Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Prise en compte de la dimension collocative dans la notation automatique de productions écrites en français langue étrangère

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Ingénierie Multilingue

par

Fabienne AUFFRET

Directeur de mémoire :

Mathieu Valette

Encadrant :

Dominique Casanova

Année universitaire 2020/2021

Attestation de non-plagiat

Déclaration sur l'honneur

Je soussignée Fabienne Auffret, déclare avoir rédigé ce travail sans aides extérieures ni sources autres que celles qui sont citées.
Toutes les utilisations de textes préexistants, publiés ou non, y compris en version électronique, sont signalées comme telles.
Ce travail n'a été soumis à aucun autre jury d'examen sous une forme identique ou similaire, que ce soit en France ou à l'étranger, à l'université ou dans une autre institution, par moi-même ou par autrui.

Date : 20 novembre 2021

Signature

Table des matières

LISTE DES FIGURES

Liste des tableaux

AVANT-PROPOS

Remerciements

Je remercie l'ensemble des professeurs du Master TAL de l'INALCO qui m'a transmis les bases des connaissances nécessaires à la réalisation de ce travail.
Je remercie, en particulier, Monsieur Mathieu Valette, enseignant à lINALCO et directeur du présent mémoire pour ses conseils et avis éclairés, ainsi que Monsieur Dominique Casanova, mon tuteur de stage, pour son accompagnement bienveillant et son aide tout au long de mon stage.
J'aimerais remercier aussi de manière plus générale la Chambre de Commerce et d'Industrie Paris Île-de-France pour son accueil, et plus particulièrement le Français des affaires.

Résumé

Ce travail s'inscrit dans la continuité de celui de plusieurs personnes, qui ont déjà construit un modèle et ses caractéristiques pour classer des copies de tests de français (pour l'obtention du statut de résident ou de la naturalisation, en France ou au Canada), ceci dans un premier temps jusqu'à un niveau intermédiaire, puis sur l'ensemble des niveaux de langue A1 à C2 du Cadre européen commun de référence pour les langues (Coe, 2001).
De nombreuses études ayant démontré que les collocations au sens large, c'est-à-dire aussi sous leur aspect « collostructions » (collocations qui prennent en compte la structure grammaticale - [Stefanowitsch and Gries, 2003], [Paquot, 2018]) permettent de départager entre eux les apprenants de niveau avancé (B2 à C2), nous allons donc créer des caractéristiques qui en sont issues, pour non seulement tenter d'améliorer le modèle existant, mais aussi utiliser des indicateurs pédagogiquement parlants car linguistiquement pertinents ([Valette and Eensoo, 2014] et non pas uniquement distributionnels.

Abréviations :

— CECRL : Cadre européen commun de référence pour les langues
— TEF : Test d'évaluation de français

Mots clés :
Collocations, apprentissage automatique, compétence linguistique, français langue étrangère

Introduction

Présentation générale

Nombreux sont les pays, comme la France et le Canada, qui exigent la maîtrise d'un niveau de compétence minimum dans la (ou une des) langue(s) nationale(s) dans le cadre du parcours d'immigration ou de l'accès à la citoyenneté et utilisent donc des tests de français afin d'évaluer le niveau d'un candidat : il est donc primordial de s'assurer de la justesse de l'évaluation.
Le rôle du Français des affaires est de promouvoir un français utile et professionnel par le biais de tests et certifications de la maîtrise de la langue française générale ou professionnelle. Son test d'évaluation de français est notamment reconnu par les pouvoirs publics français, québécois et canadiens dans le cadre des démarches d'immigration et d'accès à la citoyenneté. Ce test étant sous format électronique, il permet la constitution de corpus de textes à partir des productions écrites et orales des candidats donnant leur accord.
Le traitement automatique des langues et de l'apprentissage supervisé s'orientant depuis quelques années vers des méthodes à la fois symboliques et utilisant l'apprentissage supervisé, il nous a semblé pertinent d'en exploiter tous les aspects dans le cadre de l'élaboration d'un outil de notation automatique.
Le but de ce travail est donc d'étudier si les collocations pourraient améliorer un modèle déjà existant, qui prédit correctement 76,8 % des copies, et élargir son spectre d'application aux copies de plus haut niveau puisque de nombreux travaux démontrent que les collocations, en fonction de leur rareté, sont liées au niveau de compétence en langue.[Paquot and Granger, 2012]
En outre, leur utilisation pourrait servir non seulement un objectif utilitariste, mais aussi explicatif, puisque ces données, certes statistiques *in fine*, sont linguistiques et pourraient par conséquent contribuer à une meilleure compréhension, et donc maîtrise, des décisions de l'algorithme.
Toute amélioration de la concordance entre les scores humains et les scores automatisés permettra donc d'une part d'économiser un temps de notation substantiel et les dépenses associées à l'utilisation d'un deuxième évaluateur humain et d'autre part d'espérer échapper à la subjectivité inhérente à la correction humaine. En effet, dès les années 30, Laugier et Weinberg [Laugier and Weinberg, 1927] avaient mis en évidence, en utilisant les épreuves du baccalauréat dans toutes les matières, des écarts importants de notation de copies identiques par des correcteurs différents.
On peut donc espérer que la notation automatique, entraînée sur une quantité suffisante de copies et le corpus entier des correcteurs du Français des affaires, puisse accompagner la correction humaine et soit imperméable aux biais de voisinage (effet de contraste ou de séquence) [Leclercq et al., 2004].
La quête de notation objective étant semée d'embûches, il faudrait aussi porter une attention toute particulière à l'étape de la sélection des copies à utiliser pour l'entraînement de l'algorithme afin de ne pas laisser un quelconque biais s'y glisser.

Première partie

Contextes

CHAPITRE 1

CADRE DE L'ÉTUDE

Sommaire

1.1 Introduction

Nous présenterons d'abord le cadre dans lequel notre stage a eu lieu, au sein du département de la direction scientifique du français de affaires, une activité de la Direction de l'attractivité internationale de la Chambre de Commerce et d'industrie de région Paris Île de France.
Nous ferons une description rapide de ses activités et de leur contexte, ainsi que de ce qui a motivé les besoins qui sont à l'origine du système de notation automatique que nous décrivons et pour lequel nous avons cherché des améliorations en se fondant sur l'utilisation de collocations ([Paquot and Granger, 2012]).
Nous décrirons ensuite les besoins et contraintes spécifiques au projet de notation automatique de tests de niveau en langue française envisagé par la CCI, ainsi que les outils mis à notre disposition pour le réaliser, et nous tenterons d'en dégager une première problématique pragmatique, puis de dégager une analyse orientée recherche d'un point de vue linguistique et informatique dans le chapitre 3, en nous appuyant sur notre description de l'état de l'art de la notation automatique de textes présentée dans le chapitre 2.

1.2 Cadre pratique

1.2.1 Le Français des affaires et les tests de français

Créé en 1958, le Français des affaires est un établissement de la Chambre de commerce et d'industrie de Paris Île-de-France, dont l'activité principale consiste à concevoir, mettre en œuvre et valider des tests de français général et des diplômes de français professionnel et faire profiter de son expertise pédagogique aussi bien les enseignants que les entreprises.

Les Diplômes de français professionnel (DFP) sont spécifiques à un domaine d'activité. Ils permettent au candidat de prouver sa capacité à communiquer profession-

nellement en français à un niveau donné dans un des domaines suivants : affaires, relations internationales, tourisme, hôtellerie, restauration, santé.
Les DFP peuvent aussi être utilisés par un candidat à la naturalisation française pour justifier d'un niveau de français de niveau minimum B1.
Le Test d'évaluation du français (TEF), qui existe depuis 1998, est principalement utilisé par des candidats qui désirent immigrer en France, au Québec ou au Canada, ou à des personnes qui souhaitent obtenir la nationalité canadienne ou française.

En 2019, il existait 571 centres d'examens, parmi eux principalement des universités, des Alliances françaises et instituts français. Ils sont répartis dans 118 pays, les cinq principaux étant la France (116 centres), les États-Unis (40), le Canada (29), l'Italie (28) et la Chine (15).
Cette même année, 54 995 candidats ont passé une version du Test d'évaluation de français du Français des affaires.
Depuis janvier 2021, l'ensemble des épreuves écrites s'effectue sur ordinateurs, sauf demande particulière d'aménagement.

Le TEF permet d'attester de son niveau de français soit dans un cadre professionnel, académique, ou de mobilité. Il existe donc différentes versions selon les commanditaires, à savoir :

TEF Études : version modulaire constituée de 5 épreuves, utilisée notamment pour l'inscription dans des universités francophones ;

TEF Naturalisation : vise à valider un niveau B1 à l'écrit et à l'oral, dans le cadre des démarches d'obtention de la nationalité française ;

TEF Carte de résident : vise à valider un niveau A2 à l'écrit et à l'oral, en vue de l'attribution d'une carte de résident d'une validité de 10 ans en France ;

TEF Canada : utile pour une demande d'immigration vers le Canada ou une demande de citoyenneté ;

TEF Québec (TEFAQ) : utile pour une demande d'immigration au Québec.

TEF Express : a été mis en place en raison des mesures de confinement dues au COVID-19. Il se déroule à distance ou à domicile et ne comprend que 2 épreuves, compréhension écrite et compréhension orale. n l'absence de dispositif de surveillance des passations, il est constitué d'items différents et ne délivre pas d'attestation officielle de niveau et permet simplement de justifier de son niveau de français lors de l'inscription dans une université canadienne.

le TEF Naturalisation et le TEF Canada comportent 4 épreuves qui doivent toutes être passées en une seule session. Le candidat doit répondre à des questions (QCM) de compréhension écrite et de compréhension orale, puis rédiger le test d'expression écrite (sur ordinateur) et enfin participer à une épreuve d'expression orale face à un examinateur.
Les QCM sont évalués directement par l'ordinateur en fonction du nombre de bonnes réponses et de la difficulté des items constituant les épreuves. L'expression orale est évaluée une première fois par l'examinateur de la session (évaluation en contexte) puis une seconde fois par un examinateur extérieur, qui va écouter l'enregistrement de l'échange (évaluation hors contexte), pour obtenir la note la plus représentative du niveau du candidat.
L'expression écrite est également corrigée par deux examinateurs (sans concertation). Si les deux notes concordent, le niveau est validé, si les notes divergent d'un

niveau, le niveau retenu est la moyenne des deux scores, et en cas d'écart d'appréciation plus important, un arbitrage est demandé à un responsable pédagogique du Français des affaires, dans le but de délivrer un résultat correspondant à la performance du candidat.
Le Français des affaires délivrant des certifications officielles de niveau de français, le résultat final se doit de refléter au mieux le niveau réel du candidat.
En outre, le nombre de candidats passant les tests augmentant, les cas d'arbitrage se multiplient, et avec eux le temps d'attente des résultats.
C'est pourquoi, afin de maintenir un délai raisonnable de correction et la fiabilité des résultats, le Français des affaires envisage de mettre en place une notation automatique qui permettrait de remplacer un des deux examinateurs. Un deuxième examinateur n'interviendrait qu'en cas d'écart important entre la notation automatique et celle de l'examinateur.
Il n'est donc pas question de se substituer à l'humain mais de disposer d'une évaluation fondée sur les données précédentes et de limiter le nombre de corrections multiples des copies tout en obtenant un score fiable.
Les correcteurs de l'épreuve d'expression écrite s'appuient sur une grille de correction basée sur des critères spécifiés par l'équipe pédagogique (voir section A.1).

Les niveaux se répartissent comme suit :

Score TEF	0 - 99	100 - 199	200 - 299	300 - 399	400 - 499	500 - 599	600 - 699
Niveau CECR	<A1	A1	A2	B1	B2	C1	C2

TABLE 1.1 – Grille des niveaux CECRL en fonction des scores

Les 5 critères considérés pour l'évaluation sont les suivants :

1. Capacité à transmettre des informations ;
2. Capacité à argumenter ;
3. Syntaxe ;
4. Lexique ;
5. Cohérence et cohésion.

Un score est ainsi attribué à la copie qui est ensuite retranscrit en niveau.
Le CECRL (Cadre européen commun de référence pour les langues) définit des niveaux de maîtrise d'une langue étrangère qui constituent désormais la référence dans le domaine de l'apprentissage et de l'enseignement des langues dans de nombreux pays.
Pour le TEF Carte de Résident, la copie n'est notée que jusqu'au niveau B1 et les candidats n'ont à produire qu'un seul texte, et ce sont avec ces données qu'ont été élaborés les premiers modèles de notation automatique (au cours des stages de Benjamin Larvaron, Cynthia Gilles, Julien Mouchnino et Mouhamadou Lamine Mboup), car c'est le premier test dont l'épreuve d'expression écrite a été proposée au format électronique (dès 2018).
Afin d'étendre la notation automatique à l'ensemble des niveaux du CECRL, nous avons pour notre part travaillé dans un temps sur des copies du TEF pour la naturalisation, échantillon qui a par la suite été complété :

— -pour les niveaux <A1 et A1, par des copies du TEF pour la carte de résident (tâche identique)

— et pour les niveaux C1 et C2 par des copies de la tâche d'argumentation similaire du TEF Canada

et ce afin de disposer d'une représentation plus homogène des copies pour chacun des niveaux.
La tâche qui a été considérée pour cette étude est donc une tâche d'argumentation, commune aux différentes variantes du TEF.

1.2.2 Le rôle du département scientifique

L'équipe du développement scientifique a pour objectif de veiller à la qualité métrique des outils d'évaluation, à l'intégrité des résultats et à l'intégration des nouvelles technologies au service de l'évaluation. Elle travaille en collaboration directe avec les autres équipes (innovation pédagogique, développement, communication, opérations, innovation numérique) afin de veiller à la qualité du matériel de test utilisé, à l'intégrité des résultats délivrés et à leur bonne interprétation.
Les domaines de compétence du département scientifique sont :

1. Maintenir / renforcer la qualité métrique des tests et certifications :
   — par la validation psychométrique des nouveaux items pour générer des versions de tests de difficulté comparable à partir d'une banque d'items calibrés
   — en limitant les biais et les erreurs d'évaluation avec un suivi des évaluateurs et en vérifiant la cohérence entre les scores aux différentes épreuves
   — en contrôlant l'intégrité des passations en détectant les potentiels cas de triche et de potentiels délits d'initiés, c'est-à-dire des items dévoilés, en continu
2. Sécuriser les conditions de passation :
   — en surveillant le taux d'exposition des items (fréquence d'utilisation) ainsi qu'en alertant sur les besoins en nouveaux items
   — en proposant régulièrement de nouvelles versions
   — en identifiant les centres et/ou les régions à risque : cela consiste à surveiller et éventuellement contrôler la fréquence des suspicions de triche, des écarts importants d'évaluation, etc.
3. Optimiser l'emploi des ressources d'évaluation :
   — en exploitant pleinement la banque d'items
   — en étudiant et en mettant en œuvre des modèles de mesure alternatifs
4. Favoriser l'innovation :
   — en faisant de la veille scientifique
   — en expérimentant de nouveaux modèles à travers des stages
   — en testant de nouvelles idées/items auprès de cobayes
   — en exploitant les données (copies) des épreuves : mise en évidence des principales caractéristiques des productions des différents niveaux
5. Communiquer sur la qualité des instruments auprès :
   — des prescripteurs par des rapports d'homologation
   — de la communauté scientifique via des articles ou des participations à des colloques
   — d'un public averti au travers de notes de recherche.

Le cœur de métier de l'équipe est donc le contrôle et l'amélioration continue de la qualité des évaluations notamment par des analyses psychométriques routinières des

bases de données. Il consiste également à apporter un appui notamment à l'équipe pédagogique lorsqu'il s'agit de prendre des décisions concernant un évaluateur, par exemple.
En effet, les épreuves d'expression écrite sont évaluées par deux correcteurs, et si les deux notes concordent le niveau est validé, mais si les notes divergent d'un niveau, le niveau retenu est la moyenne des scores donnés par les évaluateurs. Si les notes diffèrent de deux niveaux, une troisième correction est demandée (arbitrage).
Un outil d'ajustement automatique des évaluations, élaboré à partir des données des profils de correction des évaluateurs a été mis en place (fondé sur les principes exposés dans [Casanova and Demeuse, 2016]) pour compenser les éventuelles différences liées à la personnalité de chaque examinateur (tendance systématique à la sévérité ou à la générosité).
Par conséquent, l'équipe améliore constamment ses outils/modèles grâce à diverses expérimentations et en partage les résultats par des publications ou des participations à des colloques scientifiques (Le Français des affaires est notamment membre de l'ADMEE-Europe, Association pour le développement des méthodologies d'évaluation en éducation, et d'ALTE, Association of Language Testers in Europe).

1.3 Objectifs

Le Français des affaires délivrant des certifications officielles de niveau de français, la note finale se doit d'être le plus juste possible et au plus près du niveau réel du candidat.
Or le nombre de candidats passant les tests augmente, contrairement au nombre de correcteurs, impliquant par conséquent un allongement du délai de correction. Pour maintenir un délai raisonnable et conserver la fiabilité des résultats, le Français des affaires souhaite mettre en place une correction automatique.
Cette correction permettrait de remplacer un des deux examinateurs et de détecter les copies pouvant être discutées. On aurait alors une correction d'un deuxième examinateur seulement si la note entre la correction automatique et la note de l'examinateur ont deux niveaux d'écart.
L'objectif de ce travail est donc à la fois d'améliorer les travaux précédents qui ont abouti à la construction d'un modèle permettant la correction automatique de tests d'expression écrite en français langue étrangère, d'étendre leur scope en terme de niveaux, et peut-être d'optimiser les caractéristiques choisies tout en utilisant des paramètres plus interprétables.

CHAPITRE 2

ÉTAT DE L'ART

Sommaire

2.1 Introduction

La construction d'un système de notation automatique de productions écrites en langue seconde est à la croisée des domaines de l'étude de l'apprentissage des langues, de la docimologie, du traitement automatique des langues et de l'apprentissage automatique.
En effet, avant de l'évaluer, il convient de définir ce que nous entendons par compétence en langue : elle se décline en réalité en plusieurs domaines, dont deux principaux qui d'après Bachman et Palmer (voir : A.2) sont la connaissance organisationnelle (principalement celle des structures formelles de la langue) et les connaissances pragmatiques (principalement la production et la compréhension de la communication). La littérature (voir ci-dessous) semble converger pour nous indiquer que les collocations sont un marqueur fort du niveau de compétence en langue (appartenant donc à la partie organisationnelle), permettant de distinguer les apprenants avancés. Comme le modèle existant élaboré au sein du français des affaires peine à reconnaître les niveaux avancés, il nous a semblé judicieux d'extraire des collocations des caractéristiques qui améliorent son efficacité pour ceux-ci.

2.2 À la croisée des chemins

La compétence lexicale (la capacité de produire et de comprendre les mots d'une langue) est une mesure essentielle du degré de maîtrise d'une langue. Pour les apprenants d'une langue seconde (L2), elle constitue un indicateur important de réussite lors de leur parcours universitaire (Daller et al. 2003). Et pourtant, elle a été décrite comme mal comprise [Crossley et al., 2011] et complexe [Zareva et al., 2005]. Nombreux sont les chercheurs qui se sont efforcés de développer des méthodes informatiques valides pour la mesurer (par exemple, [Daller et al., 2003], [McCarthy and Jarvis, 2010]), mais elle se mesure essentiellement en fréquences de mots isolés.
Beaucoup de modèles s'appuient sur les facteurs qui représentent le continuum des

connaissances lexicales d'un apprenant, depuis leur étendue jusqu'à la capacité à les organiser([Henriksen, 1999]).
Par conséquent, cet autre élément de la maîtrise d'une langue, qu'est la compétence phraséologique est désormais aussi largement reconnue comme un élément important d'une bonne maîtrise de la langue [Paquot, 2019], mais elle n'est mentionnée dans le Cadre européen commun de référence pour les langues (CECRL, Conseil de l'Europe, 2001) que pour les niveaux les plus élevés (C2) sous la forme de la maîtrise des expressions idiomatiques.
En effet, pendant longtemps, peu nombreux étaient les travaux de recherche dans le domaine du FLE (français langue étrangère) consacrés spécifiquement à la didactique des collocations, par rapport à la didactique du lexique en général.
La raison en est que leur classification et leur définition sont sujettes à débat et varient suivant les auteurs (par exemple, [Hausmann and Blumenthal, 2006], [Nesselhauf, 2005], [Gries, 2013]), et leur enseignement est souvent réduit à celui des expressions idiomatiques (par exemple « avoir une peur bleue ») ou des proverbes.
En réalité, les collocations apparaissent beaucoup plus que les expressions idiomatiques et les proverbes dans nos discours. Elles ont pour caractéristiques que leurs composants gardent souvent leur sens littéral et sont moins figés. On peut les voir comme une expression semi-idiomatique (« avoir un gros chagrin », mais « éprouver une grande douleur »), où intervient le choix du locuteur [Kahane and Polguere,].
L'insertion et la substitution sont possibles, ce qui les rend difficiles à repérer comme une unité, et elles sont source d'erreurs pour les apprenants lors de la production, orale ou écrite.
Ensuite, au cours des dix dernières années, les collocations ont fait l'objet de nombreuses études sur l'apprentissage des langues et on a constaté que les combinaisons de mots qui reviennent fréquemment dans un registre, jouent un rôle crucial dans l'acquisition du langage ([Ellis, 2012], [Vlach, 2019] etc).
Les paires de mots qui « cooccurrent plus souvent que ne le laissent prévoir leurs fréquences respectives et la longueur du texte dans lequel elles apparaissent » (Altenberg, 1998, p. 122 cité par [Paquot, 2019]), méritent une attention particulière dans l'enseignement ([O'Donnell et al., 2013], [Gablasova et al., 2017]) et restent un défi pour les niveaux les plus avancés (par exemple [Laufer and Nation, 1995], [Ebeling and Hasselgård, 2015]).
En 1993, Waller avait démontré grâce à une analyse statistique que les erreurs d'utilisation de collocation se limitaient aux textes écrits par des locuteurs non natifs, alors que par exemple, les erreurs syntaxiques, ou des erreurs lexicales autres que collocationnelles) se retrouvaient aussi dans le groupes des natifs.
Ensuite, en 1996, Howarth avait observé que des étudiants, par ailleurs très avancés, commettent un type particulier d'erreurs, à savoir des erreurs de collocation, qui, selon lui, « can lead to a lack of precision and obscure the clarity required in academic communication » (cité par [Lesniewska, 2006]). Par conséquent, les erreurs des collocations peuvent être vues comme un marqueur du texte non natif : « a foreign accent in writing » (cité par [Lesniewska, 2006].
Il n'existe pas actuellement de liste recensant des combinaisons de mots et de leurs fréquences en français. Il n'est donc pas possible de procéder comme pour la complexité lexicale.
On utilise donc souvent les Pointwise Mutual Information (PMI), mesure issue de la théorie de l'information et qui reflètent la « force » des collocations, c'est-à-dire leur sophistication.

