Un système de *learning analytics* linguistiques actionnables pour l'apprentissage de l'anglais en contexte universitaire

Rémi Venant[1] Nicolas Ballier[3] Cyriel Mallart[2] Andrew Simpkin[4] Bernardo Stearns[4]
Jen-Yu Li[2] Thomas Gaillat[2]
(1) LIUM, Université Le Mans, 72000 Le Mans, France
(2) LIDILE, Université Rennes 2, Pl. Recteur Henri le Moal,3500 Rennes, France
(3) ALTAE, Université Paris Cité, rue Thomas Mann 75013 Paris, France
(4) Insight, DSI, University of Galway, Galway, Irlande
thomas.gaillat@univ-rennes2.fr

RÉSUMÉ

Cet article présente un système d'apprentissage de l'anglais L2 à l'Université exploité dans le cadre de l'évaluation formative des écrits. Depuis un *Learning Management System* (LMS), celui-ci exploite différentes dimensions linguistiques pour la prédiction de niveau, des analyses multifactiorelles et leur synthèse sous forme de visualisations interactives pour l'enseignant. Co-construit avec ses utilisateurs, ce dernier offre différents niveaux d'explications en termes de niveau de compétences (CECR) et de caractérisation d'unités linguistiques à consolider. Il facilite le suivi des cohortes et des individus, ainsi que la génération de feedback spécifique et la conception de recommandations en classe.

ABSTRACT

A system for actionable *learning analytics* for learners of English in ESP Academic Context

This paper presents a system for English L2 at university. It is used in the context of formative writing assessment. Embedded in a Learning Management System (LMS), it exploits different linguistic dimensions for proficiency prediction, multi-factorial analyses and their synthesis in the form of interactive visualisations for the teacher. Co-constructed with its users, it offers different levels of explanation in terms of proficiency (CEFR) and linguistic units requiring consolidation. It facilitates the monitoring of cohorts and individuals as well as specific feedback generation and instruction design for the class.

MOTS-CLÉS : Learning analytics, Explicabilité, Métriques linguistiques, Tableau de bord pour l'apprentissage, CECR, anglais de spécialité.

KEYWORDS: Learning analytics, Explainability, Textual metrics,Learning Analytic Dashboard, CEFR, English for Specific Purposes.

1 Introduction

Le domaine de l'évaluation de la production écrite en langue étrangère (L2) se tourne de plus en plus vers des approches automatisées et fondées sur les données pour soutenir le feedback formatif et la prise de décision des enseignants. Bien que les systèmes traditionnels d'évaluation automatisée des productions écrites (Automatic Essay Scoring ou AES en anglais) aient principalement servi des objectifs sommatifs, leur application dans des contextes formatifs — où le feedback doit être explicable,

interprétable et actionnable sur le plan pédagogique — reste peu explorée ou fondée sur la simple détection et la correction d'erreurs sans prendre en compte la dimension cognitive de compréhension métalinguistique. Les enseignants bénéficieraient d'applications diagnostics permettant un guidage effectif de leurs étudiants.

Nous proposons un système incluant sur un tableau de bord d'indicateurs linguistiques reposant sur une pipeline de Traitement Automatique des Langues (TAL). Celle-ci intégre des outils offrant des mesures de complexité linguistique alignées sur le Cadre Européen Commun de Référence pour les langues (CECR), garantissant que l'analyse automatisée reflète des construits théoriquement fondés de la compétence en L2. Des modèles prédictifs permettent une classification des productions écrites, et éclairent les traits linguistiques sous-jacents aux prédictions, qu'ils soient positifs ou négatifs.

Au delà de la question de la performance des prédictions des modèles, ce type de système pose aussi la question de leur utilité. La question centrale de cette recherche est la suivante : Dans quelle mesure un tableau de bord d'analytics d'apprentissage, fondé sur une pipeline de TAL et des métriques de complexité linguistique, peut-il fournir des informations valides et actionnables pour les enseignants dans l'évaluation de l'écrit en L2 ?

