didactique des langues. *Rencontres de l'ASDIFLE. FLE : L'instant et l'histoire 49 et 50*. Clé International, p. 39-47.

Tiedemann, J. (2012). Parallel Data, Tools and Interfaces in OPUS ». *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, Turkey: European Language Resources Association (ELRA), p. 2214-2218.

Tsai, C.-T., Chen, J.-J., Yang, C.-Y., Chang, J. S. (2020) LinggleWrite: a coaching system for essay writing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, p. 127–133, Association for Computational Linguistics.

Wenzek, G., Lachaux, M-A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A. et Grave, E. (2020). CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In Proceedings of the Twelfth Language Resources and Evaluation Conference, p. 4003–4012, Marseille, France. European Language Resources Association.

Yuan, Zheng, Felix Stahlberg, Marek Rei, Bill Byrne et Helen Yannakoudakis (2019). Neural and FST-based approaches to grammatical error correction. *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 228-239.