Cette mesure (PMI) proposée par [Church and Hanks, 1989]) est issue de la théorie de l'information et quantifie la probabilité d'apparition d'un mot a avec un mot b, en fonction de leur fréquence.
Il a été démontré ([Durrant and Schmitt, 2009]) que les locuteurs non natifs de l'anglais utilisent moins que les natifs les éléments aux scores PMI élevés (deux éléments peu courants fortement associés) et que cette différence se retrouvait chez les apprenants intermédiaires et avancés ([Paquot and Granger, 2012]).
Il semble donc assez convaincant d'utiliser les collocations et cette mesure pour évaluer la compétence en langue et notamment distinguer les performances de candidats avancés.
Il existe différentes approches pour classer les collocations. L'approche basée sur la fréquence de cooccurrence des mots dans les corpus ([Paquot and Granger, 2012]) est celle que nous avons adoptée. On peut y distinguer les cooccurrences de surface et syntaxiques [Evert, 2005]. La cooccurrence de surface s'intéresse à la cooccurrence des mots, les cooccurrences syntaxiques nécessitent des informations supplémentaires sur la structure du texte (par exemple, verbe + objet, adj + modif + nom) qui est une approche phraséologique.
Lors de l'identification des collocations, nous pouvons également tenir compte de la distance entre les mots co-occurrents (fenêtres) et de la longueur souhaitée des unités (n-grammes).
Une autre approche est celle des réseaux de collocation [Brezina et al., 2015], qui utilisent l'approche par fenêtre pour relier un mot à un second, puis un troisième...
Ce travail utilisera l'approche par fenêtre (les bigrammes en sont un cas particulier) et prendra en compte les formes de surface, lemmes, étiquetages morphosyntaxiques et dépendances, ce qui permet d'expérimenter avec diverses structures linguistiques.
Deux mesures statistiques des collocations seront utilisées : la fréquence (comptage des cooccurrences des formes de mots) et la force d'association entre les combinaisons de mots (voir [Evert, 2005] ; [Gablasova et al., 2017]).
Il est donc clair que ce travail se positionne dans une approche opportuniste ([Valette and Eensoo, 2014]) ou hybride, où les méthodes de classification utilisent à la fois des règles et la statistique, tout ceci n'étant possible bien entendu, que grâce à la naissance de corpus annotés suffisamment volumineux.
Depuis plusieurs dizaines d'années, des systèmes de notation automatique (voir [Ramesh and Sanampudi, 2021] pour une revue complète) comme par exemple, Intelligent Essay Grader ([Foltz et al., 1999]), E-Rater ([Attali and Burstein, 2006]) sont développés et utilisés en complément des correcteurs humains pour la notation de copies en anglais.
Les modèles de systèmes de notation automatique sont généralement développés en utilisant les jugements des évaluateurs humains comme données d'entraînement, ce qui pose la question de l'annotation.
En effet, les notes peuvent varier d'un correcteur à un autre et même dans le temps pour un même correcteur ([Leclercq et al., 2004]).
Ces modèles de notation automatique ont recours à un large éventail de caractéristiques [Foltz et al., 1999], de la longueur des phrases à l'exactitude grammaticale et orthographique, la longueur du texte produit, la longueur des mots, le nombre d'étiquettes morphosyntaxiques différentes, la qualité de la langue, la diversité lexicale, la structure syntaxique et les erreurs [Dikli, 2006].
Ce n'est que plus récemment que des modèles ont été développés pour les langues européennes ([Vajjala and Lèo, 2014] par exemple). Cependant, contrairement à

l'anglais, à notre connaissance, les systèmes de notation automatique n'ont pas été utilisés à grande échelle et dans divers domaines, pour ces langues.
Les notes générées par e-rater sont prédites par un modèle de régression linéaire multiple [Attali and Burstein, 2006].
Les méthodes d'apprentissage automatique ont été appliquées à la notation automatique (souvent modélisée comme un problème de régression, de classement ou de classification ([Taghipour and Ng, 2016]) de divers types de tâches. Par exemple, en 2012 Nehm, Ha et Mayfield [Nehm et al., 2012] ont étudié l'utilisation de l'apprentissage automatique pour noter automatiquement des évaluations en biologie. Leur programme s'est révélé efficace pour évaluer les connaissances et les performances des élèves dans ce domaine scientifique.
Un système de notation automatique a aussi été développé en 2012 pour noter de courtes réponses à une question (écrites) [Heilman and Madnani, 2013] et en 2011, Yannakoudakis, Briscoe et Medlock ([Yannakoudakis et al., 2011]) ont appliqué des techniques d'apprentissage automatique pour noter des copies d'examen d'anglais langue seconde.
Des études ont été menées pour comparer les performances de différentes méthodes d'apprentissage automatique dans la notation automatisée. Actuellement, la plupart des travaux sur la création de caractéristiques pour la notation automatique sont basés sur des approches orientées machine learning ([Fonseca et al., 2018]). Elles ont principalement recours à une combinaison de variables statistiques issues du traitement du langage naturel pour la création de caractéristiques (Feature Engeeniring).
Classiquement, les modèles utilisés sont les Support Vector Machine, avec différents noyaux ([Shin, 2018]), les réseaux de neurones ([Taghipour and Ng, 2016]), et les arbres de boosting de gradient (GBT) ([Fonseca et al., 2018]).
La méthode Least Absolute Shrinkage and Selection Operator (LASSO), proposée par Tibshirani en 1996, permet d'améliorer la précision et l'interprétabilité de la régression linéaire en éliminant les caractéristiques les moins pertinentes, c'est-à-dire celles qui ont le moins d'impact sur les résultats de la régression. La machine de boosting de gradient proposée par Friedman ([Friedman, 2001] en 2001 peut être utilisée avec les arbres de régression. Selon Friedman, la précision relativement élevée, la performance constante et la robustesse du boosting peuvent représenter un avantage notable. Ces résultats suggèrent que la notation basée sur le SVM produit un accord machine-humain qui s'approche de l'accord humain-humain dans certains cas.
Nous testerons donc nos caractéristiques avec ces algorithmes.
Jusqu'à récemment, l'apprentissage automatique (ML) utilisant des méthodes basées sur l'ingénierie des caractéristiques ont prévalu dans le domaine de la notation automatique ([Ramesh and Sanampudi, 2021], [Attali and Burstein, 2006], [Dikli, 2006], [Foltz et al., 1999]).
Certaines études ont suggéré que le Deep Learning semblait avoir de meilleurs résultats ([Filho et al., 2020], [Fonseca et al., 2018]). Le Deep Learning permet de se passer de l'étape de création et extraction de caractéristiques pertinentes, qui est longue et laborieuse ([Taghipour and Ng, 2016]), mais c'est au détriment de l'interprétabilité.
C'est pourquoi les chercheurs ont tenté de s'assurer que leur modèle, et donc ses caractéristiques (les indicateurs qu'ils créent) capturent au mieux contenu de la production écrite, et c'est dans une optique similaire que ce travail se positionne.

2.3 Conclusion

Tous les points mentionnés dans les travaux cités ne seront pas forcément utilisables dans notre contexte : le style d'écriture de ce corpus de travail est très diversifié et parfois déconcertant, même normalisé. Et même si cela ne concerne que les copies de niveau les plus faibles, il est probable que l'utilisation des données syntaxiques des collocations ne soit pas pleinement exploitable ici.
Nous nous inspirerons néanmoins de ces analyses pour développer une première méthodologie adaptée aux spécificités du corpus de travail.
Par ailleurs, il est à noter que le corpus est en français et que les caractéristiques des apprenants par les collocations puissent ne pas avoir les mêmes propriétés qu'une anglais [Vandeweerd,]).

Deuxième partie

Expérimentations

CHAPITRE 3

CORPUS ET CARACTÉRISTIQUES EXISTANTES

Sommaire

3.1 Introduction

Ce travail a principalement deux objectifs : l'ajout de développement de caractéristiques issues des collocations aux modèles prédisant le niveau d'une copie, et en conséquence, une réflexion sur le choix des caractéristiques à la fois les plus efficaces et les plus interprétables.
Par conséquent, nous réaliserons une exploration des données du corpus constitué des textes des copies seuls, puis en regard avec le corpus qui a servi à créer les collocations de référence, afin de mieux le cerner et de construire les caractéristiques pertinentes reflétant le statut collocatif des copies pour nourrir la classification automatique.
Il s'agit là de comparer et explorer les critères récemment ajoutés (marqueurs de l'argumentation) qui sont « parlants » et qui serviront de caractéristiques pour les méthodes d'apprentissage supervisé employées.
Nous supposons que le choix de telles caractéristiques au regard de notre objectif

est plus robuste que, par exemple, des critères en lien avec la thématique de chaque test (débattre sur les écrans dans la vie quotidienne par exemple) critères propres au corpus et qui risqueraient d'engendrer un surapprentissage (i.e. construction d'un modèle peu généralisable) et qu'il faut donc plutôt s'orienter vers des critères fondés par exemple, sur la présence de structures argumentatives, ce qui est déjà le cas pour certaines caractéristiques existantes, et qui est en lien avec tout ce qui est collocation.

Nos données sont composées d'un ensemble de copies avec leur niveau, leur texte, le test concerné (TEF, TEF naturalisation, TEF canada), et le sujet (utilisé pour les critères issus du champ lexical, sujet du stage de Chia-Ting Kuo) leur texte a été normalisé et nous disposons déjà de caractéristiques extraites (que nous allons détailler plus loin dans la3.3.

3.2 Corpus

Les TEF concernés sont désormais presque intégralement passés au format électronique, le corpus de copies est donc en pleine expansion.
Comme l'objet de ce travail est d'améliorer un modèle existant, nous allons nous appuyer sur un échantillon de 2928 copies qui a une distribution des niveaux conforme à celle de l'ensemble des copies, accompagnées de caractéristiques qui ont été créées pour le modèle précédent et auxquelles nous ajouterons celles que nous élaborerons.
Pour constituer le corpus de collocations de référence qui serviront de référence. nous avons utilisé une partie du corpus Leipzig qui est composé d'articles de presse (nous avons annoté 150 000 phrases) et est mis à disposition du public par l'université de Leipzig.

3.2.1 Pré-traitements effectués sur les copies

Les copies ont été anonymisées, car il arrive que les candidats y incluent leur nom et/ou leur(s) prénom(s) ou courriels. Afin d'éviter de supprimer un mot existant, il a été décidé de remplacer les mots qui sont à une distance de Damerau-Levenshtein de 0 ou de 1 du prénom ou du nom et qui ne sont pas dans un dictionnaire comportant les 10 000 mots les plus fréquents de la langue française.
Les erreurs typographiques courantes et spécifiques au corpus (utilisation de virgules ou apostrophes pour représenter les accents ou texte écrit en majuscules, par exemple) sont corrigées à l'aide de règles.
De la même façon, les erreurs orthographiques sont corrigées à l'aide de règles s'appuyant sur un fichier contenant un ensemble de fautes fréquentes de la langue française et de leurs corrections et d'un fichier constitué des orthographes de type SMS (comme « C » pour « C'est », « Keskil » pour « Qu'est-ce qu'il », etc.).
Ensuite elles sont traitées par un algorithme de normalisation automatique (élaboré par Julien Mouchnino lors de son stage en 2018), entraîné sur les erreurs trouvées dans les réponses à l'épreuve d'expression écrite du TEF et leur correction (corpus créé par M. Casanova).
Ce corpus est en expansion, puisqu'il existe une interface qui permet d'afficher un mot erroné quand celui-ci n'est pas dans l'historique des corrections, et permet éventuellement à un superviseur humain (compte tenu du nombre de copies traitées par jour, l'enrichissement du corpus d'erreurs par interface n'est pas raisonnablement envisageable en continu, mais il est possible ponctuellement afin de nourrir ce

corpus) d'entrer la correction qui lui semble juste.
L'algorithme (basé sur du boosting) y propose des variantes. Il utilise comme caractéristiques basées sur un calcul des erreurs de la distance de Damerau-Levenshtein entre l'erreur et la suggestion faite grâce à une méthode de phonétisation, mais aussi de variables dépendant de la fréquence de la suggestion dans un corpus donné ou même de la fréquence d'occurrence parmi la liste des 100 mots avec le plus de variantes différentes parmi les erreurs.

Aperçu de l'interface

FIGURE 3.1 – Interface d'ajout de corrections de mots erronés au corpus utilisé pour normaliser les copies

3.2.2 Annotation du corpus Leipzig

Le corpus a été annoté à l'aide de l'analyseur syntaxique Udpipe (Straka et Straková, 2017) de R avec le modèle french-gsd (c'est à dire entraîné sur le corpus UD GSD).

	doc_id	paragraph_id	sentence_id	sentence	token_id	token	lemma	upos	xpos	feats	head_token_id	dep_rel	deps	misc	has_morph	mor
47	doc1	1	2	« 06 septembre 2020, un jour de joie pour Kpatcha G...	24	imaginaire	imaginaire	ADJ	NA	Gender=Masc|Number=Sing	23	amod	NA	SpaceAfter=No	TRUE	NA
48	doc1	1	2	« 06 septembre 2020, un jour de joie pour Kpatcha G...	25	,	,	PUNCT	NA	NA	2	punct	NA	SpacesAfter=\n	FALSE	NA
49	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	1	"Tous	"Tous	PRON	NA	Gender=Masc|Number=Plur	4	nsubj:pass	NA	NA	TRUE	NA
50	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	2	ont	avoir	AUX	NA	Mood=Ind|Number=Plur|Perso...	4	aux	NA	NA	TRUE	NA
51	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	3	été	être	AUX	NA	Gender=Masc|Number=Sing|T...	4	aux:pass	NA	NA	TRUE	NA
52	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	4	transférés	transférer	VERB	NA	Gender=Masc|Number=Plur|Te...	0	root	NA	NA	TRUE	NA
53	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	5–6	au	NA	NA	NA	NA	NA	NA	NA	NA	FALSE	NA
54	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	5	à	à	ADP	NA	NA	7	case	NA	NA	FALSE	NA
55	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	6	le	le	DET	NA	Definite=Def|Gender=Masc|Nu...	7	det	NA	NA	TRUE	Def
56	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	7	pénitencier	pénitencier	NOUN	NA	Gender=Masc|Number=Sing	4	obl:arg	NA	NA	TRUE	NA
57	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	8	de	de	ADP	NA	NA	9	case	NA	NA	FALSE	NA
58	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	9	Koléa	Koléa	PROPN	NA	NA	7	nmod	NA	SpaceAfter=No	FALSE	NA
59	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	10	,	,	PUNCT	NA	NA	4	punct	NA	NA	FALSE	NA
60	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	11	a	avoir	AUX	NA	Mood=Ind|Number=Sing|Perso...	12	aux	NA	NA	TRUE	NA
61	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	12	révélé	révéler	VERB	NA	Gender=Masc|Number=Sing|T...	4	conj	NA	NA	TRUE	NA
62	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	13	le	le	DET	NA	Definite=Def|Gender=Masc|Nu...	14	det	NA	NA	TRUE	Def
63	doc1	1	3	"Tous ont été transférés au pénitencier de Koléa, a ré...	14	quotidien	quotidien	NOUN	NA	Gender=Masc|Number=Sing	12	obj	NA	NA	TRUE	NA

FIGURE 3.2 – Aperçu des copies annotées

3.3 Caractéristiques existantes

La description des caractéristiques existantes est nécessaire pour interpréter l'apport des nouvelles caractéristiques issues des collocations.
Les travaux précédents d'élaboration d'un système de notation automatique au sein du Français des affaires ont conduit à la création des caractéristiques suivantes (leur regroupement a une part d'arbitraire qui découle de la distinction entre forme

et fond) :

3.3.1 Les caractéristiques lexicales

Les caractéristiques lexicales traduisent la richesse lexicale de la copie du candidat.
Tout d'abord nous avons la taille du vocabulaire du candidat : **nb_vocab**, elle est préférée au nombre tokens qui est moins informatif du fait de la répétition possible de mots. Afin de mesurer plus finement la complexité lexicale de la copie, les travaux précédents ont utilisé les corpus FLELex.
FLELex est un lexique de français langue étrangère (FLE) mis en place par Dr Thomas François qui donne les fréquences d'apparition des lemmes dans les méthodes d'apprentissage du français langue étrangère aux différents niveaux du CECR (c'est à dire A1, A2, B1, B2, C1, C2). Il a été obtenu à partir d'un corpus de 777 000 mots provenant de manuels scolaires et de textes simplifiés destinés aux apprenants de FLE.

Lemme	POS	A1	A2	B1	B2	C1	C2	Total
Voiture	Nom	633.3	598.5	428.7	202.7	271.9	25.9	461.5
Abandonner	Ver	35.5	62.3	104.8	79.8	73.6	28.5	78.2
sous réserve de	PREP	0	62.3	0	0	0	0	0,03
kilo	NOM	40,3	29,9	10,2	0	1,6	0	19,8

TABLE 3.1 – Extrait de FLElex : les lemmes et leurs fréquences respectives par niveaux

Ces fréquences ont servi à créer les caractéristiques suivantes :

ind1_flelex : Au départ ind_flelex, variable correspondant à la moyenne de la fréquence totale d'apparition des lemmes trouvés dans la table FLELex ainsi que dans la copie du candidat, divisée par la fréquence maximale de FLELex.
En d'autres termes, cette variable donne une idée de la richesse des mots utilisés par le candidat. Elle a été ensuite remplacée par ind1_flelex = (1 - ind_flelex) afin de faciliter son interprétation.

ind_A1, ind_A2, ..., ind_C2 : Ces variables correspondent aux parts des lemmes de la copie figurant dans la table FLELex avec une probabilité cumulée (en partant du niveau A1) pour les niveaux A1, A2, ..., C1, C2, supérieure à 0,5.

ind_alt_A1, ind_alt_A2, ..., ind_alt_C2 : Elles correspondent aux parts des lemmes de la copie figurant dans la table FleLex avec une fréquence d'apparition au niveau A1, A2, ..., C1, C2, supérieure à la fréquence totale.

La table grande_base_mots (qui contient des fréquences d'utilisation de mots dans la langue française) a servi pour créer les caractéristiques **Base1**, **Base2**,... **Base6**. Chacune d'entre elles correspond à la fréquence du nombre de mots trouvés dans grande_base_mots par rapport aux tranches ci-dessous :

— Tranche 1 : les 50 mots les plus courants pour base1 ;
— Tranche 2 : du 50e au 100e mot le plus courant pour base2 ;
— Tranche 3 : du 100e au 500e mot le plus courant position pour base3 ;

— Tranche 4 : du 500e au 1000e mot le plus courant pour base4 ;
— Tranche 5 : du 1000e au 3000e mot le plus courant pour base5 ;
— Tranche 6 : au delà du 3000e mot le plus courant position pour base6.

Ainsi, l'indice base1 correspond au nombre total de mots de la copie appartenant à la Tranche 1 divisé par le nombre total de mots de la copie. Ainsi de suite pour les autres variables base1, base2, ..., base6.

Autres caractéristiques que l'on peut considérer comme lexicales :
— **vocab_norm** donne la profondeur du vocabulaire de la copie.
— **connecteurs_categ** donne le nombre de catégories de connecteurs mobilisées.
— **modif_diff** : nombre de types de modificateurs différents utilisés.
— **connecteurs_diff** : nombre de connecteurs logiques différents.
— **conj_coord_diff** : indicateur de la diversité des conjonctions de coordination employées.

3.3.2 Caractéristiques reflétant l'adéquation du texte à la tâche

Il s'agit de mesurer si la réponse du candidat correspond au sujet de l'épreuve et s'il argumente.

nb_champ_lex : correspond au nombre de mots de chaque copie appartenant au dictionnaire du champ lexical du sujet (créé en lemmatisant l'énoncé). Cette variable est utilisé en calculant au moyen d'un word2vec la similarité entre les mots du textes et une liste de mots, établie *ad hoc* à partir de l'énoncé du sujet, que les candidats sont susceptibles d'utiliser. Le nombre de mots dont la similarité maximale est supérieure à 0.8 sont comptabilisés.

nb_opinion : correspond au nombre de lemmes de la copie appartenant au dictionnaire d'expressions d'opinion (créé manuellement).

diff_opinion : correspond au nombre de marqueurs d'opinion différents. En d'autres termes, ces deux dernières caractéristiques quantifient la capacité du candidat à donner son opinion et la profondeur de son vocabulaire pour ces marqueurs.

Expression : caractéristique résumant les marqueurs d'expression.

3.3.3 Caractéristiques issues des n-grammes

Un ensemble de copies a été annoté avec UDpipe et les fréquences des bigrammes (et trigrammes pour les étiquettes morphosyntaxiques) y ont été calculées pour chaque niveau, ce qui permet de calculer les probabillités conditionnelles d'apparition d'un lemme en fonction de ce qui le précède.
Ces probabilités (par exemple, pour les lemmes : PAi_lemme) servent à créer les caractéristiques : **upos_infA1,upos_A1,upos_A2,upos_B1,upos_B2,upos_C1, upos_C2,phrase_infA1,phrase_A1,phrase_A2,phrase_B1, phrase_B2,phrase_C1,phrase_C2,dep_infA1,dep_A1,dep_A2, dep_B1,dep_B2,dep_C1,dep_C2,lemme_infA1,lemme_A1, lemme_A2,lemme_B1,lemme_B2,lemme_C1,lemme_C2**.
On y ajoute quatre caractéristiques qui les agrègent, avec une pondération par niveau :

score_lemme : $100(PA1\_lemme + 2PA2\_lemme + 3PB1\_lemme + 3.5PB2\_lemme)$
score_upos, score_dep, score_phrase : même principe de calcul

3.3.4 Caractéristiques reflétant la cohérence du texte

La cohérence du texte est l'un des critères de notation utilisés dans la notation des essais ([Attali and Burstein, 2006]) et est liée à l'aspect fluidité utilisé dans la recherche sur les langues secondes. Elle reflète donc un niveau de compétence linguistique.
Les connecteurs de discours (mots tels que : et, bien que, mais, cependant, etc.) qui relient les parties d'un texte sont fréquemment utilisés comme mesure de la cohérence du texte dans la littérature et des recherches ont également souligné l'importance d'étudier l'utilisation des connecteurs dans les textes d'apprenants ([Granger and Tyson, 1996], [Vajjala, 2017]).
Les caractéristiques s'inspirent notamment de Coh-Metrix ([gra, 2014]) et utilisent l'indicateur qui indique s'il y a au moins une phrase qui reprend un stem présent dans les autres phrases du texte (compte-tenu de la spécificité de notre corpus).

coh_connecteurs : connecteurs logiques et conjonctions de coordination ;

noms_com : part de phrases du texte qui ont un nom (type) en commun avec la précédente ;

lem_com : part de phrases du texte qui ont un lemme (nom ou pronom personnel) en commun avec la précédente ;

stem_com : part de phrases du texte qui ont un stem (nom ou pronom personnel) en commun avec la précédente ;

coh_redondance : cohérence calculée sur la base de la quantité de mots communs du texte (coh_redondance = (stem_com + stem_com_reste + noms_com_reste)/3) ;

diff_anaphores : il s'agit du nombre d'anaphores différents : « lui », pronom « le », « ceci », « celui-ci », «le », « leur ».