Contrairement aux systèmes AES traditionnels, souvent perçus comme des boîtes noires, dont seules les performances sont publiées, cette étude adopte un cadre de validité basé sur l'argumentation (Kane, 2013) pour évaluer l'inférence d'utilité (Chapelle & Lee, 2021)— c'est-à-dire le degré auquel les résultats du tableau de bord sont utiles, interprétables et actionnables pour les enseignants. Nous n'abordons que succintement l'évaluation des performances de prédictions dans la mesure ou celle-ci fait l'objet d'un article distinct (en cours de soumission). Neanmoins la section 5 expose brièvement certains résultats.

La section 2 propose un apport théorique sur la notion d'utilité, les mesures de complexité linguistique et les tableaux de bord (TBA). Le dispositif est présenté en section 3 ; son TBA en section 4, son évaluation par les utilisateurs en section 5. La discussion en section 6 précède la conclusion.

2 Cadre théorique

2.1 La validité dans les sytèmes d'évaluation automatique des écrits en L2

La question de la validité joue un rôle central dans les études portant sur les systèmes de scoring d'écrits en langue seconde (L2). Une évaluation exhaustive implique de prendre en compte plusieurs dimensions reflétant au mieux ce construit qu'est la compétence de production écrite. Cependant la plupart des études sur ce type de système se focalise sur les seules performances de classification des données, ignorant ainsi des questions qui peuvent se poser relativement à l'interprétabilité des scores, leur généralisation et leurs usages.

Parmi les cadres théoriques proposés pour remédier à ces limites, la validation basée sur les arguments (Kane, 2013; Williamson *et al.*, 2012) se distingue, en intégrant des preuves multiples de validité. Ce cadre met en lumière des inférences clés — explication, évaluation, généralisation, extrapolation et utilisation — qui doivent être étayées pour soutenir les affirmations sur la validité des scores d'évaluation. (Chapelle *et al.*, 2015) approfondissent cette approche en soulignant que chaque inférence contribue à structurer un argument de validité, notamment dans les contextes où les scores guident l'enseignement et soutiennent l'apprentissage. Parmi ces inférences, l'inférence d'utilisation

(*utilization inference*) est particulièrement cruciale : elle concerne la manière dont les scores ou les retours générés par le système sont effectivement utilisés par les utilisateurs (enseignants et apprenants). Contrairement aux inférences de généralisation ou d'extrapolation, qui portent sur la fiabilité et la transférabilité des scores, l'inférence d'utilisation évalue si les décisions ou actions basées sur ces scores sont pertinentes, efficaces et adaptées à leur contexte d'usage.

L'inférence d'utilisation exige une analyse approfondie de la façon dont les enseignants et les apprenants intègrent les retours du système de scoring dans leurs pratiques. Par exemple, un système conçu pour un usage formatif doit non seulement fournir des scores précis, mais aussi des retours actionnables qui permettent aux enseignants d'adapter leur pédagogie et aux apprenants d'améliorer leurs compétences. Pourtant, cette inférence est souvent négligée dans les études sur les AES, où l'accent est mis sur la précision des scores plutôt que sur leur utilité pratique.

2.2 Mesures exploitées dans des systèmes d'évaluation automatique en L2

Les systèmes de prédiction de compétence en langue remontent aux années soixante (Page, 1968) mais leur application en langues étrangères (L2) est plus récente. Depuis deux décennies, les méthodes de Traitement Automatique des Langues (TAL) ont permis la conception de systèmes de *scoring* reposant sur des algorithmes d'apprentissage supervisé. À l'inverse des approches récentes fondées sur des *Large Language Models*, la plupart des approches traditionnelles reposent sur des propriétés explicites de la langue. Elles constituent des traits qui reposent sur des mesures fréquentielles d'unités textuelles (Yannakoudakis *et al.*, 2012). Les unités correspondent à des n-grammes de mots, des étiquettes morpho-syntaxiques, des structures syntaxiques de phrases, des relations de dépendance ou encore des ressources lexicales pour l'extraction et le décompte de catégories grammaticales (Yannakoudakis *et al.*, 2012, 2018; Vajjala & Lõo, 2014; Vajjala & Rama, 2018; Pilán *et al.*, 2016).