3.3.5 Caractéristiques issues de la normalisation

Pénalités

Ces variables nous renseignent sur les différentes erreurs de syntaxe telles que : les erreurs orthographiques, grammaticales, ou absence de ponctuations (virgules).

ratio_elision : taux de fautes d'élision.

ratio_norm : taux de mots normalisés.

Note :
Lors de la phase de normalisation évoquée précédemment, une gravité est assignée aux mots normalisés, et qui pondère le calcul des erreurs de la copie.
Exemple :
— janty > gentil
— gentile > gentil

3.3.6 Caractéristiques syntaxiques et grammaticales

Les fréquences d'erreurs d'accord en genre (en nombre) correspondent au nombre d'erreur d'écart en genre (en nombre) divisé par le nombre total d'écart en genre (en nombre).

nb_punct_diff : nombre de ponctuations différentes utilisées.

ind_virgule : indicateur de l'utilisation de virgules.

nb_temps_verb : indicateur de la diversité des temps employés.

respect_genre_nombre : reflète le respect des accords genre nombre.

Les informations sur tous ces indicateurs sont rassemblées de façon synthétique dans le tableau A.1.

3.3.7 Accord inter-évaluateur

L'échantillon de copies que nous avons utilisé est issu de données de 2020, mais il faut souligner que l'accord inter-évaluateur est relativement stable dans le temps.
L'échantillon est essentiellement constitué de copies du TEF pour la naturalisation (donc des niveaux intermédiaires), mais complété, pour les niveaux inférieurs à A1 et A1 par des copies du TEF pour la carte de résident et, pour les niveaux C1 et C2 par des copies du TEF CANADA.
Il convient de noter qu'il y a un léger biais dans les données exploitées lié au fait que la correction automatique ne s'appuie actuellement que sur la tâche d'argumentation alors que les évaluations de référence du TEF CANADA concernent les deux tâches (narration + argumentation).
Toutefois les copies TEF CANADA utilisée ayant évaluées au niveau C1 ou C2, il y a peut de risque que le texte argumentatif soit sensiblement plus faible ou meilleur que le texte narratif.
Ces copies sont fort peu représentées dans notre échantillon.

Définitions :

- Accord exact = pourcentage de cas où les deux évaluateurs ont donné le même niveau
- Accord adjacent = pourcentage de cas où les deux évaluateurs ont donné le même niveau ou dont l'écart en nombre de niveaux est de 1
- Cas d'arbitrage = pourcentage de cas ou la différence de niveaux entre évaluateurs est élevée.
- Fréquence des accords par niveau = pour chaque niveau, nombre de cas où les deux évaluateurs ont donné le niveau considéré / nombre de cas ou au moins un des évaluateurs a délivré le niveau considéré.

Les statistiques sur les différences d'évaluation (avant et après compensation du profil de sévérité de l'évaluateur) que fournit le Français des affaires sont les suivantes :
M. Casanova, mon encadrant, a calculé, pour chaque copie et chaque évaluateur, l'écart-type entre les notes délivrées selon qu'il y avait accord ou non dans le niveau délivré. Il s'agit de voir si les copies pour lesquelles il y avait accord avaient une notation plus homogène.

Pour le TEF/TEFAQ/TEF CANADA, la moyenne des écart-types est inférieure en cas d'accord avant ajustement (compensation de la sévérité suivant le profil de l'évaluateur), mais les copies ont été sélectionnées sur la base de l'accord après ajustement, où l'écart est moins évident (du moins quand on regarde accord versus écart > 1 niveau).

Avant compensation			**Après compensation (données dont a été extrait l'échantillon pour la notation automatique) :**	
	Ecart-type scores critères		Ecart-type scores critères	
	Moyenne	Variation	Moyenne	Variation
Accord	0,584	0,416	0,591	0,405
Ecart 1 niveau	0,61	0,406	0,608	0,415
Ecart >1 niveau	0,612	0,416	0,589	0,442

TABLE 3.2 – Écart-type entre les notes délivrées - TEF/TEFAQ/TEF CANADA.

Pour le TEF Naturalisation, dont sont principalement extraites les copies de l'échantillon. Le constat est similaire : une homogénéité un peu plus grande lorsqu'il y a accord avant compensation, mais qui est amoindrie par la compensation :

Avant ajustement			**Après ajustement (données dont a été extrait l'échantillon pour la notation automatique) :**	
	Ecart-type scores critères		Ecart-type scores critères	
	Moyenne	variation	Moyenne	variation
Accord	0,462	0,355	0,472	0,351
Ecart 1 niveau	0,49	0,349	0,485	0,348
Ecart >1 niveau	0,485	0,368	0,458	0,414

TABLE 3.3 – Écart-type entre les notes délivrées - TEF Naturalisation.

Pour estimer grossièrement les erreurs de classement des copies, d'une part on utilise uniquement des classements qui coïncident après compensation et on compare le niveau après ajustement et le niveau avant ajustement. L'erreur est de 4 %, avec un écart maximal de 1 niveau.

Il en va de même pour les classements qui convergent avant compensation quand on compare le niveau avant ajustement avec le niveau après ajustement.

On calcule ensuite « l'exactitude de la décision « (entendre exactitude des classements) qui est une mesure psychométrique ([Livingston and Lewis, 1995]. Elle donne :

— entre 82 et 86 % de bons classements à partir des données avant ajustement
— entre 81 et 86 % de bons classements à partir des données après ajustement

Selon chacune de ces deux méthodes, les erreurs de classement ne sont jamais supérieures à 1 niveau.

Le corpus d'entrainement a été constitué de sorte à ne comporter que des copies auxquelles les évaluateurs avaient attribué le même niveau. Cela permet d'avoir une confiance relativement élevée envers le niveau délivré.

3.4 Construction des caractéristiques issues des collocations

Dans ce travail, les collocations choisies, qui doivent refléter la complexité phraséologique, seront mesurées pour les combinaisons de mots contenant des modificateurs adjectivaux, des modificateurs adverbiaux, des verbes (avec auxiliaire ou non),

des noms, des adjectifs et des conjonctions de subordination. La sophistication est étudiée sous l'angle du rapport entre le nombre de collocations de référence et le total des mots, ainsi que le rapport entre le nombre de cooccurrences du texte correspondant à des collocations de référence, ceci pour chaque quartile des valeurs des PMI (Pointwise Mutual Information) associées à chaque combinaison de mots.
Cette mesure est empruntée à la théorie de l'information et compare la probabilité d'observer un mot a et un mot b ensemble, avec les probabilités d'observer a et b indépendamment (mesure proposée par [Church and Hanks, 1989]).

Dans notre analyse, la richesse en collocations d'un texte de TEF a été évaluée à l'aide de l'information ponctuelle mutuelle (PMI) en ne prenant en compte que les noms, les adjectifs, les modificateurs adverbiaux et adjectivaux, les verbes et les auxiliaires, ainsi que les conjonctions de coordination.
Comme il a été montré que cette mesure fait ressortir les collocations composées de mots étroitement associés de fréquence moyenne à faible (c'est-à-dire des mots connus des apprenants avancés) et que des études ([Paquot, 2018]) notent que les scores PMI moyens des collocations utilisées par des apprenants de l'anglais augmentent en continu de B2 à C1 et C1 à C2, cela laisse penser que plus les apprenants sont compétents, plus ils utilisent des collocations aux scores PMI élevés.
D'ailleurs, la littérature recommande couramment d'adopter MI = 3 comme valeur minimale pour le statut collocatif (utilisé dans de nombreux travaux, [Paquot, 2018], [Durrant and Schmitt, 2009], [Granger and Bestgen, 2014]). C'est pourquoi nous avons choisi de tester la construction de nos collocations avec cette valeur.
Nous testerons également des modèles construits avec une PMI > 1, décrite comme moins discriminante, notamment par [Paquot, 2018] (nous avons des copies de tout niveau), mais qui pourraient aider à classer les niveaux plus faibles.
Les scores PMI de chaque collocation ont été calculés sur la base d'un extrait de 150 000 phrases issues du corpus Leipzig, et qui seront nos collocations de référence (notre objectif étant de faire ressortir les niveaux les plus avancés, le choix d'un corpus de presse nous a paru pertinent).
En pratique, le corpus Leipzig a été étiqueté par UDpipe puis les scores PMI ont été calculés pour chaque paire de mots y apparaissant avec une fréquence de 5, 10, 20 et 100 (pour différents tests de création de caractéristiques). Les scores PMI ont ensuite été découpés en quartiles (grade1, grade2, etc.).
La formule utilisée par UDpipe pour calculer les scores PMI est la suivante :

$$\text{PMI} = \log_2\left(\frac{\text{P(w1 et w2)}}{\text{P(w1) P(w2)}}\right) \tag{3.1}$$

Dans un premier temps, avons testé différentes approches pour créer ces paires de mots : avec ou sans patrons morphosyntaxiques, sans contrainte grammaticale séquentielle, avec une tolérance (définition d'une longueur de la fenêtre de cooccurrence autour du token traité) ou non. Comme nous allons utiliser les collocations sous forme de lemmes, tokens, upos et dépendances, nous espérons y trouver une mesure de la complexité lexicale et syntaxique.

Exemples de nos collocations-bigrammes de référence (lemmes) et de leurs valeurs suivant leur score PMI (élévé ou proche de 3) :

keyword	ngram	left	right	freq	freq_left	freq_right	pmi	md	lfmd
sin die	2	sin	die	8	14	8	17.20991	-0.8073549	-18.82462
escape game	2	escape	game	6	6	17	16.92980	-1.5025003	-19.93481
redorer blason	2	redorer	blason	6	10	11	16.82087	-1.6114347	-20.04374
tenant aboutissants	2	tenant	aboutissants	8	25	8	16.37341	-1.6438562	-19.66112
fake news	2	fake	news	6	17	9	16.34484	-2.0874628	-20.51977
arabes unis	2	arabes	unis	23	23	27	16.26238	-0.2313255	-16.72503
next gen	2	next	gen	6	9	21	16.03999	-2.3923174	-20.82462
bouc émissaire	2	bouc	émissaire	6	9	23	15.90874	-2.5235620	-20.95587
no deal	2	no	deal	7	15	19	15.66980	-2.5401083	-20.75002
Air austral	2	Air	austral	7	12	24	15.65470	-2.5552152	-20.76513
in extremis	2	in	extremis	12	42	13	15.50947	-1.9228321	-19.35514
foie gras	2	foie	gras	7	16	25	15.18077	-3.0291463	-21.23906
business model	2	business	model	6	41	9	15.07475	-3.3575520	-21.78986

FIGURE 3.3 – Aperçu des collocations aux PMI proches de 3

keyword	ngram	left	right	freq	freq_left	freq_right	pmi	md	lfmd
gouvernement pourrait	2	gouvernement	pourrait	7	2031	914	3.000597	-15.209315	-33.41923
préservation de	2	préservation	de	28	30	247401	3.001240	-13.208673	-29.41858
beaucoup d'	2	beaucoup	d'	67	1444	12299	3.001248	-11.949930	-26.90111
même jour	2	même	jour	36	4288	2225	3.001513	-12.845829	-28.69317
dans bureau	2	dans	bureau	33	24916	351	3.001549	-12.971324	-28.94420
a toutefois	2	a	toutefois	62	34079	482	3.001979	-12.061092	-27.12416
d' activité	2	d'	activité	31	12368	664	3.002099	-13.060971	-29.12404
pour continuer	2	pour	continuer	36	26786	356	3.002267	-12.845075	-28.69242
été atteints	2	été	atteints	6	11427	139	3.003130	-15.429175	-33.86148
été remise	2	été	remise	6	11427	139	3.003130	-15.429175	-33.86148
fait vivre	2	fait	vivre	9	6636	359	3.003254	-14.844088	-32.69143
où passe	2	où	passe	6	3202	496	3.003276	-15.429029	-33.86133
dernier serait	2	dernier	serait	9	3293	723	3.004154	-14.843189	-32.69053

FIGURE 3.4 – Aperçu des collocations aux PMI élevées

Les collocations avec une PMI > 3 sont les plus caractéristiques d'une écriture de niveau universitaire (par exemple, redorer + blason, tenants + aboutissants, plein + fouet, corroborer + résultat). Elles comprennent également des paires de mots relativement courants qui sont des tournures presque idiomatiques (par exemple, à juste titre).
En revanche, les collocations avec PMI > 1 sont moins complexes et utilisent des mots très courants, voire neutres (avoir, seulement) qui ne sont pas des marqueurs d'un discours de niveau universitaire. Ce sont aussi des collocations avec des mots de fréquence moyenne, mais peu fréquemment associés.
Celles de PMI autour de 1 incluent des mots fréquents, plus fréquemment associés (pour + cacher, avoir + symptôme) qui ne seront sans doute pas discriminantes des niveaux de langue avancés.
Ce sont ces considérations qui nous ont conduit à tester des sous-ensembles de collocations de référence ne comprenant que les collocations de PMI >1 ou PMI >3.
Chaque paire de mots extraite des textes de chaque copie a ensuite été recherchée dans la liste des collocations de référence. Les collocations appartenant aux différents quartiles ont été comptées et leur moyenne a été calculée sur chaque quartile (par exemple, somme des PMI des collocations de grade 1 divisée par le nombre total

de collocations de grade 1).
Ces opérations ont été faites pour des collocations portant sur les tokens, les lemmes, les étiquettes morphosyntaxique, les dépendances. On a utilisé des bigrammes, des trigrammes, des 4 et 5-grammes (seulement ceux-ci sont utiles pour les étiquettes morphosyntaxique car les bigrammes ou trigrammes ont des PMI <3, car très fréquemment associés) afin de capturer les constructions grammaticales.
Ces calculs ont été effectués avec R.
La liste ci-dessous résume les différentes étapes de l'ingénierie des caractéristiques :

1. Construction des collocations de référence (script R), choix du nombre de répétions, d'une structure syntaxique ou des étiquettes parmi lesquelles elles seront choisies, et calcul des PMI (UDpipe) qui seront gradées en niveaux correspondants aux quartiles (grades)

2. Annotation des copies (UDpipe)

3. Extraction des collocations des copies présentes dans le corpus de référence (UDpipe)

4. Calcul de leur nombre et taux par grade, ainsi que de la moyenne des PMI sur chaque grade (script R)

Caractéristiques ajoutées :

tx_collocN_VAR : nombre de N-grammes pour les VAR (token, lemme, upos, deprel) divisé par le nombre total de tokens de la copie, à la fois total (sans _gradeJ) ou pour chaque grade.

collocN_VAR_moy_pmi_gradeJ : moyenne des PMI des N-grammes (pour les VAR) appartenant au grade J (J =1,2,3,4) c'est à dire à un des quartiles de répartition des PMI du corpus de référence (quartiles calculés après élimination des PMI < x, x ayant varié dans nos tests).

Exemple :

tx_colloc2_tokens, tx_colloc2_tokens_grade1, colloc2_upos_moy_pmi_grade3, tx_colloc3_tokens_grade1, tx_colloc3_lemmes_grade4, colloc3_lemmes_moy_pmi_grade1, tx_colloc4_deprel_grade3, tx_colloc5_deprel_grade1, colloc5_deprel_moy_pmi_grade4

3.5 Exploration des corpus et des caractéristiques

Nous allons présenter une exploration des corpus et des caractéristiques afin de pouvoir à la fois mieux adapter ces dernières au moment du va-et-vient entre expérimentation et leur affinement, mais aussi pour leur interprétation ultérieure.
Quelques extraits du corpus consultables à l'annexe A.4, montrent immédiatement le poids que peut avoir la diversité lexicale.

3.5.1 Corpus des copies

Tout d'abord nous vérifions la distribution des niveaux de notre sélection de copies :

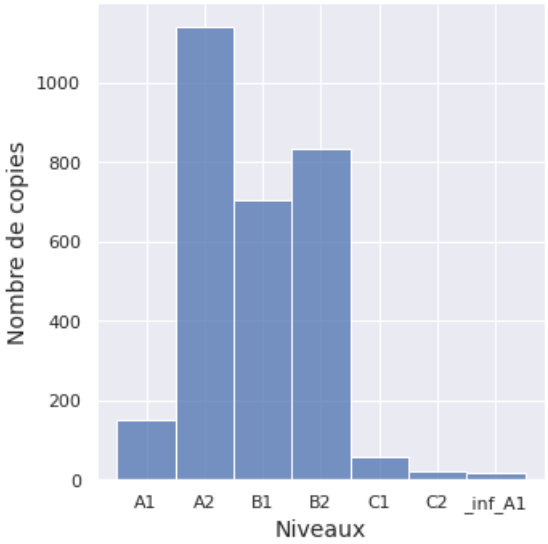

FIGURE 3.5 – Distribution des niveaux de notre corpus de copies_A1

qui correspond à celle du corpus de toutes les copies.

Nous donnons ici un aperçu de l'allure typique des distributions des données existantes, allure qui peut influencer certaines classifications et nécessiter une transformation.

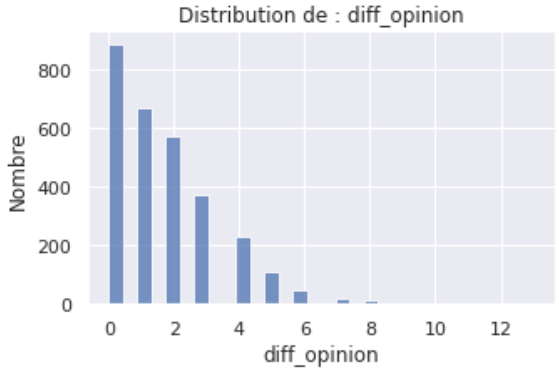

FIGURE 3.6 – Marqueurs d'opinion

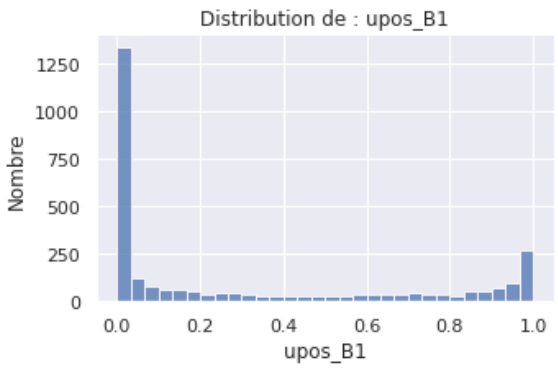

FIGURE 3.7 – Upos de niveau B1

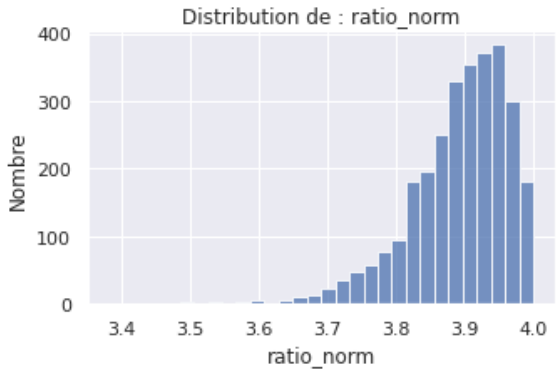

FIGURE 3.8 – Marqueur de normalisations

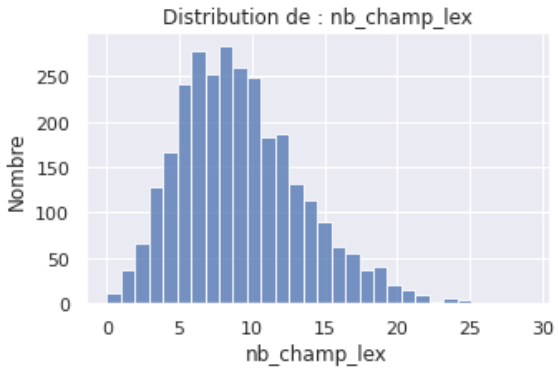

FIGURE 3.9 – Distribution de champ_lex

Toutes

les distributions ne sont pas centrées ni normales. Pour celles qui sont dissymétriques à droite et/ou ont une variabilité qui augmente lorsque la moyenne augmente, une transformation logarithmique pourrait permettre d'avoir une distribution normale, utile dans une optique d'amélioration pour certains classifiers.

D'autres caractéristiques existantes ont des outliers :

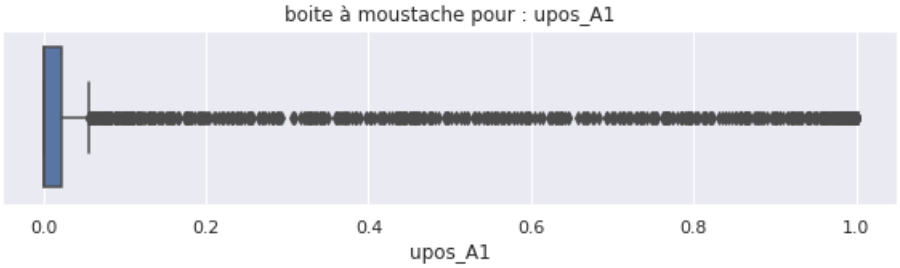

FIGURE 3.10 – Boîte à moustaches pour la caractéristique upos_A1

3.5.2 Comparaison avec le corpus de référence

Une analyse de la fréquence des étiquettes morphosyntaxiques nous montre que leur répartition varie entre les deux corpus :

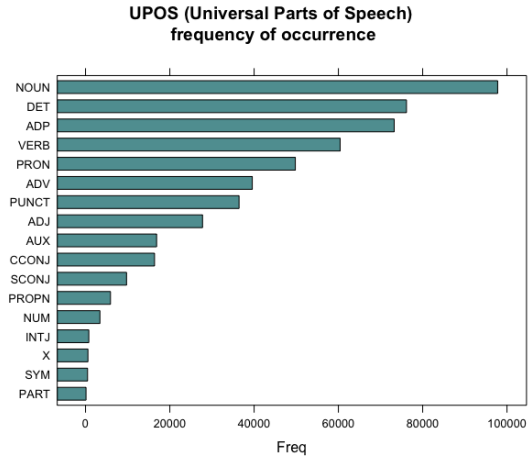

FIGURE 3.11 – Échantillon des copies

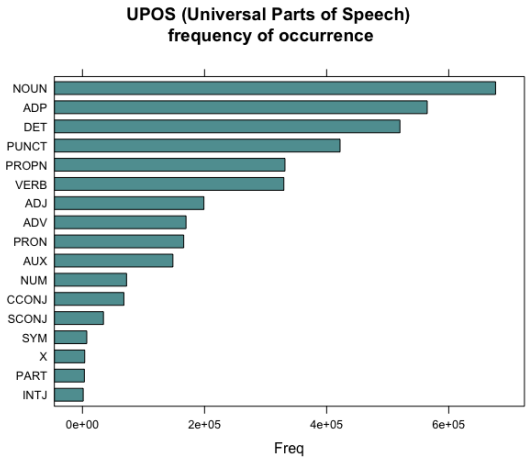

FIGURE 3.12 – Extrait du corpus Leipzig

Les candidats utilisent plus de déterminants et moins de modificateurs que dans les articles de presse, ils sont moins à l'aise avec une expression nuancée. D'autre part, la tâche considérée étant une argumentation, ils utilisent plus de pronoms.
La présence de ponctuations en moins grand nombre (rappel : la normalisation a supprimé les signes de ponctuation employés pour écrire les accents) reflète à la fois la présence de phrases plus courtes (les niveaux intermédiaires sont ceux les plus représentés) que dans un corpus journalistique, ainsi qu'une méconnaissance des règles de leur emploi à l'écrit.
Enfin, le corpus journalistique comprend évidemment bien plus de noms propres, trait inhérent dû aux particularités du traitement de l'actualité.

Dans notre construction de collocations, nous prendrons en compte justement les collocations formées avec des étiquettes moprhosyntaxiques de répartition similaire entre les deux corpus.

Un aperçu de la différence lexicale (sans surprise, étant donné leur nature) entre les deux corpus nous laisse penser que ce seront plutôt des caractéristiques qui ne prennent pas en compte les formes de surface qui pourraient être les plus efficaces (voir annexe : A.6)
Il en va de même pour les adjectifs (voir annexe : A.6).

Les collocations les plus fréquentes sont aussi évidemment différentes, puisqu'elles dépendent du champ lexical, mais cette différence se réduit considérablement quand on réduit le choix des collocations à un certain patron morphosyntaxique (phrases verbales simples). C'est d'ailleurs une des pistes que nous allons explorer (patrons ou une simple restriction sur les étiquettes morphosyntaxiques) :

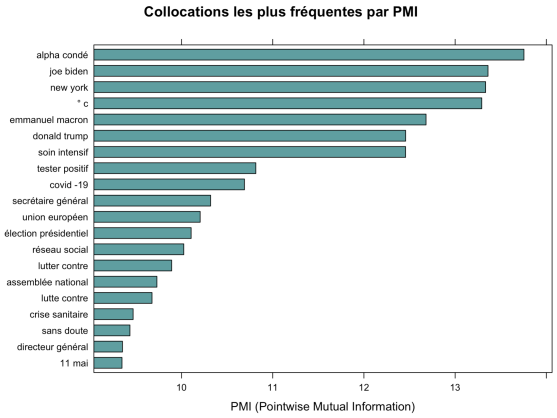

FIGURE 3.13 – Collocations (lemmes) les + fréquentes (PMI) - copies

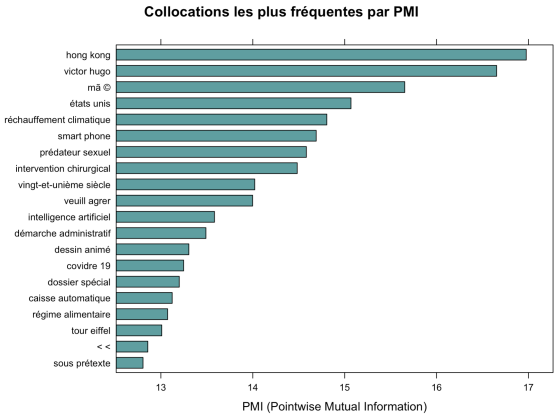

FIGURE 3.14 – Collocations (lemmes) les + fréquentes (PMI) - référence

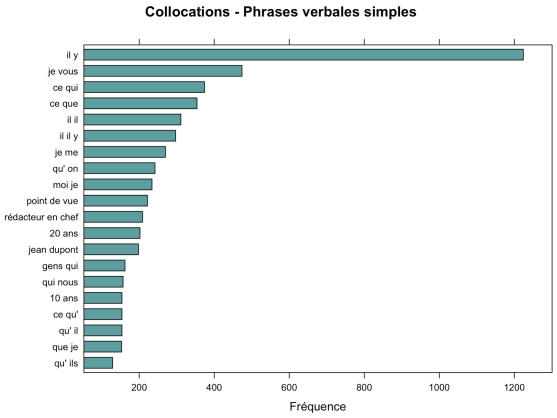

FIGURE 3.15 – Phrases verbales ou nominales simples - copies

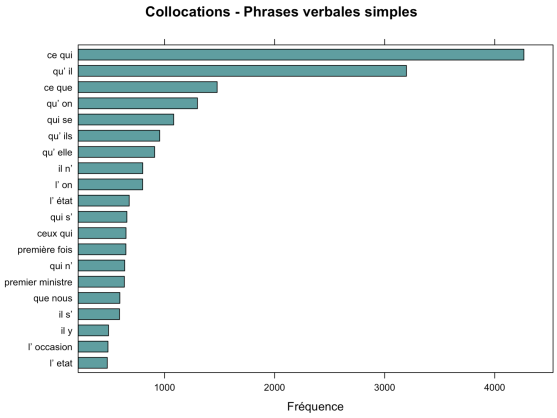

FIGURE 3.16 – Phrases verbales ou nominales simples - référence

Un survol lexicométrique (voir le figure A.6) du corpus des copies montre que les collocations argumentatives plus rares sont plus présentes dans les copies avancées : la présence de « En effet » ou de « cette affirmation » est discriminante des copies de niveau C2 alors que « À mon avis » ou « par exemple » l'est pour les copies plus faibles alors que « point de vue » est plus représenté dans les copies de niveau B1 à C1.
On a une confirmation presque intuitive de l'utilité des collocations catégorisées par leur force d'association.
Par ailleurs, malgré ses spécificités (voir les extraits : A.4), il est envisageable

d'exploiter les dépendances (voir figureA.5), nous cherchons certes à mieux évaluer les copies avancées, donc avec un étiquetage fiable, mais il semble que même celui des copies moins avancées puisse être exploitable.
Enfin, un graphe de lemmes en réseau, en fonction de leur associations (uniquement adjectifs et noms) montre que notre corpus de copies est doté d'un champ lexical délimité par l'énoncé de la tâche d'argumentation (voir annexe A.6). Il se pourrait que les caractéristiques liées au champ lexical aient un poids important dans la notation automatique.

3.5.3 Exploration des caractéristiques

Comme nous l'avons vu, les caractéristiques sont nombreuses. Nous avons procédé à un examen de leurs corrélations à l'aide du calcul des coefficients de Pearson et le tracé de matrices de corrélation (toutes les variables regroupées sur une seule matrice ne donnaient pas un résultat lisible), et des visualisations d'analyses multivariées, afin de nous conforter ou non dans le choix des caractéristiques à l'avenir et pouvoir éventuellement en réduire le nombre pour les modèles type SVM qui sont sensibles à la corrélation, ainsi que pour être éventuellement mieux outillés pour interpréter les résultats.

En voici un aperçu :

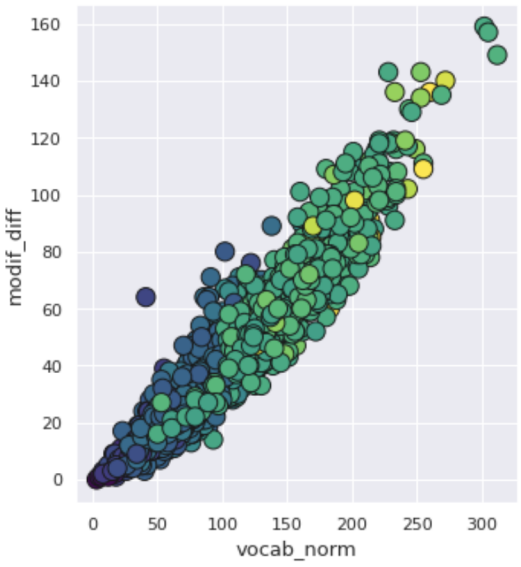

FIGURE 3.17 – Profondeur du vocabulaire / Modificateurs différents

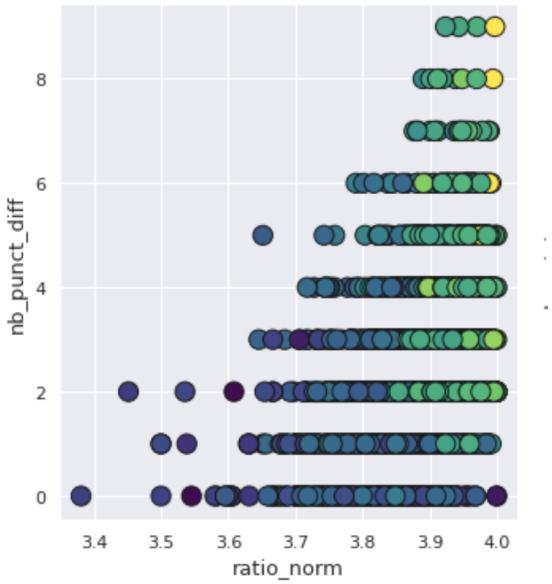

FIGURE 3.18 – Erreurs / Ponctuations différentes

Note : les copies aux scores les plus élevés sont de couleur claire, celles plus faibles sont plus foncées.
On voit que la profondeur du vocabulaire est positivement corrélée à la note (sans surprise), mais aussi à la quantité de modificateurs différents utilisés. On pourrait envisager, à l'avenir, de regrouper ces deux indicateurs.
Pour le taux d'erreurs ratio_norm (son calcul le fait varier positivement avec le score) on voit qu'il est meilleur pour les meilleurs copies et que la diversité de la ponctuation est effectivement un marqueur du niveau.

On a de nombreuses caractéristiques corrélées, mais qui reflètent chacune assez bien le niveau.

On peut le constater sur la matrice de corrélation des données originales en annexe : A.14.
Les distributions de nos caractéristiques sont soit symétriques, soit très déséquilibrées, cela étant dû d'une part au choix de ne prendre que des PMI > 3 et d'autre part à notre segmentation des taux en quartiles (grades dans le nom de de nos caractéristiques).
Note : les bigrammes d'étiquettes morphosyntaxiques étant très courants, il n'ont jamais une PMI > 3 et ne sont donc pas discriminants pour notre modèle.
La visualisation des distributions peut permettre de comprendre l'importance de chaque caractéristique pour l'algorithme (certaines prennent très peu de valeurs différentes par exemple) :

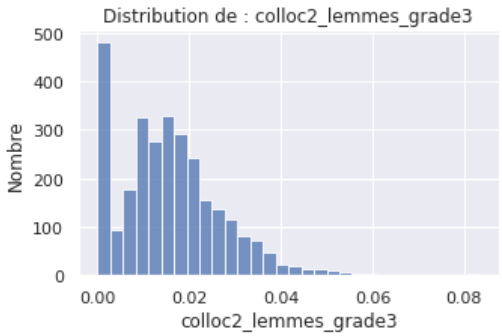

FIGURE 3.19 – Distribution des taux de bigrammes de lemmes dans le 1er quartile des PMI

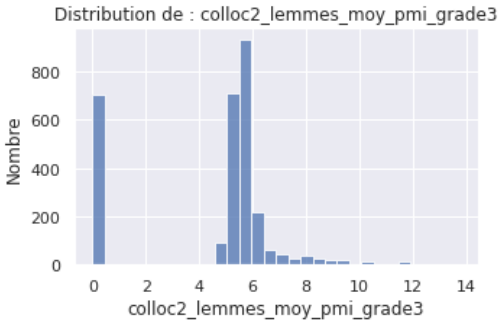

FIGURE 3.20 – Distribution des moyenne des PMI (du 3e quartile) des bigrammes de lemmes

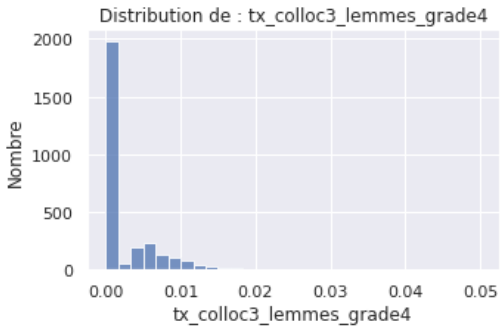

FIGURE 3.21 – Distribution des taux de 5-grammes de lemmes dans le 4e quartile des PMI

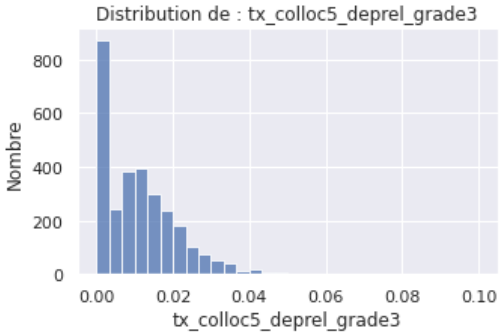

FIGURE 3.22 – Distribution des taux de 5-grammes de lemmes dans le 3e quartile des PMI

Par ailleurs la visualisation des distributions (sous forme de boîte à moustaches) par niveaux nous démontre la pertinence de certains indicateurs :

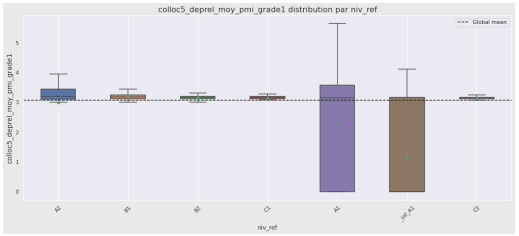

FIGURE 3.23 – 5-grammes de dépendances de grade 1

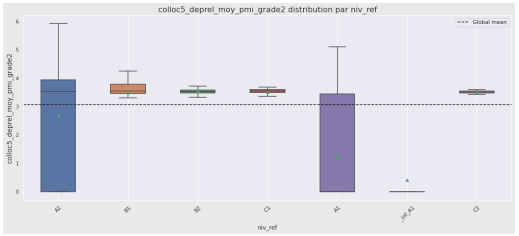

FIGURE 3.24 – 5-grams de dépendances de grade 2

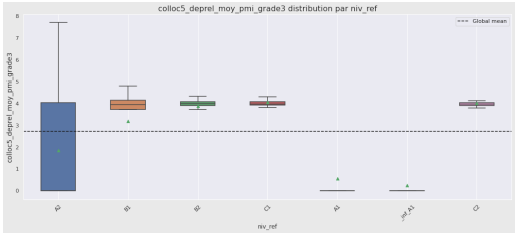

FIGURE 3.25 – 5-grams de dépendances de grade 3

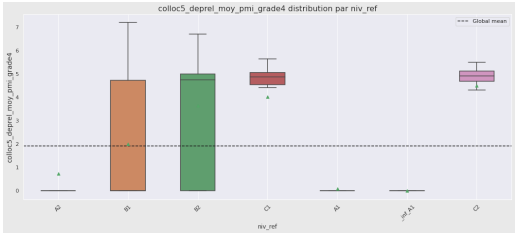

FIGURE 3.26 – 5-grams de dépendances de grade 4

D'autres départagent les niveaux de façon moins marquée : voir annexe A.8.
Les caractéristiques qui départagent le mieux les niveaux sont celles des n-grammes ($n \geq 3$), ou des indicateurs de grade extrême (appartenance au premier quartile ou au quatrième).
Les caractéristiques aux PMI dans le premier quartile discriminent les niveaux faibles.
Un exemple ci-dessous :

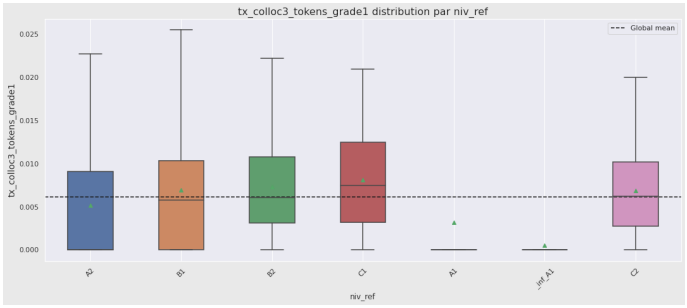

FIGURE 3.27 – Distribution des trigrammes de grade 1 (PMI dans le premier quartile)

Les nouvelles caractéristiques n'ajoutent pas de nouvelles corrélations fortes, voir annexe A.19.

CHAPITRE 4

RÉSULTATS

Sommaire

4.1 Introduction

Comme nous l'avons vu, la littérature montre que les combinaisons de mots, aussi bien sous forme de collocations que de collostructions jouent un rôle crucial dans l'acquisition et la maîtrise de la langue ([O'Donnell et al., 2013], [Stefanowitsch and Gries, 2003]) et que la complexité d'une langue seconde ne peut se réduire à la complexité syntaxique.
Les mesures de la complexité lexicale, par exemple, sont toutes des mesures basées sur un seul mot. Un mot simple et fréquent, utilisé avec divers sens et diverses constructions plusieurs fois dans une seule copie, ne contribuera pas à la diversité lexicale de cette copie.
La polysémie d'un mot n'entre en jeu et ne peut donc être évaluée qu'avec la prise en compte de son contexte.
Un exemple simple peut être donné avec le verbe « mettre », qui peut signifier aussi bien ranger, accrocher, s'habiller, enfiler, poser, poster, glisser, que disposer : j'ai mis la table, elle a mis les enfants à l'école, il ont mis deux jours à arriver à Paris, mettons que ce soit vrai, il a mis son poème en musique, je l'ai mise au courant, mal à l'aise, la télé lui a mis ces idées dans la tête, mets la radio plus fort, etc.
Par conséquent, les évaluations de copies de français langue seconde ont besoin de mesures qui prennent en compte ces combinaisons.
C'est pourquoi l'ajout des collocations, au sens large, aux caractéristiques déjà existantes peut permettre de saisir la complexité et donc le niveau d'une production écrite, car les mesures traditionnelles de la complexité ne parviennent pas à saisir une grande partie de la complexité de l'utilisation de la langue [Paquot, 2018].
Par ailleurs, nous avons choisi de prédire le score de la copie et non son niveau pour deux raisons :

Des expériences ont montré les limites du système d'appréciations globales du type « Très bien, Bien, Satisfaisant, Faible, Insuffisant » qui avaient été adoptées dans l'idée d'augmenter la concordance entre évaluateurs, mais ne reflètent pas assez, surtout pour les catégories intermédiaires (l'appréciation « moyenne » étant sur-représentée [Leclercq et al., 2004]), toute la palette des nuances possibles.
Ensuite, il est envisagé, à l'avenir, de ne prédire que certains critères qui s'additionnent pour donner la note finale (exemple) et ainsi réaliser une prédiction plus fine (il n'a pas été possible d'extraire assez rapidement suffisamment de données annotées critère par critère pour pouvoir travailler immédiatement sous cet angle).

4.2 Les outils et métriques de comparaison

Nous prendrons le jeu de données dont nous avons décrit les caractéristiques ci-dessus (existantes) comme modèle de référence (baseline).
Comme les caractéristiques utilisant des n-grammes sont liées à celles issues des collocations, nous ferons la comparaison de notre modèle, augmenté à la fois avec et sans n-grammes, pour évaluer l'apport intrinsèque de nos nouvelles caractéristiques.
Les comparaisons pour évaluer ce qu'apportent les collocations seront faites à l'aide de la librairie Scikit-learn de Python (le script du modèle précédent a été fait en R) avec divers outils, appliqués sur le modèle de base, puis sur le modèle augmenté . Ces outils seront :

- un SVR (Support Vector Regression) ;
- une régression Elastic Net ;
- une régression Random Forest ;
- et un XGBoost.

Les machines à vecteurs de support (SVM) sont des modèles qui ont pour principaux points forts une efficacité dans les espaces à dimension élevées (et nous avons 78 caractéristiques au départ auxquelles nous en ajoutons au maximum 62) et leur polyvalence (différentes fonctions de noyau peuvent être spécifiées pour la fonction de décision).
Les SVR sont un type de SVM (Support Vector Machine) qui prend en charge la régression linéaire et non linéaire, via le choix du noyau.
Un SVR nous donne la possibilité de définir le niveau d'erreur acceptable pour notre modèle (ε - epsilon-) et trouvera une ligne appropriée (ou un hyperplan dans les dimensions supérieures) pour s'adapter aux données (d'où le nom en français de séparateur à vaste marge).
Nos tests nous ont mené à choisir un noyau gaussien pour notre SVR.
Le SVR est sensible à la présence de caractéristiques corrélées (plus d'une solution).
Elastic Net de Scikit combine des régressions Lasso et Ridge pondérées. Ce sont des régressions avec un paramètre de pénalité qui vise à minimiser la complexité et/ou à réduire le nombre de caractéristiques utilisées dans le modèle final. L'objectif, comme toujours, est de réduire l'erreur sur l'ensemble de test.
Le Random Forest se base sur l'assemblage d'arbres de décision indépendants qui ont chacun une vision partielle des caractéristiques (feature sampling) et des observations (tree bagging) par un tirage aléatoire.
La prédiction finale est obtenue en faisant la moyenne de tous les arbres.
Le défaut du Random Forest est qu'il peut être vu comme une boîte noire qui donne des résultats peu facilement, c'est-à-dire peu explicatifs, c'est pourquoi nous regarde-

rons aussi les caractéristiques qui y ont le plus de poids.
L'eXtreme Gradient Boosting combine les résultats d'un ensemble de modèles plus simples et plus faibles afin de fournir de meilleurs résultats en termes de prédictions. Il utilise une approche séquentielle (Boosting) et construit le modèle par incréments en mettant l'accent sur les observations que les modèles précédents ont mal classées, tout en optimisant leur paramètres par descente de gradient.
Ainsi, un modèle apprend des erreurs d'un autre, ce qui renforce l'apprentissage, mais en contrepartie, il est sujet au suraprentissage.
Tous ces algorithmes seront paramétrés au mieux grâce à un choix des hyperparamètres fait grâce à la méthode classique d'optimisation GridSearchCV de Scikit-learn qui teste et compare tout un éventail de combinaisons possibles.
Les données seront séparées en données d'entraînement et de test en proportions respectives de 80 % et 20 % et l'entraînement utilisera une validation croisée à cinq plis.
Les métriques de mesure de comparaison des modèles :
La racine de l'erreur quadratique moyenne (en anglais Root Mean Square Error) est une mesure des différences entre les valeurs prédites (par un modèle ou estimateur) et les valeurs observées (ou vraies valeurs). Elle représente la racine carrée de la moyenne arithmétique des carrés des écarts entre prévisions du modèle et observations :

$$\text{RMSE} = \sqrt{\frac{1}{n}\Sigma_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}$$

où y_i est une valeur prise par une variable expliquée et $\hat{y}_i$ la valeur telle qu'elle aurait été prévue par le modèle (elle peut être normalisée dans certains calculs par l'écart-type).
Pourquoi prendre la moyenne ? Elle nous permet d'estimer l'écart type σ de l'erreur pour une observation unique typique plutôt qu'une sorte d'« erreur totale ». En divisant par n, nous gardons cette mesure de l'erreur cohérente lorsque nous passons d'une petit nombre d'observations à un plus grand nombre (elle devient simplement plus précise lorsque nous augmentons le nombre d'observations).
Pour résumer, la racine de l'erreur quadratique moyenne est une bonne mesure à utiliser si nous voulons estimer l'écart type σ d'une valeur observée typique par rapport à la prédiction de notre modèle, en supposant que nos données observées puissent être décomposées comme : valeur observée = valeur prédite + bruit avec une distribution centrée (ces écarts sont appelés résidus lorsque les calculs sont effectués sur l'échantillon de données qui a été utilisé pour l'estimation ou ils sont appelés erreurs - de prédiction - lorsqu'ils sont calculés sur les données de test).
Si les réponses prédites sont très proches des réponses réelles, la racine de l'erreur quadratique moyenne (RQEM) sera faible (la valeur nulle, presque jamais atteinte en pratique, indiquerait un ajustement parfait aux données). Si les réponses prédites et réelles diffèrent considérablement - au moins pour certaines observations - elle sera grande.
Une valeur de zéro indiquerait un ajustement parfait aux données. Puisqu'elle est mesurée sur la même échelle, avec les mêmes unités que les observations, elle permet donc de comparer différents modèles basés sur les mêmes observations y.
Mais quid de la comparaison avec le découpage des données en un ensemble de données d'entraînement et de test ? Ou de la comparaison de différents indicateurs ?
Une solution possible est de la normaliser et c'est ce que nous avons fait. Nous avons choisi de la normaliser avec l'écart-type (on peut la normaliser par la moyenne,

l'écart interquartile ou l'écart entre y_{min} et y_{max}).
En effet, la normalisation avec l'écart-type représente le rapport entre la variation non expliquée par l'estimation et la variation globale de y. Si l'estimation explique toute la variation de y, rien n'est inexpliqué et le REQM, normalisé ou non, est égal à zéro.
Si l'estimation explique une partie et laisse une autre partie inexpliquée, mais qui est à une échelle similaire à la variation globale, le rapport sera d'environ 1. Tout ce qui est au-delà indiquera une variation ou un bruit beaucoup plus important que dans la variable elle-même et par conséquent une faible prévisibilité.