Au-delà des fréquences d'unités, de nombreux indices reposent sur le cadre conceptuel *Complexity, Accuracy, Fluency* (CAF). La complexité lexicale, syntaxique et cohésive ont été exploités pour la modélisation des niveaux de langue, mettant en relation diversité et sophistication avec niveau de langue (Kyle *et al.*, 2018; Kyle & Crossley, 2015; Lu, 2012). Des indices de complexité grammaticale mesurent différentes unités syntaxiques sous formes de ratios (Kyle, 2016; Lu, 2014; Crossley *et al.*, 2016), avec, par exemple, le nombre de propositions en fonction du nombre de mots. Les indices de cohésion textuelles prennent en compte l'usage de connecteurs ou les répétitions de mots entre phrases ou paragraphes (Crossley & McNamara, 2012; Crossley *et al.*, 2019). Par ailleurs, les erreurs ont également été intégrées dans des modèles de prédiction de niveau (Ballier *et al.*, 2019b).

Ces indices ont aussi été exploités à des fins de visualisation. Dans ce cas, l'objectif est de faciliter l'exploration des caractéristiques linguistiques de textes d'apprenants, comme les progrès effectués sur des points grammaticaux (Rudzewitz *et al.*, 2019). De manière similaire, (Yannakoudakis *et al.*, 2012) ont développé un outil d'analyse de traits linguistiques déterminants d'un niveau donné. (Ballier *et al.*, 2019a) avaient proposé un prototype de visualisation des productions individuelles reposant uniquement sur les mesures de complexité en comparaison du corpus de natifs ICE-GB (Nelson *et al.*, 1998). Le système VizLing (Gaillat *et al.*, 2023) approfondissait cette voie et proposait des visualisations comparant textes d'étudiants avec cohortes de référence de différents niveaux, et ce, en fonction de différents critères de complexité linguistique.

2.3 Explicabilité et tableaux de bord

L'apport de modèles statistiques reposant sur un grand nombre de variables peut être inopérant pédagogiquement sans accompagnement à leur exploitation, dans le contexte de la pratique de l'enseignant. Un premier verrou concerne la compréhension nécessaire du modèle nécessaire à l'enseignant pour une prise de décision (Amro & Borup, 2019), tant comme facteur de confiance que pour permettre son intégration dans sa pédagogie. L'explicabilité, enjeu central en Intelligence Artificielle, notamment en Éducation, vise à fournir des bases théoriques et des méthodes pour faciliter la compréhension des décisions. Définie comme le degré auquel un humain peut comprendre la cause d'une décision (Miller, 2019), elle est explorée à travers diverses approches complémentaires. Ces approches se distinguent par des critères tels que la complexité algorithmique ou l'universalité (application à tout modèle ou dépendance à des propriétés spécifiques). Cependant, le critère principal semble être l'orientation de ces explications, globales ou locales (Molnar, 2020). Les premières tendent à décrire le fonctionnement du modèle dans son ensemble, tandis que les secondes se concentrent sur l'interprétation des décisions individuelles du modèle (Rachha & Seyam, 2023). Elles s'appuient sur des comparaisons contrastives (Kim *et al.*, 2016) pour analyser, par exemple, pourquoi deux étudiants spécifiques ont été évalués différemment, ou sur l'étude des changements minimaux nécessaires sur une instance pour modifier sa prédiction (Guidotti, 2024).

Dans le contexte des TBA appliqués à l'apprentissage d'une langue étrangère, la plupart des approches sont centrées sur les apprenants (Rudzewitz *et al.*, 2019; Attali & Burstein, 2006). Elles adoptent un angle d'analyse souvent fondé sur les erreurs de langue, omettant ainsi les caractéristiques positives potentiellement présentes. Nous proposons un TBA conçu pour les enseignants de langue étrangère et reposant sur des mesures linguistiques objectives incluant caractéristiques négatives et positives des productions.