Nous utilisons aussi le coefficient de détermination (R2) qui donne des informations sur la qualité des prédictions d'un modèle. Un R2 de 1 indique que les prédictions de régression correspondent parfaitement aux données.
Pour le modèle final nous calculerons les erreurs de classification.

4.3 Tests et résultats

Nous avons construit différents modèles :
- Modèle de base sans n-grammes ;
- Modèle de base avec n-grammes ;
- Modèle de base + collocations à PMI > 3 ou PMI >1 avec plus de 5, 10, 20 répétitions (sans patron morphosyntaxique, mais avec une restriction sur les étiquettes morphosyntaxique, ou l'utilisation de divers patrons morphosyntaxique : voir la partie sur la construction des caractéristiques 3.4) lors de la constitution du corpus des collocations académiques ;
- Modèle test avec uniquement les nouvelles caractéristiques ;
- Modèle de base + le meilleur modèle de collocations, avec et sans n-grammes.

4.3.1 Aperçu des expérimentations de régression

Les résultats sont comparés dans les tableaux ci-dessous, avant de choisir les plus performants et les comparer en termes de précision et de rappel :
Note : l'algorithme XGBoost est en moyenne dix fois plus rapide que tous les autres, nous n'inclurons pas de données de durée d'exécution dans nos résultats.

Le modèle utilisant des collocations ayant une PMI >3 et ayant eu une fréquence minimum de 5 dans le corpus Lepizg est le plus performant (avec XGboost) par rapport au modèle de base sans les caractéristiques issues des n-grammes :

Modèle de base sans les n-grammes				**Modèle de base + 5 répétitions, PMI >3 sans n-grammes**		
Modèle	**R2_score**	**RMSE_relative**		**Modèle**	**R2_score**	**RMSE_relative**
Elasticnet Regression	0,7902	0,4581		**Elasticnet Regression**	0,7910	0,4572
Support Vector Machine	0,7999	0,4473		**Support Vector Machine**	0,7837	0,4651
Random Forest Regressor	0,8088	0,4373		**Random Forest Regressor**	0,8121	0,4335
XGBoost	0,8156	0,4294		**XGBoost**	0,8167	0,4281

TABLE 4.1 – Comparaison du modèle de base sans les n-grammes / avec les collocations ajoutées

Les autres modèles ont des performances proches :

Modèle de base +100 répétitions, PMI >3 sans n-grammes		
Modèle	R2_score	RMSE_relative
Elasticnet Regression	0,7893	0,4591
Support Vector Machine	0,7842	0,4645
Random Forest Regressor	0,8107	0,4351
XGBoost	0,8107	0,4350

Modèle de base + 20 répétitions, PMI >3 sans n-grammes		
Modèle	R2_score	RMSE_relative
Elasticnet Regression	0,7944	0,4534
Support Vector Machine	0,7842	0,4646
Random Forest Regressor	0,8103	0,4355
XGBoost	0,8135	0,4318

Modèle de base + 20 répétitions, PMI >1 sans n-grammes		
Modèle	R2_score	RMSE_relative
Elasticnet Regression	0,7905	0,4578
Support Vector Machine	0,7930	0,4550
Random Forest Regressor	0,8049	0,4417
XGBoost	0,8114	0,4343

Modèle de base + 10 répétitions, PMI >3 sans n-grammes		
Modèle	R2_score	RMSE_relative
Elasticnet Regression	0,7936	0,4543
Support Vector Machine	0,7791	0,4700
Random Forest Regressor	0,8130	0,4324
XGBoost	0,8122	0,4334

TABLE 4.2 – Comparaison de quelques modèles testés sans n-grammes

Pour en savoir plus sur nos nouvelles caractéristiques, nous les testons seules, pour comprendre quel pourcentage de la variance des données elles expliquent. Nous ajoutons aussi un exemple de modèle avec utilisation de patron lors de la construction des collocations. On voit que les collocations ont un bon pouvoir explicatif, et que les modèles avec patron sont moins performants que sans (voir premier tableau ci-dessus) :

Collocations 5 répétitions, PMI >3 seules		
Modèle	R2_score	RMSE_relative
Elasticnet Regression	0,5518	0,6695
Support Vector Machine	0,5600	0,6633
Random Forest Regressor	0,5986	0,6336
XGBoost	0,6082	0,6259

Modèle de base + 5 répétitions, PMI >3 sans n-grammes+ patron		
Modèle	R2_score	RMSE_relative
Elasticnet Regression	0,7872	0,4612
Support Vector Machine	0,7922	0,4559
Random Forest Regressor	0,8035	0,4433
XGBoost	0,8148	0,43037

Enfin, nous comparons le modèle de base avec n-grammes et avec l'ajout des caractéristiques issues des collocations :

Modèle de base avec les n-grammes		
Modèle	R2_score	RMSE_relative
Elasticnet Regression	0,8068	0,4395
Support Vector Machine	0,4334	0,7527
Random Forest Regressor	0,8247	0,4187
XGBoost	0,7314	0,5182

Modèle de base avec toutes les caractéristiques		
Modèle	R2_score	RMSE_relative
Elasticnet Regression	0,8052	0,4414
Support Vector Machine	0,4502	0,7415
Random Forest Regressor	0,8264	0,4167
XGBoost	0,7314	0,5182

Pour les deux, c'est avec une régression Random Forest que les résultats sont les meilleurs.
Les performances du SVM étant dégradées en présence de variables corrélées (et il en y en a quelques unes dans les caractéristiques existantes).
L'ancien modèle a un R2 légèrement moindre mais un meilleur score en termes d'erreurs.
L'allure générale, pour ces modèles, des valeurs prédites / valeurs réelles est la suivante :

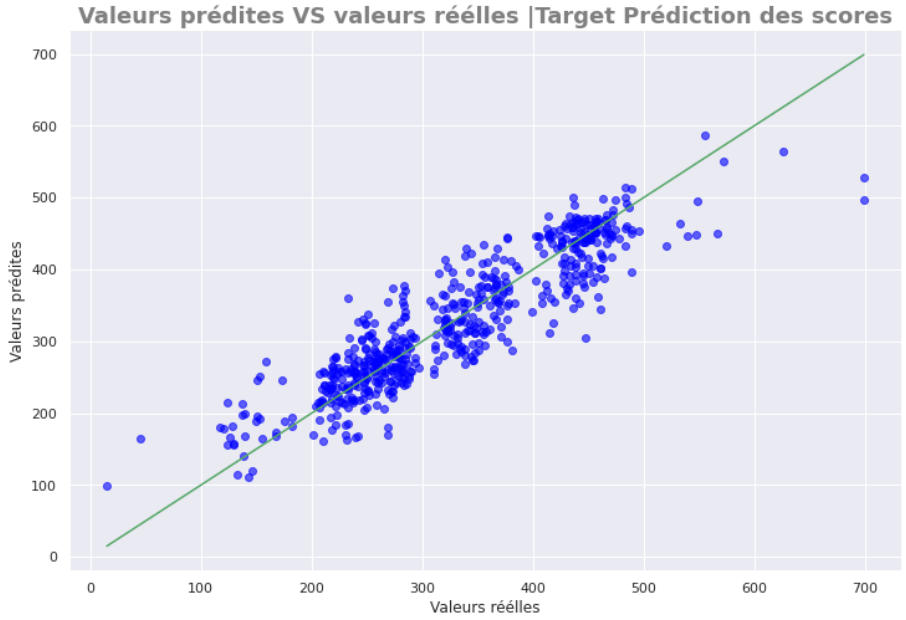

FIGURE 4.1 – Visualisation des différences entre valeurs prédites et valeurs réelles

On voit que les copies avancées (C1 et C2) sont sous-notées, alors que les copies faibles (A1 et moins) sont sur-notées.

Nous allons nous pencher sur les résultats de la régression et les transformer en classement.

4.3.2 Classements

Le classement par le modèle de base avec les caractéristiques n-grammes a la matrice de confusion suivante :

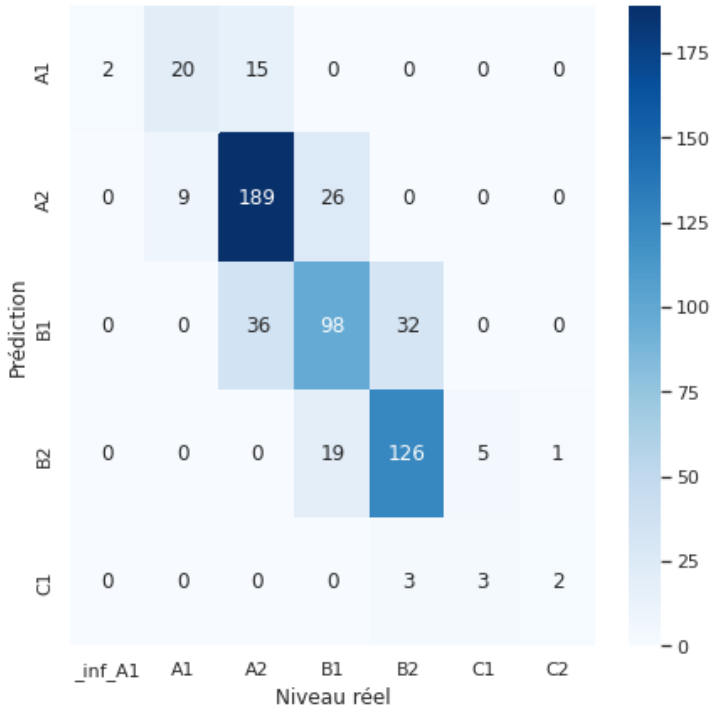

FIGURE 4.2 – Matrice de confusion pour le modèle augmenté

On a une précision de 0,744 et les rappels et précisions suivants pour chaque niveau :

	precision	recall	f1-score	support
A1	0.5405	0.6897	0.6061	29
A2	0.8438	0.7875	0.8147	240
B1	0.5904	0.6853	0.6343	143
B2	0.8344	0.7826	0.8077	161
C1	0.3750	0.3750	0.3750	8
C2	0.0000	0.0000	0.0000	3
_inf_A1	0.0000	0.0000	0.0000	2
accuracy			0.7440	586
macro avg	0.4549	0.4743	0.4625	586
eighted avg	0.7508	0.7440	0.7455	586

FIGURE 4.3 – Métriques de mesure du classement pour le modèle augmenté

Pour le modèle de base avec les n-grammes, le passage aux niveaux donne ceci :

On a une précision de 0,745 et les rappels et précisions suivants pour chaque niveau :

FIGURE 4.4 – Matrice de confusion pour le modèle de base avec les n-grammes

	precision	recall	f1-score	support
A1	0.5641	0.7586	0.6471	29
A2	0.8514	0.7875	0.8182	240
B1	0.5906	0.7063	0.6433	143
B2	0.8425	0.7640	0.8013	161
C1	0.2500	0.2500	0.2500	8
C2	0.0000	0.0000	0.0000	3
_inf_A1	0.0000	0.0000	0.0000	2
accuracy			0.7457	586
macro avg	0.4427	0.4666	0.4514	586
weighted avg	0.7556	0.7457	0.7477	586

FIGURE 4.5 – Métriques de mesure du classement pour le modèle de base avec les n-grammes

Les modèle sont presque équivalents au point de vue macroscopique avec un léger avantage pour le modèle augmenté.
Si on regarde par niveaux, on voit que pour le niveau C1, le modèle avec les collocations (modèle augmenté) est bien plus pertinent (3 sur 8 de bien classés contre 2 sur 8 : rappel de 0,36 contre 0,25).
Pour le niveau B2, le modèle augmenté a une précision légèrement moins bonne (0,83 / 0,84), mais un meilleur rappel (0,78 / 0,76). Pour les niveaux A1 et B1, c'est le contraire, les rappels du modèle augmenté sont respectivement très et légèrement

inférieurs à ceux du modèle de base (respectivement 0,69 / 0,76 pour A1 et 0,79 / 0,71 pour B1 alors que la tendance est inverse au-delà du niveau A2 (où les performances sont égales) : on a donc la confirmation que les collocations discriminent les niveaux avancés de compétence en langue.
En revanche, les deux modèles classent mal toutes les copies de niveau C2, et de la même façon (2 en avec un niveau d'écart (C1), et une copie avec deux niveaux d'écart, en B2).
Un calcul (voir A.20), fait par M. Casanova, a comparé sur R les performances de modèles construits avec et sans les variables collocatives sur un échantillon plus équilibré de copies (Cet échantillon comporte 5800 copies : 400 réputées de niveau A1, 400 réputées de niveau C2 et 1000 pour chacun des autres niveau. Il s'agit de l'échantillon qu'il a utilisé pour l'apprentissage des modèles n-grammes).
Les variables basées sur l'exploitation des modèles n-grammes ont donc été retirées des modèles utilisé pour la comparaison. Plusieurs constats ont pu être faits à cette occasion :

— si la modélisation SVM s'applique sans difficulté aux variables initiales, l'ajout de variables collocatives pose des problème de convergence (alerte signalée par la librairie utilisée) qui conduisent à une dégradation des performances
— dans tous les cas le modèle Random Forest (appliqué à 20 répartitions 80/10 différentes de l'échantillon en échantillons d'apprentissage et de test) produit les meilleurs résultats
— une amélioration de 0,5 % de accords de classement et du kappa est constatée. Cette amélioration est du même ordre que celle apportée par un autre stage réalisé en 2021 pour introduire de nouvelles variables linguistiques.

4.3.3 Expérimentation avec un nouvel échantillon

Un nouvel échantillon de 5800 copies, accompagné des caractéristiques pré-existantes a été obtenu après tous les tests précédents et nous a conduit à ajouter les résultats de cette expérimentation tardive, qui confirme une piste évoquée dans la discussion.
Tout d'abord, il convient de noter que comme nous avons toujours la même allure des données prédites / réelles, le nombre de copies aux extrémités étant artificiellement augmenté (l'échantillon est équilibré en niveaux, elles sont donc sur-représentées par rapport à la distribution réelle), la précision baisse, car c'est aux extrémités que nos algorithmes performent moins bien.

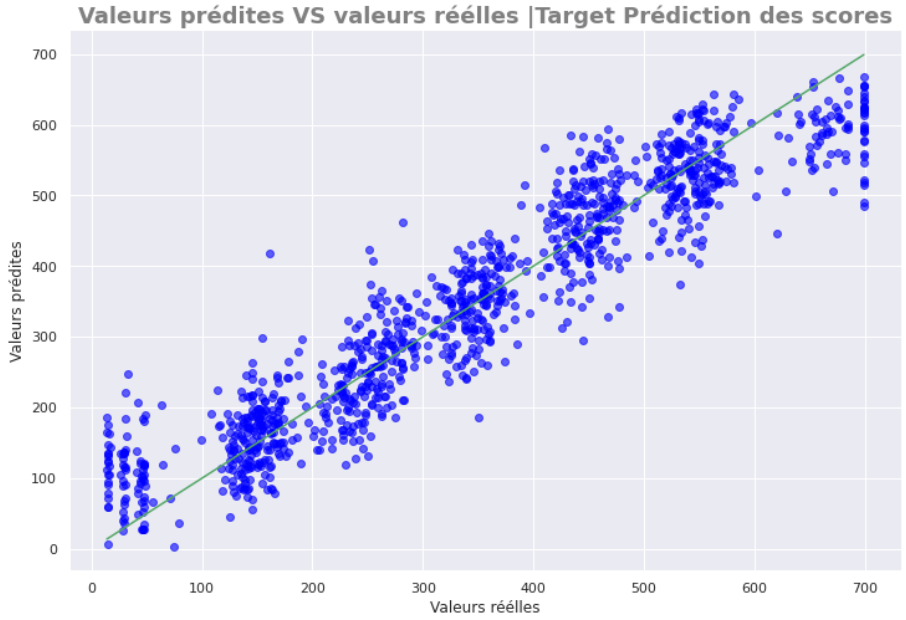

FIGURE 4.6 – Visualisation des différences entre valeurs prédites et valeurs réelles XGBoost- échantillon équilibré 5800 copies

On obtient les métriques suivantes, pour une PMI > 5, car celle >3 n'améliorait pas suffisamment le modèle aux extrémités, la piste de PMI plus élevées et de la particularité du français, par rapport à l'anglais semble se confirmer :

Modèle de base				**Modèle de base + 5 répétitions, PMI >5**		
Modèle	R2_score	RMSE_relative		Modèle	R2_score	RMSE_relative
XGBoost	0,88799	0,33468		XGBoost	0,889682	0,33214

On obtient les matrices de confusion suivantes :

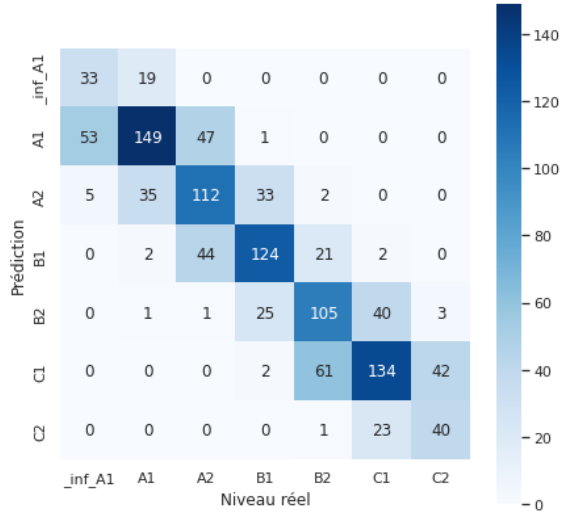

FIGURE 4.7 – Matrice de confusion pour le modèle de base -5800 copies

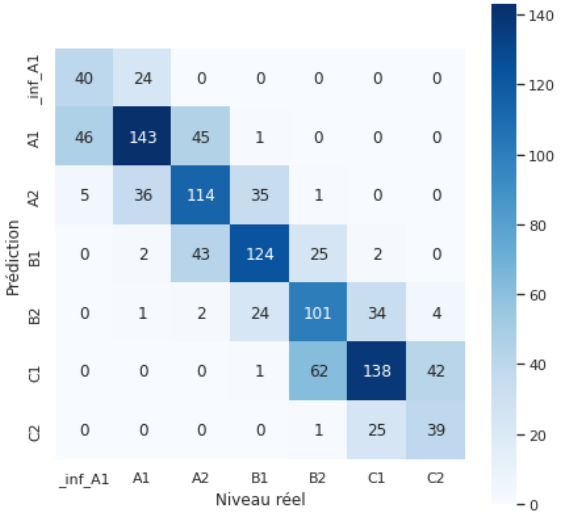

FIGURE 4.8 – Matrice de confusion pour le modèle augmenté -5800 copies

Et les précisions et rappels suivants :

	precision	recall	f1-score	support
A1	0.5960	0.7233	0.6535	206
A2	0.5989	0.5490	0.5729	204
B1	0.6425	0.6703	0.6561	185
B2	0.6000	0.5526	0.5753	190
C1	0.5607	0.6734	0.6119	199
C2	0.6250	0.4706	0.5369	85
_inf_A1	0.6346	0.3626	0.4615	91
accuracy			0.6009	1160
macro avg	0.6082	0.5717	0.5812	1160
weighted avg	0.6037	0.6009	0.5962	1160

FIGURE 4.9 – Rapport de classification pour le modèle de base-5800 copies

	precision	recall	f1-score	support
A1	0.6085	0.6942	0.6485	206
A2	0.5969	0.5588	0.5772	204
B1	0.6327	0.6703	0.6509	185
B2	0.6084	0.5316	0.5674	190
C1	0.5679	0.6935	0.6244	199
C2	0.6000	0.4588	0.5200	85
_inf_A1	0.6250	0.4396	0.5161	91
accuracy			0.6026	1160
macro avg	0.6056	0.5781	0.5864	1160
weighted avg	0.6040	0.6026	0.5991	1160

FIGURE 4.10 – Rapport de classification pour le modèle augmenté-5800 copies

On voit que le modèle augmenté performe mieux sur les niveaux B2 et C1 : meilleur rappel notamment et il sous-classe moins les C1. Il convient de noter que la différence sur le niveau C2 (en faveur du modèle de base) est minime.
Par contraste, le modèle augmenté améliore aussi le rappel du niveau inférieur à A1.

4.3.4 Importance des variables pour les algorithmes

Pour tous les modèles (de base, augmenté, PMI >3 ou PMI >5) on constate une importance de la diversité lexicale (vocab_norm), des pénalités appliquées (ratio_norm) et du dictionnaire de mots FLELex (ind1_Flelex), ce qui confirme l'intuition que nous en avions eu au moment de l'exploration des données (voir 3.5.3 et A.8) Un coup d'oeil sur l'importance des variables (voir l'annexe A.20 pour le classement entier) nous apprend que les deux algorithmes (XGBoost et Random Forest) ont accordé plus de poids, parmi les nouvelles variables.