3 Le dispositif de *learning analytics*

3.1 Contexte du projet

Pour des cours en anglais de spécialité à l'université, les enseignants sont confrontés à l'hétérogénéité des parcours antérieurs, et doivent traiter des problématiques générales (grammaire, orthographe) et spécialisées liés aux domaines d'enseignement comme la terminologie (maîtrise du vocabulaire spécialisé). Le projet *Analytics for Language Learning* (A4LL) a pour ambition d'inciter les étudiants à la rédaction pour développer leur compétences écrites, *a fortiori* à l'heure de l'IA générative, qui peut détourner d'une pratique régulière de l'écrit et poser problème pour les évaluations sommatives. Pour cela, A4LL a pour ambition de proposer un système de collecte des écrits étudiants, d'évaluation et d'analyse exploratoire en quasi temps réel à l'enseignant pour l'assister dans son évaluation, dans les retours qu'il pourra faire à l'étudiant et les actions pédagogiques qu'il pourra mener [1].

Dans son ensemble, le système illustré en Figure 1 s'apparente à une pipeline de traitement en quatre blocs : 1) une application *Learning Tools Interoperability* (LTI) [2] intégrable dans un système de

1. Il est à noter que le dispositif comprend une collecte, connue des apprenants par un formulaire de consentement éclairé, des traces numériques clavier susceptibles d'être exploitées pour détecter des textes purement recopiés cf (Velentzas *et al.*, 2024).

2. https://www.1edtech.org/standards/lti

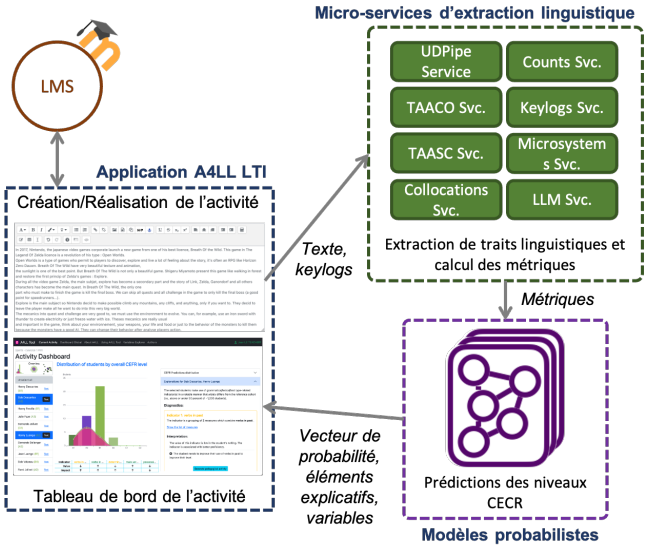

FIGURE 1 – Vue d'ensemble du système Analytics for Language Learning (A4LL)

gestion de l'apprentissage (LMS), pour spécifier et réaliser une activité de rédaction (par exemple, rédiger un texte d'une taille minimal et dans un délai imparti), 2) différentes mesures linguistiques calculées sur la base de ce texte, 3) un modèle prédictif général des niveaux CECR et des sous-modèles par dimension linguistique et 4) un TBA, intégré dans l'application LTI. Dans cet article, nous nous focalisons sur la restitution à l'enseignant, à travers le TBA, et sur la présentation et l'exploration des mesures.

La technicité de l'ensemble du dispositif, et la conception des modèles prédictifs dépassent le cadre de cet article. Cependant, au regard du rôle central des modèles prédictifs, il est important de détailler brièvement la méthode TAL utilisée. Il s'agit d'une approche par apprentissage supervisé. Nous avons utilisé deux corpus d'apprenants. Le corpus EFCAMDAT (Shatz, 2020), comprend 723 282 productions écrites en ligne d'apprenants d'anglais L2 issues de tâches variés, et le corpus CELVA.Sp (Mallart *et al.*, 2023), constitué de 1746 textes rédigés par des étudiants universitaires français apprenant l'anglais de Spécialité (ASP). Pour l'apprentissage machine, un sous-ensemble de 100 000 textes a été extrait aléatoirement du EFCAMDAT en conservant les proportions des niveaux du CECR, puis divisé en jeux d'entraînement (80 %) et de test (20 %). De même le CELVA.Sp à fait l'objet d'une division 80-20. Afin de traiter l'adaptation au domaine cible, le jeu d'entraînement CELVA.Sp a été combiné avec des textes similaires du jeu d'entraînement EFCAMDAT à l'aide d'une analyse en composantes principales (ACP) et d'un appariement par score de propension (PSM) pour créer un jeu de données hybride d'entraînement final équilibré. Les jeux de tests n'ont pas été mélangés afin de garantir l'interpretabilité des résultats concernant l'adaptabilité au domaine. Le modèle Elastic Net, une régression logistique multinomiale avec régularisation L1 et L2, a été utilisé pour la classification CECR générique. Pour une analyse fine par dimension du CECR, des modèles Random Forest ont été entraînés sur les données de chaque dimension du même jeu d'entraînement final. Les métriques d'évaluation incluaient la précision équilibrée (*balanced accuracy*), la précision, le rappel, le score F1 ainsi que des courbes de calibrage pour évaluer l'alignement entre les probabilités prédites et les niveaux CECR observés.