Pour le modèle avec PMI > 5 :

- colloc3_tokens_moy_pmi_grade2
- tx_colloc2_tokens_moy_pmi_grade2
- tx_colloc2_tokens
- colloc5_deprel_moy_pmi_grade1

Pour le modèle avec PMI > 3 :

- colloc3_tokens_moy_pmi_grade2
- colloc5_deprel_moy_pmi_grade3
- tx_colloc2_tokens_moy_pmi_grade4

Il est intéressant de noter que les dépendances ont un poids assez important et que ce sont essentiellement les moyennes des PMI (sur chaque quartile) qui sont plus informatifs pour l'algorithme, qui a déja des informations de fréquence pure dans les autres caractéristiques n-grammes.
On voit aussi une différence suivant le seuil de PMI choisi, qui inévitablement élimine les bigrammes courants.

4.4 Conclusion

Les résultats obtenus sont déjà satisfaisants, et on constate que nos caractéristiques les améliorent légèrement, et ce, malgré la présence de caractéristiques certes issues de n-grammes, mais qui ne sont que statistiques.
Comme il n'existe pas en français de dictionnaire de collocations avec leur fréquence, leur efficacité dépend du corpus de référence utilisé, des types de relations ou d'étiquettes morphosyntaxiques prises en compte et du choix la mesure de la force d'association.
Il est sans doute possible d'obtenir des résultats équivalents, voire supérieurs en utilisant un réseau neuronal, ce qui ferait gagner du temps : plus de besoin d'élaborer des caractéristiques pour entraîner le modèle, mais qui est plus opaque et donc moins maîtrisable, ce qui n'est pas notre objectif de départ.

CHAPITRE 5

DISCUSSION

Sommaire

5.1 Introduction

Nos expérimentations, aussi bien lors la création des caractéristiques qu'au cours de leur adaptation au moment de la mise en œuvre des algorithmes de machine learning, nous ont démontré qu'il y avait encore de nombreuses pistes à explorer. Elles sont l'occasion d'une réflexion sur ce qui fait la compétence en langue, mais aussi sur ce qui sous-tend la notation humaine et comment les modéliser toutes deux avec des caractéristiques linguistiquement pertinentes afin de garder la main sur les décisions de l'algorithme.

5.2 Réflexions et pistes d'amélioration

Ce travail a étudié des indicateurs de la présence de collocations, sous la forme de taux de n-grammes, de tokens, de lemmes et d'étiquettes morphosyntaxiques.
Ces indicateurs ont été testés sous diverses formes (avec patrons morphosyntaxiques ou simplement distributionnels) et ils se sont révélés efficaces pour la détermination des compétences en français langue seconde, notamment pour les apprenants avancés, et donc pour améliorer la prédiction des scores des apprenants de niveau B2 et C1.
Il a donc a été montré que les mesures fondées sur des collocations de références étaient plus pertinentes pour l'évaluation de la compétence en langue pour les niveaux avancés que celles de la diversité lexicale et même argumentative.
Nos résultats montrent, même s'ils n'entraînent qu'une amélioration modeste des résultats du modèle précédent, que l'utilisation des moyennes d'information ponctuelles mutuelles sur chaque quartile sont aussi utiles que les taux de n-grammes par grade (voir résultat section importance des variables) et les caractéristiques créées expliquent 60 % de la variance des scores des copies à elles seules déjà, ce qui est encourageant.
Il est à noter aussi que les PMI pourraient être différentes et peut-être plus adaptées à des apprenants avancés, si calculées avec un autre corpus de référence.
On pourrait aussi envisager d'utiliser un jour un dictionnaire de collocations (équi-

valent des listes de fréquences de mots), comme il en existe en anglais.
Nous avons vu que les collocations 5-grammes et 4-grammes de dépendances avaient une bonne importance pour les algorithmes (aussi bien Random Forest que XGBoost), mais que celles de forme *adjectif + nom*, ou *verbe + conjonction* n'étaient pas assez puissantes pour distinguer les différents niveaux de compétence.
Par conséquent, il pourrait être fructueux d'explorer plus avant la piste des patrons de dépendances en se concentrant sur des relations plus complexes, ou en ajoutant une caractéristique avec des *modificateurs adverbiaux + verbe ou adjectif* avec une PMI > 6 (Read cité par[Paquot, 2018]), ce que conifrme notre début d'expérimentation avec l'échantillon de 5800 copies et une PMI > 5.
Nous avons donc vu que les collocations avec des PMI entre 1 et 3 (et c'est encore plus marqué entre 0 et 1) donnaient de moins bons résultats pour discriminer les niveaux, car ce sont celles constituées de mots à haute fréquence, donc un vocabulaire que l'on peut qualifier de basique.
La création de nos caractéristiques, première approche pour l'utilisation des collocations, n'a testé que peu de patrons pour former ces indicateurs et d'autres devraient être explorés dans des travaux futurs, car la littérature démontre que certaines combinaisons de mots couplées aux dépendances, comme par exemple *verbe + objet* permettent de distinguer les apprenants avancés ([Paquot, 2018]).
Mais il convient de garder à l'esprit que ces études portent sur l'anglais langue seconde et que le français, langue fléchie, n'a pas les mêmes marqueurs de niveaux ([Vandeweerd,]). Il est donc possible qu'il soit utile de tester des seuils différents pour les mesures de la force d'association.
L'évaluation de la qualité d'une collocations pourrait utiliser des modèles de langue comme GPT3, qui calculerait la probabilité du mot suivant d'après la séquence de mots le précédant, ce qui pourrait permettre de noter la vraisemblance du texte évalué. Pour ce qui concerne les autres pistes possibles d'amélioration, comme la tâche de notation porte sur des copies argumentatives, un examen plus poussé des collocations vues sous l'angle collostruction ([Paquot, 2018]) pourrait permettre l'extraction de relations sémantiques caractérisant l'existence d'un lien logique entre énoncés qu'ils soient mis en relation explicitement par un connecteur logique ou non (par exemple absence de « donc »), comme dans l'exemple suivant :
« Le téléphone est devenu trop présent dans notre quotidien. On passe moins de temps à parler aux autres. »
Comme ce type de relation rhétorique est implicite, on peut imaginer utiliser des approches statistiques portant sur des phrases adjacentes (voir les travaux de [Braud and Denis, 2013]), tout en gardant comme objectif la meilleure discrimination des copies avancées (l'étiquetage des dépendances pour les copies faibles n'étant pas fiable). Une autre piste d'amélioration de nos résultats est celle de la fragmentation de la notation automatique par critère d'évaluation, avec des caractéristiques *ad hoc* adaptées à chaque bloc de critères.

Conclusion générale

Dans ce mémoire, nous avons étudié l'utilisation des collocations pour améliorer un modèle de notation automatique de copies de tests de français. En effet, si un algorithme classe bien les copies, c'est qu'il les a « compris » à sa façon, et s'il ne les classe pas bien, il est utile de comprendre ce qu'il « dit » et « voit » et de « dialoguer » avec le modèle pour ajuster son classement.
C'est pourquoi des caractéristiques plus explicites peuvent être préférables pour permettre un ajustement « raisonné » des données.
Par ailleurs, la notation automatique de copies de test de français n'échappe évidemment pas aux questions qui se posent en matière d'utilisation de l'intelligence artificielle puisque le succès de cette notation est conditionné par le choix du type et de la quantité de données qui alimente le modèle algorithmique.
Il convient donc ne pas donner trop de responsabilité ni de « carte blanche » à des algorithmes souvent vus comme infaillibles et « neutres », alors que la composante humaine, et ses biais, est inévitablement présente dans leur construction.
Créer des caractéristiques adaptées que l'on maîtrise y contribue.

BIBLIOGRAPHIE

[gra, 2014] (2014). Coh-Metrix: Theoretical, Technological, and Empirical Foundations. In Graesser, A. C., McNamara, D. S., McCarthy, P. M., and Cai, Z., editors, *Automated Evaluation of Text and Discourse with Coh-Metrix*, pages 5–6. Cambridge University Press, Cambridge. – Cité page 34.

[Attali and Burstein, 2006] Attali, Y. and Burstein, J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment*, 4(3). – Cité pages 23, 24 et 34.

[Braud and Denis, 2013] Braud, C. and Denis, P. (2013). Automatically identifying implicit discourse relations using annotated data and raw corpora (Identification automatique des relations discursives « implicites » à partir de données annotées et de corpus bruts) [in French]. In *Proceedings of TALN 2013 (Volume 1: Long Papers)*, pages 104–117, Les Sables d'Olonne, France. ATALA. – Cité page 60.

[Brezina et al., 2015] Brezina, V., McEnery, T., and Wattam, S. (2015). Collocations in context:a new perspective on collocation networks. – Cité page 23.

[Casanova and Demeuse, 2016] Casanova, D. and Demeuse, M. (2016). Évaluateurs évalués : évaluation diagnostique des compétences en évaluation des correcteurs d'une épreuve d'expression écrite à forts enjeux. *Mesure et évaluation en éducation*, 39(3):59–94. – Cité page 19.

[Church and Hanks, 1989] Church, K. W. and Hanks, P. (1989). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th annual meeting on Association for Computational Linguistics -*, pages 76–83, Vancouver, British Columbia, Canada. Association for Computational Linguistics. – Cité pages 23 et 37.

[Crossley et al., 2011] Crossley, S., Salsbury, T., McNamara, D., and Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. – Cité page 21.

[Daller et al., 2003] Daller, H., van Hout, R., and Treffers-Daller, J. (2003). Lexical Richness in the Spontaneous Speech of Bilinguals. *Applied Linguistics*, 24(2):197–222. – Cité page 21.

[Dikli, 2006] Dikli, S. (2006). Automated Essay Scoring. *Turkish Online Journal of Distance Education 7(1)*. – Cité pages 23 et 24.

[Durrant and Schmitt, 2009] Durrant, P. and Schmitt, N. (2009). To What Extent Do Native and Non-Native Writers Make Use of Collocations? *International Review of Applied Linguistics in Language Teaching (IRAL)*, 47(2):157–177. – Cité pages 23 et 37.

[Ebeling and Hasselgård, 2015] Ebeling, O. and Hasselgård, H. (2015). Learners' and native speakers' use of recurrent word-combinations across disciplines. In

Learner Corpus Research: LCR2013 Conference Proceedin, pages 87–106. – Cité page 22.

[Ellis, 2012] Ellis, N. (2012). Formulaic Language and Second Language Acquisition: Zipf and the Phrasal Teddy Bear. *Annual Review of Applied Linguistics*. – Cité page 22.

[Evert, 2005] Evert, S. (2005). The statistics of word cooccurrences : word pairs and collocations. – Cité page 23.

[Filho et al., 2020] Filho, A. H., Concatto, F., do Prado, H. A., and Ferneda, E. (2020). Comparing Feature Engineering and Deep Learning Methods for Automated Essay Scoring of Brazilian National High School Examination. – Cité page 24.

[Foltz et al., 1999] Foltz, P., Laham, D., and Landauer, T. (1999). The intelligent essay assessor: Applications to educational technology. *undefined*. – Cité pages 23 et 24.

[Fonseca et al., 2018] Fonseca, E., Medeiros, I., Kamikawachi, D., and Bokan, A. (2018). Automatically Grading Brazilian Student Essays. In *PROPOR*. – Cité page 24.

[Friedman, 2001] Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, 29(5):1189–1232. – Cité page 24.

[Gablasova et al., 2017] Gablasova, D., Brezina, V., and McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence: Collocations in Corpus-Based Language Learning Research. *Language Learning*, 67(S1):155–179. – Cité pages 22 et 23.

[Granger and Bestgen, 2014] Granger, S. and Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. – Cité page 37.

[Granger and Tyson, 1996] Granger, S. and Tyson, S. (1996). Connector usage in the English essay writing of native and non-native EFL speakers of English. *World Englishes*, 15(1):17–27. – Cité page 34.

[Gries, 2013] Gries, S. (2013). 50-something years of work on collocations: What is or should be next .... – Cité page 22.

[Hausmann and Blumenthal, 2006] Hausmann, F. J. and Blumenthal, P. (2006). Présentation : collocations, corpus, dictionnaires. *Langue française*, 150(2):3–13. – Cité page 22.

[Heilman and Madnani, 2013] Heilman, M. and Madnani, N. (2013). ETS: Domain Adaptation and Stacking for Short Answer Scoring. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 275–279, Atlanta, Georgia, USA. Association for Computational Linguistics. – Cité page 24.

[Henriksen, 1999] Henriksen (1999). Vocabulary Semantics. – Cité page 22.

[Kahane and Polguere,] Kahane, S. and Polguere, A. Formal Foundation of Lexical Functions. – Cité page 22.

[Laufer and Nation, 1995] Laufer, B. and Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. – Cité page 22.

[Laugier and Weinberg, 1927] Laugier, H. and Weinberg, D. (1927). Le Facteur subjectif dans les notes d'examen. *L'Année psychologique*, 28(1):236–244. – Cité page 11.

[Leclercq et al., 2004] Leclercq, D., Nicaise, J., and Demeuse, M. (2004). Docimologie critique: des difficultés de noter des copies et d'attribuer des notes aux élèves. In Demeuse, M., editor, *Introduction aux théories et aux méthodes de la mesure en sciences psychologiques et en sciences de l'éducation*, pages 273–292. Les éditions de l'Université de Liège. – Cité pages 11, 23 et 48.

[Lesniewska, 2006] Lesniewska, J. (2006). Collocations and second language use. *Universitatis Iagellonicae Cracoviensis*, Instytut Filologii Angielskiej. – Cité page 22.

[Livingston and Lewis, 1995] Livingston, S. A. and Lewis, C. (1995). Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *Journal of Educational Measurement*, 32(2):179–197. – Cité page 36.

[McCarthy and Jarvis, 2010] McCarthy, P. M. and Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2):381–392. – Cité page 21.

[Nehm et al., 2012] Nehm, R. H., Ha, M., and Mayfield, E. (2012). Transforming Biology Assessment with Machine Learning: Automated Scoring of Written Evolutionary Explanations. *Journal of Science Education and Technology*, 21(1):183–196. – Cité page 24.

[Nesselhauf, 2005] Nesselhauf, N. (2005). *Collocations in a Learner Corpus*. John Benjamins Publishing. Google-Books-ID: ucI5AAAAQBAJ. – Cité page 22.

[O'Donnell et al., 2013] O'Donnell, M. B., Römer, U., and Ellis, N. C. (2013). The development of formulaic sequences in first and second language writing: Investigating effects of frequency, association, and native norm. *International Journal of Corpus Linguistics*, 18(1):83–108. – Cité pages 22 et 47.

[Paquot, 2018] Paquot, M. (2018). Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations. *Language Assessment Quarterly*, 15(1):29–43. – Cité pages 9, 37, 47 et 60.

[Paquot, 2019] Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, 35(1):121–145. – Cité page 22.

[Paquot and Granger, 2012] Paquot, M. and Granger, S. (2012). Formulaic Language in Learner Corpora. *Annual Review of Applied Linguistics*, 32:130–149. – Cité pages 11, 15 et 23.

[Ramesh and Sanampudi, 2021] Ramesh, D. and Sanampudi, S. K. (2021). An automated essay scoring systems: a systematic literature review. *Artificial Intelligence Review*, pages 1–33. – Cité pages 23 et 24.

[Shin, 2018] Shin, E. (2018). A Neural Network approach to Automated Essay Scoring: A Comparison with the Method of Integrating Deep Language Features using

Coh-Metrix. Department of Educational Psychology University of Alberta. – Cité page 24.

[Stefanowitsch and Gries, 2003] Stefanowitsch, A. and Gries, S. T. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2):209–243. – Cité pages 9 et 47.

[Taghipour and Ng, 2016] Taghipour, K. and Ng, H. (2016). A Neural Approach to Automated Essay Scoring. In *EMNLP*. – Cité page 24.

[Vajjala, 2017] Vajjala, S. (2017). Automated Assessment of Non-Native Learner Essays: Investigating the Role of Linguistic Features. *International Journal of Artificial Intelligence in Education*. – Cité page 34.

[Vajjala and Lėo, 2014] Vajjala, S. and Lėo, K. (2014). Automatic CEFR Level Prediction for Estonian Learner Text. – Cité page 23.

[Valette and Eensoo, 2014] Valette, M. and Eensoo, E. (2014). Approche textuelle pourle traitement automatique du discours évaluatif. *Langue francaise*, 184(4):109–124. – Cité pages 9 et 23.

[Vandeweerd,] Vandeweerd, N. Applying Phraseological Complexity Measures to L2 French: A Partial Replication Study | DIAL.pr - BOREAL. – Cité pages 25 et 60.

[Vlach, 2019] Vlach, H. A. (2019). Learning to Remember Words: Memory Constraints as Double-Edged Sword Mechanisms of Language Development. *Child Development Perspectives*, 13(3):159–165. – Cité page 22.

[Yannakoudakis et al., 2011] Yannakoudakis, H., Briscoe, T., and Medlock, B. (2011). A New Dataset and Method for Automatically Grading ESOL Texts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 180–189, Portland, Oregon, USA. Association for Computational Linguistics. – Cité page 24.

[Zareva et al., 2005] Zareva, A., Schwanenflugel, P., and Nikolova, Y. (2005). RELATIONSHIP BETWEEN LEXICAL COMPETENCE AND LANGUAGE PROFICIENCY: Variable Sensitivity. *Studies in Second Language Acquisition*, 27(4):567–595. – Cité page 21.

ANNEXE A

DOCUMENTATION

A.1 Grille d'évaluation

tef
TEST D'ÉVALUATION DE FRANÇAIS

GRILLE D'ÉVALUATION EE – 5 critères

FRANÇAIS DES AFFAIRES
FRENCH FOR BUSINESS

Critères	< A1	A1	A2	B1	B2	C1	C2
1 SECTION A Capacité à transmettre des informations	Absence d'observables, ou texte inintelligible.	Informations confuses, difficiles à comprendre. Peu d'éléments pertinents.	Récit très simple mais compréhensible ; descriptions sommaires, plutôt concrètes.	Récit simple, et direct comportant quelques détails ; le sens général est assez clair.	Récit clair et détaillé ; faits décrits de manière assez précise.	Récit rédigé avec aisance ; informations développées avec précision.	Récit limpide et approprié ; informations rédigées avec finesse dans un style sûr et personnel.
	☐	☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐
2 SECTION B Capacité à argumenter	Absence d'observables, ou texte inintelligible.	Texte confus. Peu d'éléments en lien avec la tâche à réaliser.	Texte compréhensible. Arguments très simples et pas toujours pertinents.	Texte exprimant un point de vue de façon simple mais assez claire. Arguments peu développés, notamment pour les aspects plus abstraits.	Texte clair ; arguments développés à l'aide de détails et/ou d'exemples significatifs.	Texte efficace et rédigé avec aisance ; point de vue confirmé à l'aide d'arguments secondaires.	Texte limpide, rédigé dans un style sûr et personnel ; arguments finement nuancés et appuyés.
	☐	☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐
3 Syntaxe (phrases, temps, modes, orthographe grammaticale)	Absence d'observables ou de structures grammaticales.	Structures élémentaires, répétitives, pas ou peu maîtrisées.	Structures simples ; erreurs élémentaires systématiques ; l'ensemble est compréhensible.	Phrases simples correctes, quelques phrases complexes ; temps et modes courants maîtrisés ; erreurs fréquentes mais ensemble clair.	Structures plus complexes, assez bien utilisées ; temps et modes variés et corrects ; les erreurs ne gênent pas la compréhension.	Structures et formes verbales variées et adéquates ; quelques erreurs sur les phrases les plus complexes.	Grande variété de structures complexes très bien maîtrisées et utilisées avec justesse ; des erreurs rares et difficiles à repérer.
	☐	☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐
4 Lexique (étendue, maîtrise, orthographe lexicale)	Absence d'observables, ou quelques mots ou expressions isolés.	Lexique très limité, répétitif et très approximatif voire erroné.	Lexique restreint, assez simple ; erreurs et confusions fréquentes.	Lexique plus varié mais surtout concret ; utilisation et maîtrise assez justes ; des erreurs sérieuses et quelques emprunts à d'autres langues.	Lexique assez varié, permettant l'expression d'idées complexes ; lacunes, erreurs ou approximations ne conduisant pas à des malentendus.	Lexique riche, précis et bien maîtrisé dans l'ensemble, permettant de s'exprimer sans restriction.	Lexique très riche, nuancé, très bien adapté et maîtrisé.
	☐	☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐
5 Cohérence et cohésion (organisation des textes, fluidité du discours)	Absence d'observables, ou production incohérente.	Quelques phrases juxtaposées ; quelques connecteurs élémentaires (et, mais).	Série de phrases reliées par des articulateurs simples ; logique des textes pas toujours évidente.	Textes organisés simplement, à l'aide de connecteurs courants ; les idées s'enchaînent mais pas toujours de façon logique.	Textes cohérents, progression des idées clairement marquée ; liens logiques variés mais pas toujours adéquats.	Textes fluides, bien structurés, mettant en valeur les points importants ; aucune rupture logique dans la progression des idées.	Textes très fluides, finement articulés ; le lecteur est parfaitement guidé.
	☐	☐	☐ ☐	☐ ☐	☐ ☐	☐ ☐	☐