3.2 Présentation synthétique des mesures

Le prototype de notre dispositif repose sur un ensemble de mesures textométriques permettant des comparaisons entre les textes. Une chaîne de traitement automatique inclut des outils (Straka *et al.*, 2016; Qi *et al.*, 2020) d'annotation morphosyntaxique en dépendance (et l'extraction de traits correspondants) en suivant le schéma de *Universal Dependencies* (de Marneffe *et al.*, 2021). D'autres outils exploitent ces annotations pour produire des mesures au niveau du texte relatives à différents domaines linguistiques. Du point de vue grammatical, les outils s'appuient sur l'axe syntagmatique pour des calculs de complexité syntaxique et de quelques patrons d'erreurs. D'autres outils s'appuient sur l'axe paradigmatique pour des calculs relatifs aux alternances potentielles entre formes de même fonction (Mallart *et al.*, 2025). Du point de vue lexical, la dimension phraséologique est prise en compte par des mesures collocationnelles et de similarité en fonction de types de textes spécifiques d'un corpus de référence (Davies, 2009). Du point de vue cohésif, les outils appréhendent la notion de répétition inter-phrase et inter-paragraphe ainsi que celle des connecteurs logiques. Du point de vue comportemental, les traces claviers permettent de mesurer les saisies continues (*burst*) et les temps de pause, et ce en fonction de catégories grammaticales données (Al Sawar *et al.*, 2025). Au total, le système repose sur 591 mesures. Le Tableau 2 présente en annexe un extrait illustratif comprenant description, type et outil de quelques mesures. Si les mesures permettent de capturer une partie de la complexité de la L2, le tableau montre que leurs descriptions restent difficiles à interpréter pour des enseignants. Or, afin de favoriser la prise de décision, il est essentiel de s'assurer de l'interprétabilité des mesures. Nous avons donc conçu une taxonomie des mesures mettant en correspondance les mesures avec des notions conceptuelles propres au métier d'enseignant de langues étrangères.

3.3 Taxonomie des mesures

La pipeline de TAL repose sur le cadre conceptuel CAF qui offre une approche holistique de la compétence de production en langue (Housen, 2014). Il repose sur des mesures permettant de modéliser les strates d'interlangue (Wolfe-Quintero *et al.*, 1998; Norris & Ortega, 2009; Plonsky & Gonulal, 2015), regroupées par familles telles que complexité et exactitude lexicales, grammaticales ou cohésives. Bien qu'utiles pour identifier les facteurs significatifs des niveaux de langue, ces mesures souffrent cependant d'un manque d'explicabilité du fait de leur compositionnalité (Biber *et al.*, 2020) et de leur décorrélation des marqueurs langagiers enseignés en classe. Cette compositionnalité efface le lien avec les marqueurs linguistiques sous-jacents, tels que le génitif, le prétérit ou le passif en anglais, pourtant essentiels dans l'enseignement des langues étrangères (Norris *et al.*, 2015). Il existe donc un hiatus entre la significativité des mesures et leur explicabilité. Pour répondre à ce problème, nous proposons une taxonomie des mesures qui permet de relier les mesures de complexité linguistique à des unités linguistiques interprétables et alignées avec des paramètres de description du CECR (comme la "diversité du vocabulaire" ou la "compétence morpho-syntaxique"), facilitant ainsi leur interprétation par les enseignants.