FIGURE A.1 – Grille d'évaluation des tests TEF

A.2 Travaux sur la compétence langagière

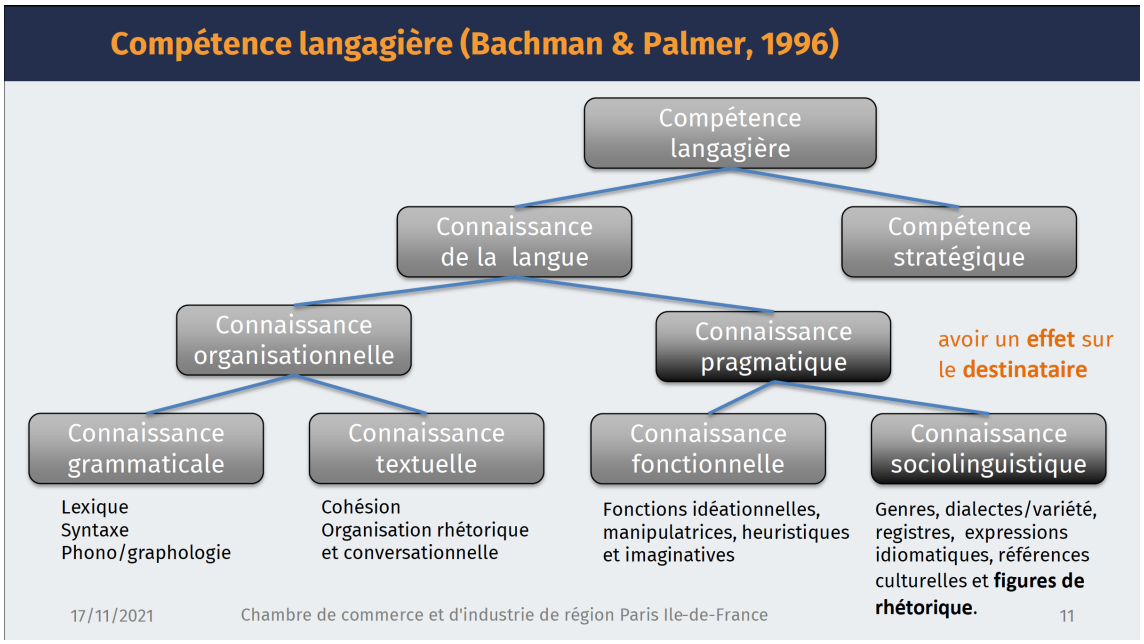

FIGURE A.2 – Compétences en langues

A.3 Les tableaux

Variable	Description
ratio_norm	1 - (erreurs normalisation/ nb tokens))
respect_genre_nombre	1-(nb_ecarts / nb_cas_accord)
vocab_norm	vocabulaire du candidat
nb_champ_lex	nombre de mot parmi la liste associée au sujet
ind_virgule	indicateur présence virgules
ratio_elision	1-(sqrt(nb_err_elision/nb_tokens))
ind1_FleLex	fréquence moyenne des mots utilisés
Expression	nb de mots dans liste exprimant une opinion
ind_A1	Fréquence de mots de niveau A1 selon FLeLex
ind_A2	Fréquence de mots de niveau A2 selon FLeLex
ind_B1	Fréquence de mots de niveau B1 selon FLeLex
ind_C1	Fréquence de mots de niveau C1 selon FLeLex
ind_C2	Fréquence de mots de niveau C2 selon FLeLex
ind_alt_A1	Fréquence de mots de niveau A1_alt selon FLeLex
ind_alt_B1	Fréquence de mots de niveau B1_alt selon FLeLex
ind_alt_B2	Fréquence de mots de niveau B2_alt selon FLeLex
ind_alt_C1_C2	Fréquence de mots de niveau C1-C2_alt selon FLeLex
Base1	Fréquence de mots de tranche 1 dans base mots
Base2	Fréquence de mots de tranche 2 dans base mots
Base4	Fréquence de mots de tranche 4 dans base mots
Base5	Fréquence de mots de tranche 5 dans base mots
Base6	Fréquence de mots de tranche 6 dans base mots
nb_temps_verb	Nombre de temps verbaux différents utilisés
coh_connecteurs	connecteurs logiques et conjonctions de coordination
coh_redondance	cohérence sur base de mots communs du texte
ratio_connecteurs	connecteurs_total/nb_tokens
connecteurs_diff	nb de connecteurs logiques différents
connecteurs_categ	Nombre de catégories de connecteurs mobilisées
ratio_conj_coord	conj_coord_total/nb_phrases
conj_coord_diff	nb de conjonctions de différentes
diff_anaphores	Nombre d'anaphores différents
diff_relatifs	diff_relatifs_simples + diff_relatifs_complexes
modif_diff	nb types de modificateurs différents utilisés
ratio_punct	nb_punct/nb_tokens
nb_punct_diff	nombre de ponctuations différentes utilisées
nb_opinion	nombre de marqueurs d'opinion
diff_opinion	nombre de marqueurs d'opinion différents
nb_champ_lex	nombre de mos se situant dans le champ lexical du sujet

TABLE A.1 – Définitions des caractéristiques pré-existantes

A.4 Extraits du corpus des copies

Inférieur à A1 : « Oui j aime apprendre une langue etrangere. parce que j aime tres bien apprendre français et c'est important dans ma vie. » ;

A1 : « A mon avis, Le telephone portable utilisee bien mais c'est ne pas de problème

les familles, ensuit pas bien utiliseé des portable c'est beacoup de problèm pour de tout les familles, pour le examble, un personne à part d'voyage on a appelle avec demande quelle que chose » ;

A2 : « Tout dabord, je mappelle charlotte Dupont, jeune femme âgéede 32, vivant dans la ville de meylan qui est une petite ville. D'ou j expliquerai en quelques mots la difference de ne pas se faire des amis dans les grandes villes mais plutôt dans des petites villes. tout d'abord les petites villes sont beaucoup plus habitées par des personnes âgé » ;

B1 : « Ce que je pense lorsque les enfants n'obéissent pas à leurs parents c'est parce que il y a manque d'éducation et compréhension touts parents doit véillé à l'éducation de ses enfants c'est 'est très important. MOI je ferais tout ce est à mon pouvoir à ce que mon enfant m'obéit. TOUT ceci dépend des parents, pourque les enfants obéissent aux parents c'est dés le bas age que tout cela commence. DONC les paraents doivent être impliquer c'est 'est très important, pour cela nous sommes les premiers responsable de la désobéissance de nos enfants. TOUTS parents doivent donner le meilleuir d'eux-mêmes c'est tres important pour l'éducation des enfants donc nous sommes res » ;

B2 : « Depuis son apparition, le téléphone portable est devenu un objet incontournable dans notre vie quotidienne. C'est pourquoi les donner aux enfants est une décison compliquée à prendre qui mérite une profonde réfléxion. Il est tout à fait possible de donner ces appareils aux enfants, mais nous devons considérer les points suivants. En premier ligne, il faut considérer l'âge de l'enfant. Même que l'éducation numérique démarre depuis très jeune age, l'usage journalière de cet équipement peut être dangereuse à son dévéloppement intelectuel. Le contrôle par un adulte est impératif dans ce cas. Autrement, un téléphone portable est utile pour un adolescent, qui pourra se communiquer avec ses amis et ses parents, ou même accéder aux applications en ligne. Pour eux, comme pour les adultes, cet appareil répresente la continuité de ses intéractions régulières et ce serait inutile de les priver de l'accès. Finalement, en tout cas, il est important de surveiller l'usage pour éviter tout excès, soit dans le temps devant l'écran, soit dans les contenus accédés. Ce contrôle a pour but de préserver l'enfant et d'assurer son bon développement. » ;

C1 : « Je ne suis pas d'avis que le fait de rester à la maison soit le meilleur moyen de passer ses vacances. Je trouve que les gens ont du mal à se déconnecter de leur quotidien en restant chez-soi, malgré le fait qu'ils ne vont plus au travail. Moi-même je fais partie de ces gens-là. Dans le cas où je n'ai rien prévu pour la journée, je passe mon temps en m'occupant de la maison, parlant au mes proches par téléphone ou regardant la télé ; ce qui est loin d'être la meilleure façon de passer mes vacances ! En revanche, les vacances passées ailleurs sont toujours pleines de nouvelles impressions. Habitant en Europe, nous avons la possibilité de découvrir des nouveaux pays, des nouvelles cultures et rencontrer des personnes et des professions différentes aux prix tout à fait raisonnables. Je suis persuadée que le fait de sortir de sa coque nous permet tout d'abord d'obtenir de l'inspiration et des idées pour les nouveaux projets. Dans certains cas, le fait de voyager nous permet aussi d'apprecier d'avantage notre vie quotidienne et y remarquer des petits plaisirs qui s'échappaient à notre regard avant. Et quel plaisir est le rétour à la maison ! » ;

C2 : « Cher monsieur le rédacteur en chef, Je me permets de vous écrire afin de partager mon opinion vis-à-vis de l'article publié dans votre édition du 25 septembre dernier, et qui avait pour intitulé Ëparents d'aujourd'hui : figures ancestrales d'autorité ou victimes modernes ?Ën effet, une affirmation poignante y a été exprimée, selon laquelle, je cite : äujourd'hui, les parents n'ont plus d'autorité sur leurs enfants.̈ Je suis tout à fait d'accord avec cette affirmation et y adhère complètement, mais je pense cependant que votre article manquait de substance. En effet, il aurait été intéressant et pertinent de pousser plus loin ce débât, et d'essayer de répondre au pourquoi. Pourquoi les parents, aujourd'hui, n'ont-ils plus d'autorités sur leurs enfants ? Dans la société actuelle, la plupart du temps les deux parents occupent une activité professionnelle à temps plein, et passent beaucoup moins de temps avec leurs enfants au sein du foyer. Cet éloignement a pour effet de donner à l'enfant un faux sens d'indépendance, et il est beaucoup moins enclin à respecter l'autorité de parents trop souvent absents. Le temps passé par les enfants devant les films et les jeux vidéos de nos jours joue également un rôle important dans la relation parents- enfants. Les enfants v̈iventd̈ans une realite virtuelle, et sont deconnectes de la réalité. » ;

A.5 Aperçu de l'étiquetage de deux phrases pour un niveau inférieur à A1, et A1

Phrase 1 : « Oui j aime apprendre une langue etrangere. parce que j aime tresbien apprendre français et c'est important dans ma vie »
Phrase 2 : « Des parents qu'ont plus d'autorités sur leurs enfants ça provoque que ces derniers sont plus violents et ingérable. »

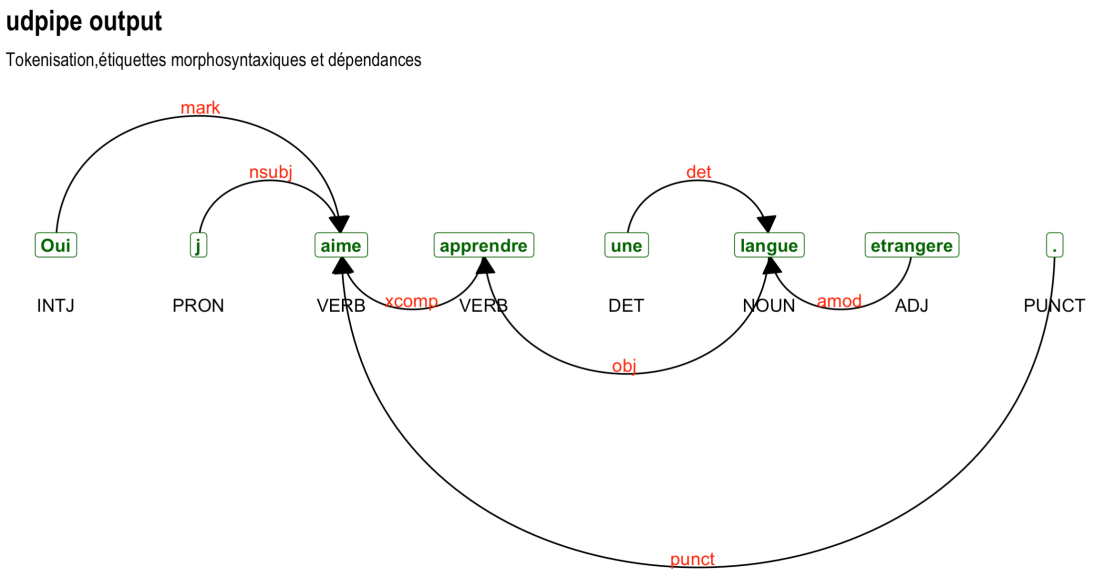

FIGURE A.3 – Phrase de niveau inférieur à A1

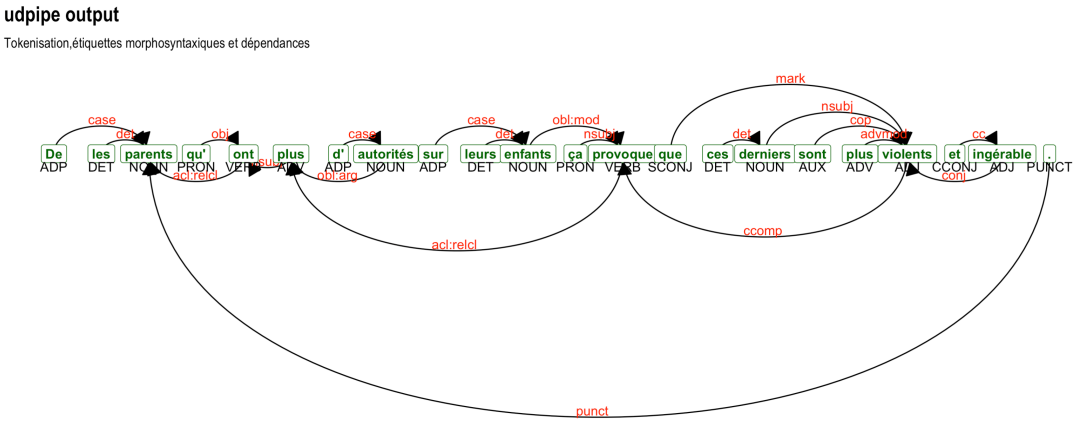

FIGURE A.4 – Phrase de niveau A1

A.6 Comparaison des corpus

Comparaison des noms les plus fréquents dans les deux corpus :

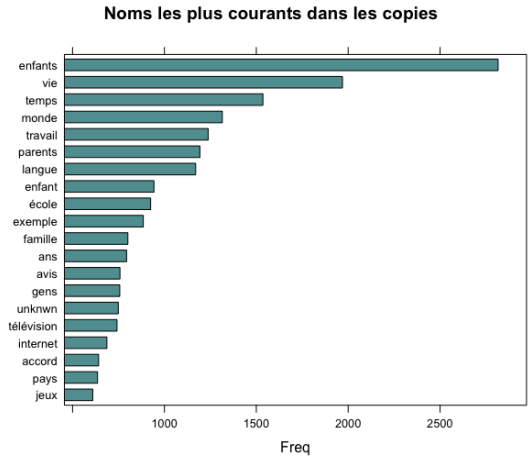

FIGURE A.5 – Corpus des copies

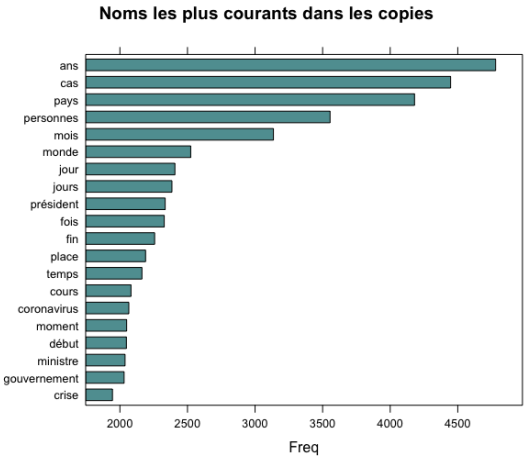

FIGURE A.6 – Corpus de référence

Comparaison des adjectifs les plus fréquents dans les deux corpus :

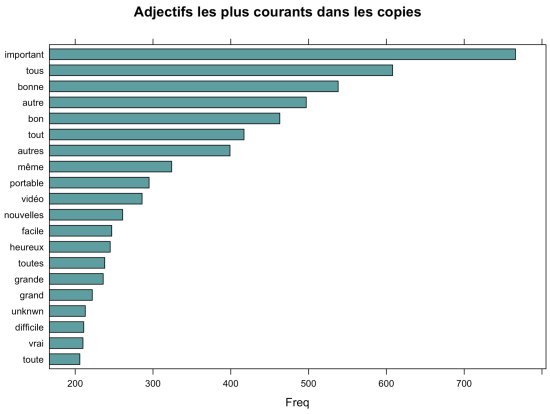

FIGURE A.7 – Adjectifs les plus courants dans le corpus des copies

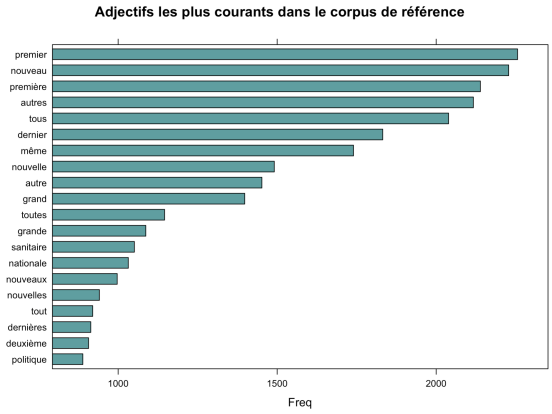

FIGURE A.8 – Adjectifs les plus courants dans le corpus de référence

Comparaison des collocations les plus fréquentes dans les copies et de leur forces d'association à l'intérieur d'une phrase :

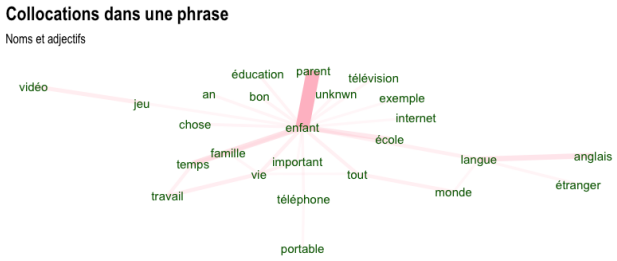

FIGURE A.9 – Corpus des copies

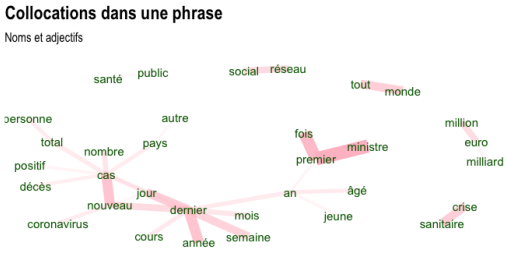

FIGURE A.10 – Corpus de référence

Comparaison des collocations les plus fréquentes dans toutes les copies / la références et de leur forces d'association en général :

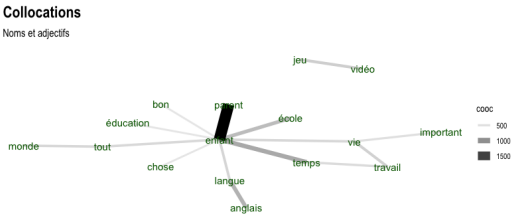

FIGURE A.11 – Corpus des copies

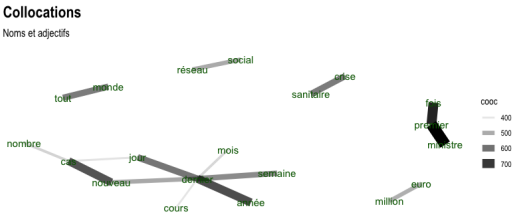

FIGURE A.12 – Corpus de référence

Visualisation des fréquences relatives de quelques collocations dans Lexico 5 :

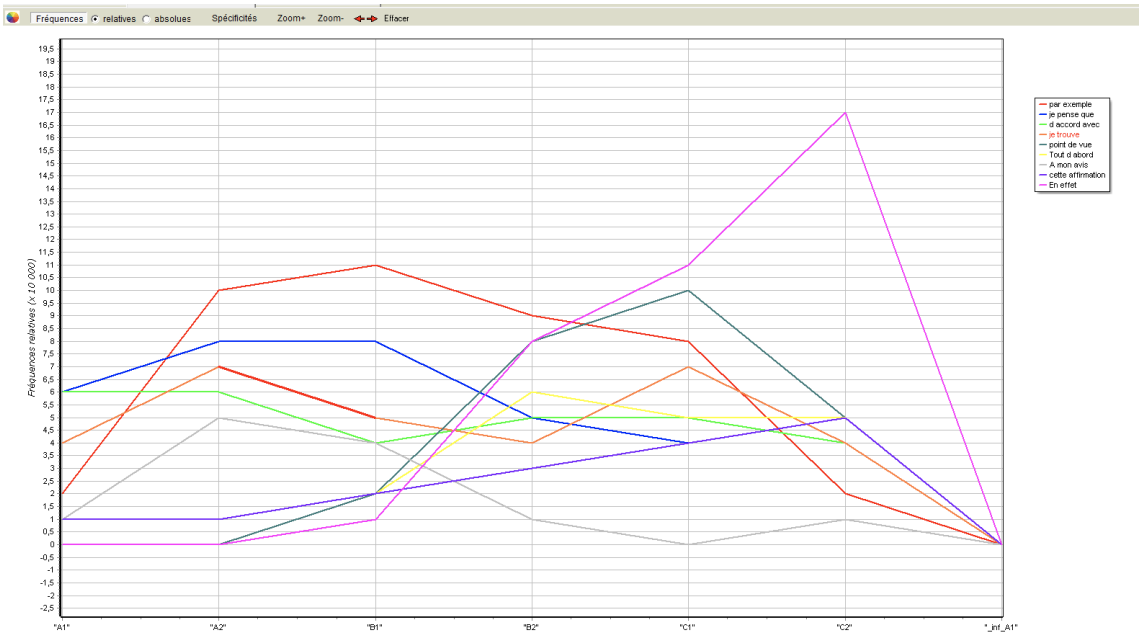

FIGURE A.13 – Fréquence de segments argumentatifs suivant les niveaux de la copie

A.7 Exploration des caractéristiques existantes

Ici, ce à quoi on peut prêter attention, ce sont les zones rouge foncé (corrélation positive forte) ou bleue foncée (corrélation négative forte) pour éventuellement dans l'avenir regrouper ou supprimer un des deux indicateurs concernés. (comme par exemple ici, diff_opinion et nb_opinion.
et surtout les indicateurs : noms_com, lem_com, stem_com, noms_com_reste, lem_com_reste, stem_com_reste, mot_com_moyenne, motComReste_moyenne.

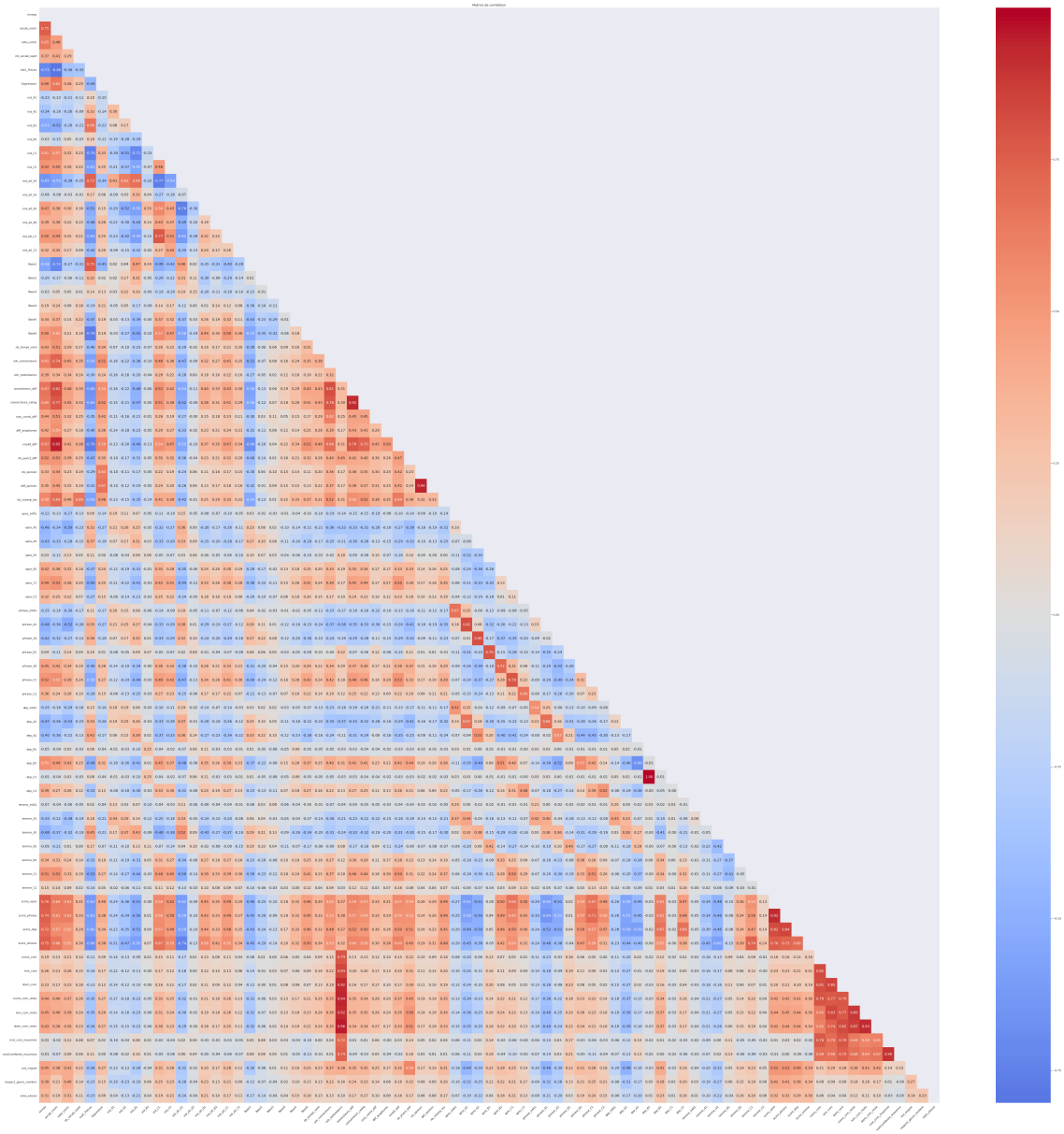

FIGURE A.14 – Matrice de corrélation des données de départ

A.8 Exploration des caractéristiques issues des collocations

Distributions des différents grades pour la moyenne des PMI pour les trigrammes de lemmes :

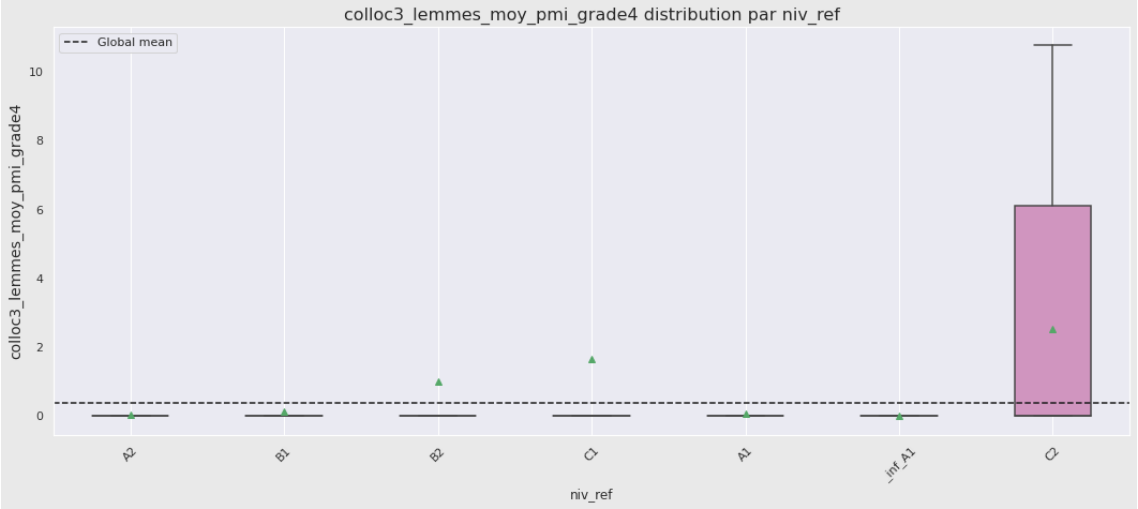

FIGURE A.15 – Trigrammes lemmes de grade 4

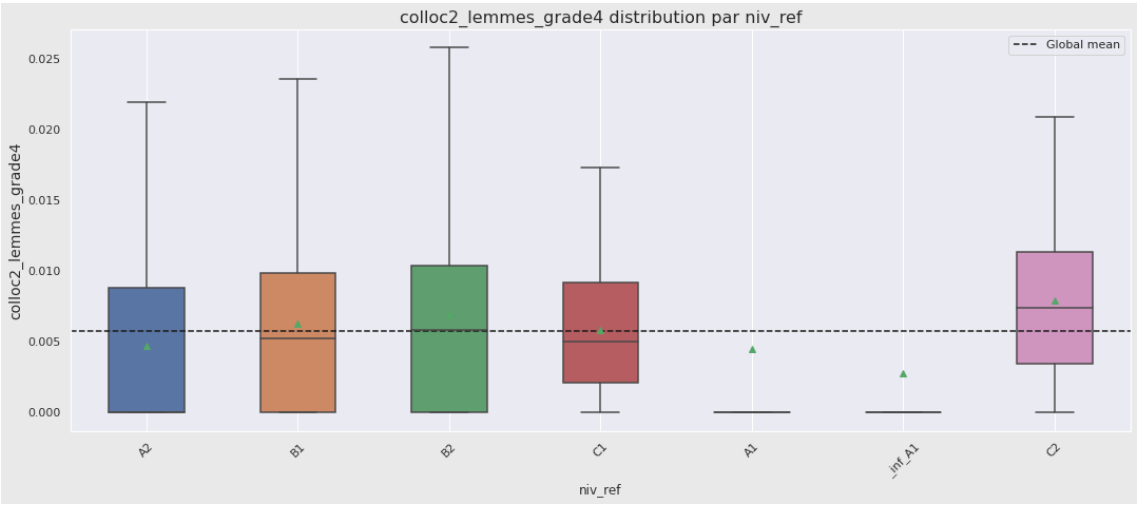

FIGURE A.16 – Trigrammes de lemmes de grade 1

Micro analyse multivariée sur la caractéristique issues du dictionnaire FLElex (mots classés par niveau de français) et vocabulaire ou connecteurs :

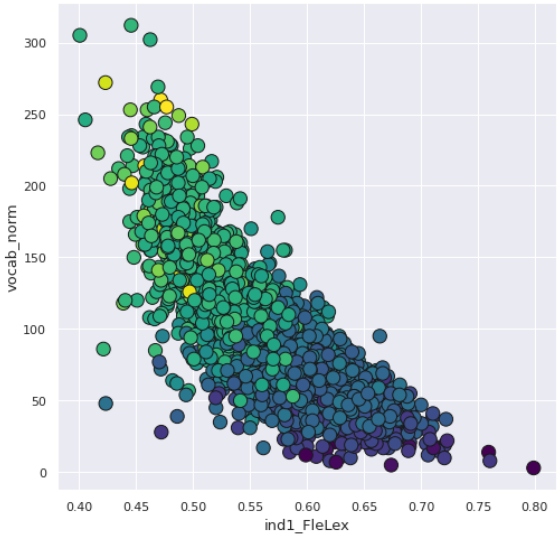

FIGURE A.17 – Variation de l'indice FLElex en fonction du vocabulaire

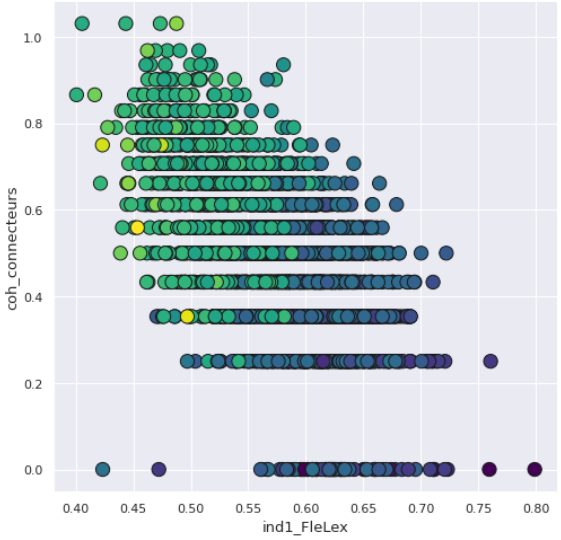

FIGURE A.18 – Variation de l'indice FLElex en fonction des connecteurs

Note : les niveaux les plus avancés sont de couleur claire, les plus faibles de couleur foncée.

A.9 Résultats

Matrice de corrélation des variables avec l'ajout des caractéristiques issues des collocations : l'intérêt de cette image est de voir que l'ajout des nouvelles caractéristiques n'ajoute pas de nouvelles corrélations fortes.

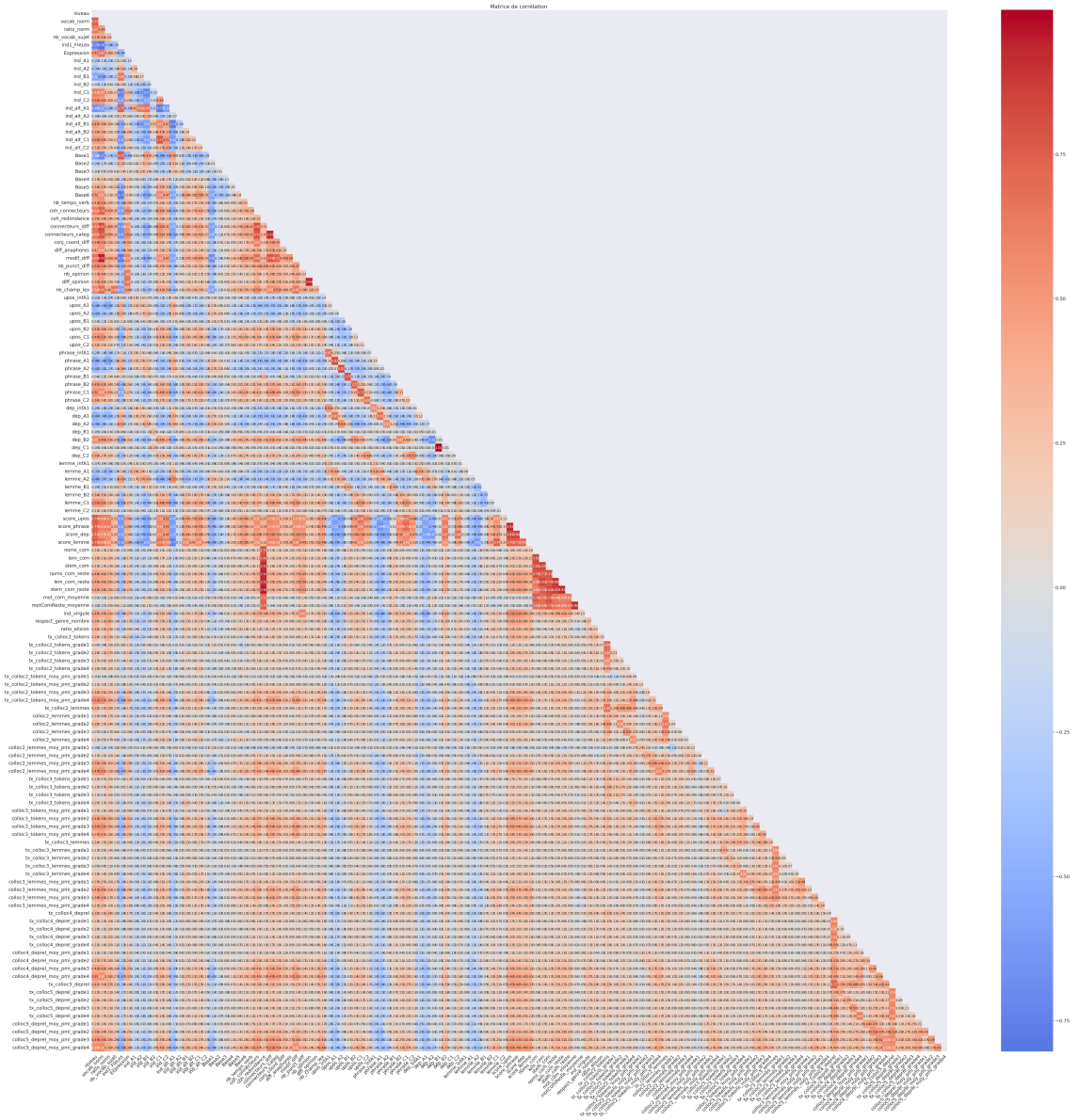

FIGURE A.19 – Matrice de corrélation des données finales

Comparaison des accords suivant l'année, entre modèle de base (2020) et modèle augmenté (2021) :

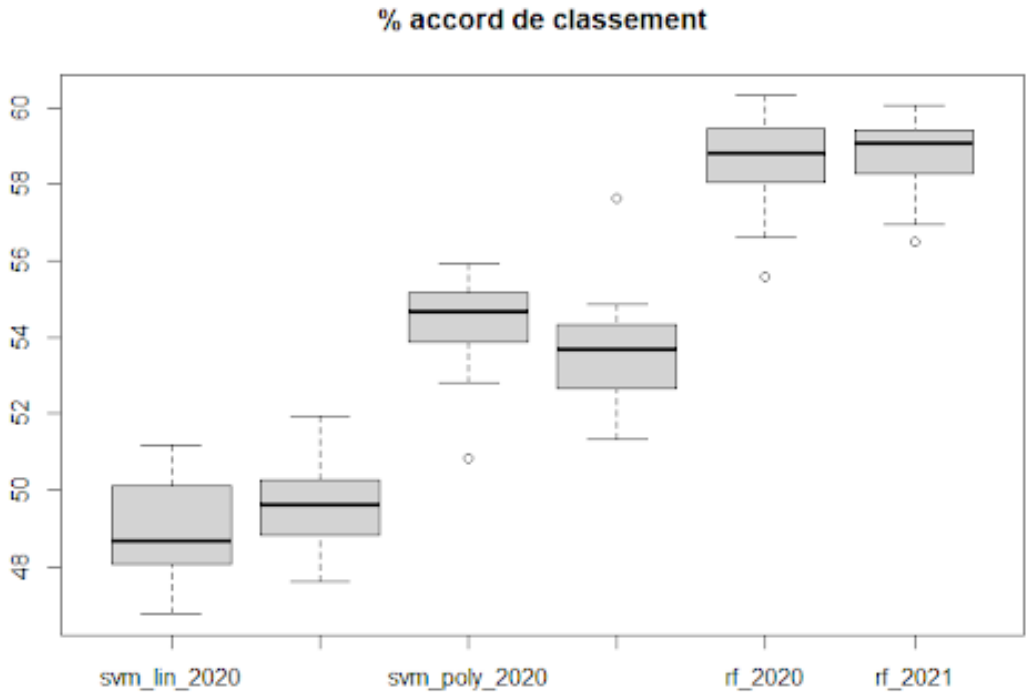

FIGURE A.20 – Comparaison des accords suivant l'année, par modèle (auteur : D.Casanova)

Importances des caractéristiques pour les modèles avec PMI > 3 et PMI > 5 :

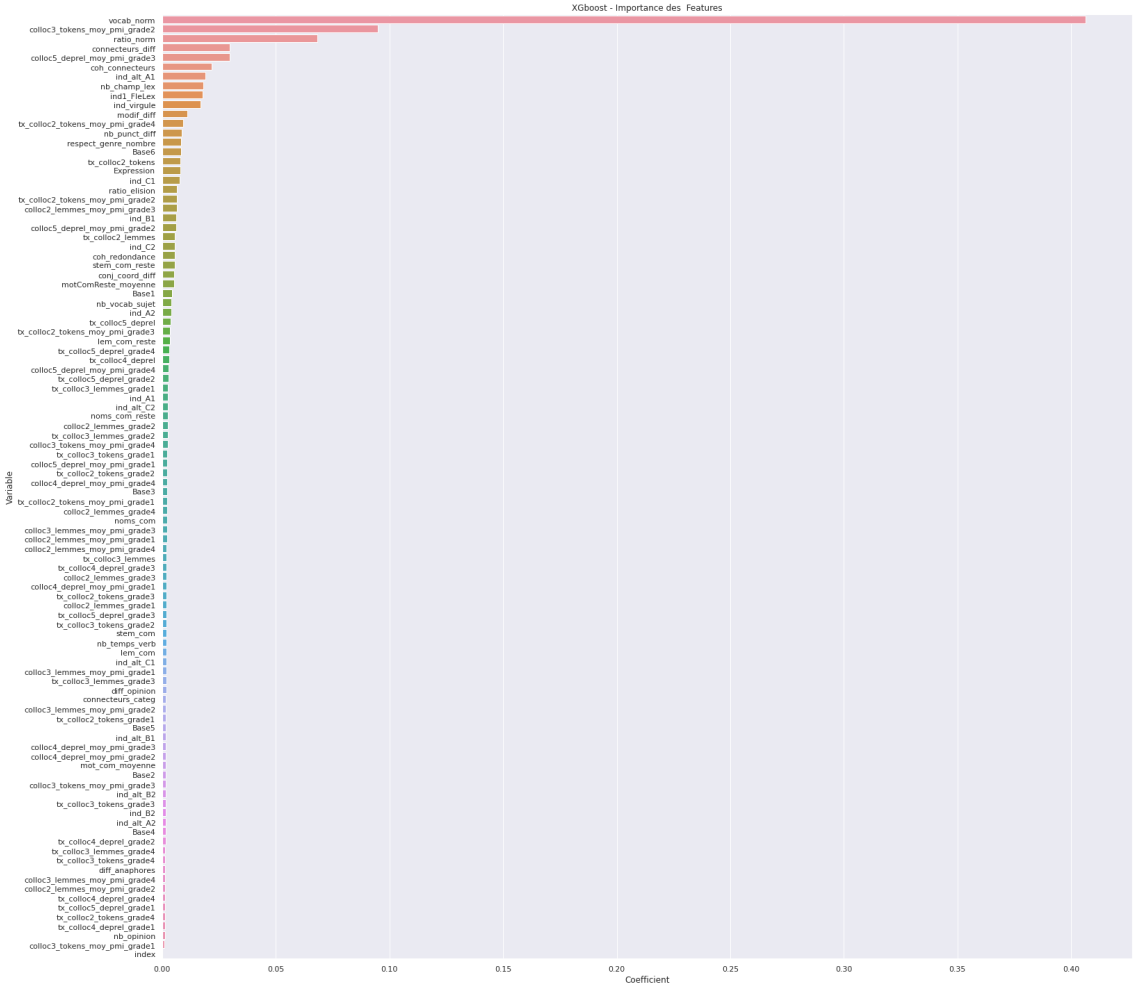

FIGURE A.21 – Importance des caractéristiques - PMI>3- XGBoost-5800. copies

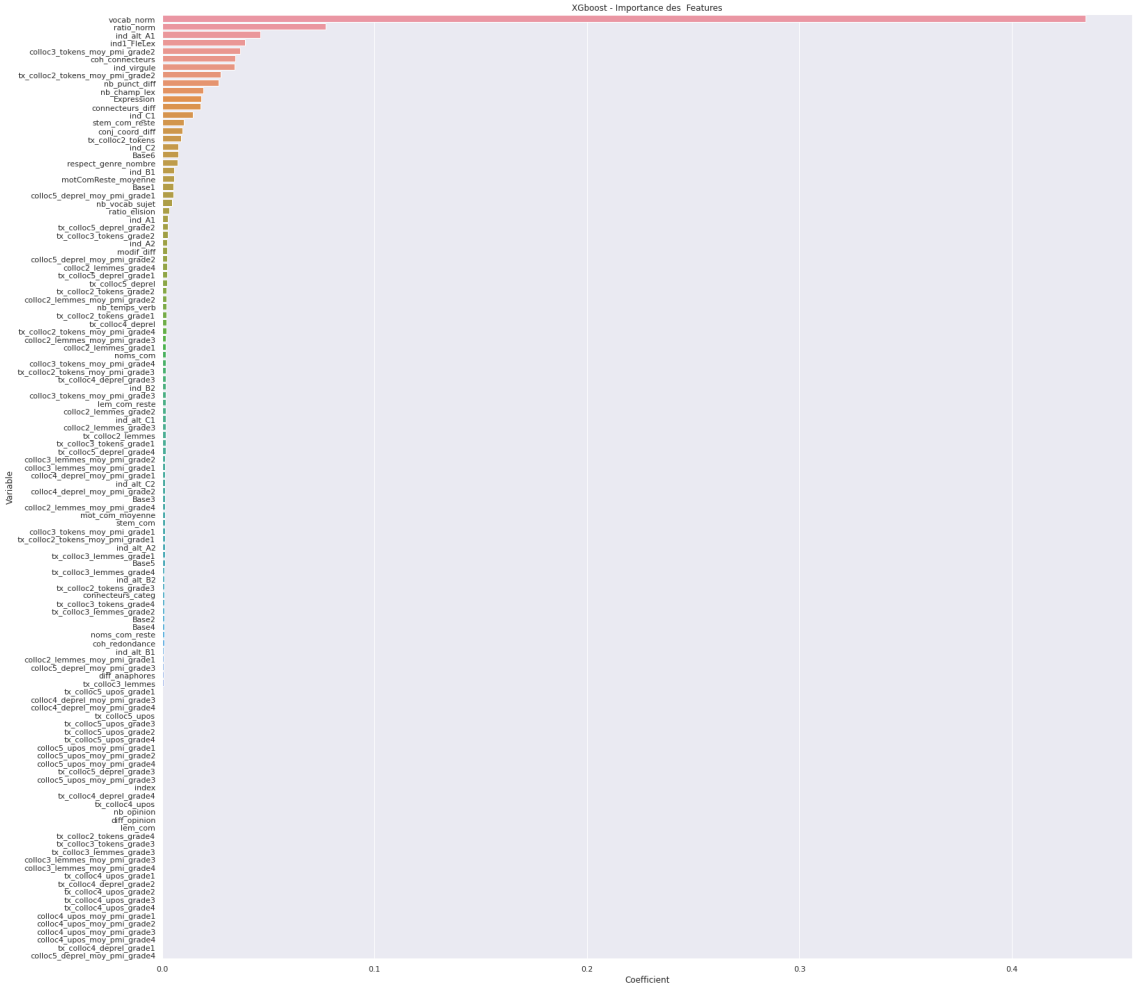

FIGURE A.22 – Importance des caractéristiques - PMI>5- XGBoost-5800. copies

INDEX