Cette taxonomie répond à deux besoins qui correspondent à deux paradigmes conceptuels qu'utilisent les enseignants. Le premier correspond aux catégories traditionnelles de l'analyse grammaticale, lexicale et discursive. Les enseignants sont formés à manier ces notions dans le déroulé de leurs cours. Ils sont en mesure d'expliquer le fonctionnement de la *voix passive* par exemple. Le second paradigme conceptuel, notamment dans les tâches d'évaluation écrite, provient du Cadre Européen Commun de Référence (CECR) (Council of Europe, 2001) en langue. Des paramètres guident les évaluateurs dans leurs tâches d'évaluation. Ils sont exprimés en termes de fonctions de communication de type : "Peut

rédiger des textes détaillés officiels ou pas sur une gamme étendue de sujets relatifs à son domaine d'intérêt ...". Dans les deux cas, les mesures exploitées dans notre dispositif restent inappropriées du point de vue de leur actionnabilité (*utility* en anglais) du fait de leur détachement apparent avec les deux paradigmes des enseignants de langue. Nous avons donc élaboré une correspondance entre les mesures et les deux paradigmes conceptuels. Pour le premier, l'objectif est de catégoriser chaque mesure en fonction de la combinaison logique des notions lexico-grammaticales ou discursive concernées (Cf. Tableau en annexe 3). La mesure *nombre de collocations* est, par exemple, caractérisée par la combinaison d'unités linguistiques (UL) *collocations include verb with noun*. Pour le second paradigme, les mesures sont caractérisées en fonction des paramètres CECR (Cf. Tableau 4) de l'expression écrite (Council of Europe, 2001, p.184) principalement concernés. Par exemple, le nombre de collocations relève de la catégorie *étendue du vocabulaire* (*vocabulary range*).

L'évaluation de la taxonomie a porté sur deux types d'appariement : l'association des mesures aux unités linguistiques (UL) et l'alignement des mesures avec les paramètres du CEFR. Pour le premier type, les résultats montrent une bonne concordance entre les annotateurs, avec un kappa de 0,793 pour l'accord inter-annotateurs, et des valeurs de 0,82 et 0,69 lors de la comparaison avec la taxonomie initiale. Seules 5 mesures sur un échantillon de 40 ont révélé des divergences, principalement dues à des ambiguïtés entre les étiquettes (ex. : les subordonnées relatives adnominales confondues avec les subordonnées générales). En revanche, pour l'appariement avec les paramètres du CEFR, l'accord inter-annotateurs est modéré (kappa = 0,44), avec des accords individuels de 0,36 et 0,27 par rapport à la taxonomie. Seules 13 mesures sur 40 ont été classées de manière identique, les désaccords étant souvent liés à des confusions entre portée morpho-syntaxique, précision grammaticale ou maîtrise du vocabulaire.

4 Le tableau de bord (TBA)

4.1 Vue générale du tableau de bord

Le TBA actuel (Figure 2) est contextuel à une activité de rédaction de texte particulière. Cette photographie instantanée des réalisations des apprenants permet de situer la production des apprenants sur l'échelle CECR et met en relation des phénomènes de macro-analyse telle que l'analyse de la cohésion et de micro-analyse telle que l'adéquation aux normes orthographiques. Ce TBA propose une première approche exploratoire descendante de la classe vers l'individu, puis permet un seconde approche centrée sur les concepts linguistiques mobilisés (par axe et dimensions).

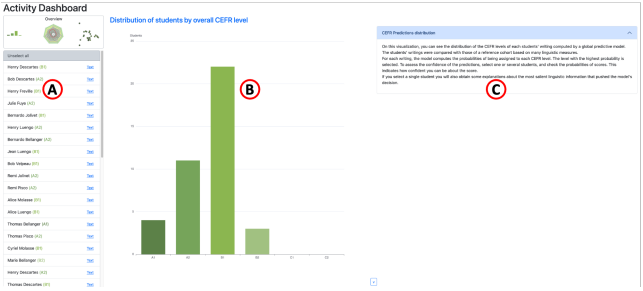

FIGURE 2 – Le TBA (vue d'ensemble)

Le TBA se divise en 3 colonnes : à gauche (**A** sur la Figure 2) est présentée la liste des apprenants, qui permet leur sélection et la consultation éventuelle de leur texte ; les visualisations sont affichées au centre (**B**). Un panneau escamotable à droite (**C**) présente les explications accompagnant ces visualisations. Le système offre trois visualisations : 1) l'histogramme de distribution du groupe classe selon les niveaux CECR, 2) un diagramme en radar des niveaux CECR estimés des apprenants par dimension selon l'axe d'analyse choisi (Compétences CECR ou domaines linguistiques) et 3) une représentation géométrique des étudiants selon leur similarité de modèle, selon l'axe ou la dimension d'analyse choisie.

Un moteur explicatif accompagne ces visualisations pour fournir des explications globales et locales. L'ensemble a été conçu pour inciter à la navigation descendante via des boutons de contrôle du défilement. Lors de la navigation, un rappel des 3 visualisations (présenté dans la colonne de gauche) permet à tout moment de revenir sur l'une d'entre elles.

4.2 Distribution globale du groupe

À l'échelle du groupe, cet histogramme reflète la répartition des niveaux CECR estimés par le modèle global. Lorsqu'un ou plusieurs étudiants sont sélectionnés, leurs courbes de densité de probabilité apparaissent en surimpression (Figure 6), servant d'indicateurs de confiance du modèle. Par exemple, pour deux étudiants de niveau B1, l'un (**A** sur la figure) présente une courbe étroite autour de son niveau, traduisant une décision plus certaine, tandis que pour l'autre (**B**), la courbe est plus large et s'étend vers B2, reflétant une hésitation du modèle entre B1 et B2. Un clic sur une barre permet de sélectionner tous les étudiants d'un même niveau, facilitant l'identification de caractéristiques communes dans les explications et visualisations subséquentes.

4.3 Niveaux CECR par dimensions d'analyse

Comme indiqué précédemment (Cf. Section 3.3), les dimensions des axes de compétences CECR ou des domaines linguistiques ne sont pas représentées de manière égale en termes de mesures, certaines étant potentiellement reléguées à l'arrière-plan des explications du modèle global. Toutefois, il serait didactiquement erroné de conclure à leur manque d'importance ; situer les apprenants sur ces dimensions est donc pertinent. Des modèles spécifiques ont été entraînés (avec la même méthode que celle du modèle global) pour prédire le niveau des apprenants pour chaque dimension des deux axes d'analyse. Des classificateurs ordinaux type *Random Forest*, en suivant (Frank & Hall, 2001), ont été privilégiés pour éviter des probabilités incohérentes (ex. : 40% pour A1 et 60% pour C1), tout en identifiant précisément les mesures importantes associées à chaque niveau CECR. Seuls les modèles atteignant une précision moyenne supérieure à 0,4 (le seuil de décision aléatoire sur 6 classes étant de 0,16) ont été retenus.

Tous les axes étant exprimés sur la même échelle (niveaux CECR), un diagramme radar a été choisi pour la visualisation (Figure 3), outil fréquent en *learning analytics* (Kaczynski *et al.*, 2008) et choisi par les enseignantes des précédents groupes de co-conception. Un bouton de sélection (**A** sur la figure) permet de choisir l'axe d'analyse. Le diagramme radar affiche en rouge les moyennes du groupe classe. Lorsqu'aucun étudiant n'est sélectionné, tous apparaissent sous forme de fines lignes grises, offrant une vue de la variété des profils. Si un ou plusieurs étudiants sont sélectionnés, leurs lignes s'affichent en violet. Un survol de la souris détaille les niveaux estimés des dimensions (**B**).

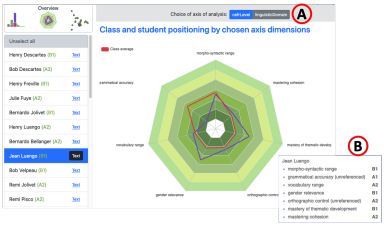

FIGURE 3 – Niveaux CECR par dimensions d'un axe d'analyse

Cette visualisation permet ainsi d'identifier les disparités entre étudiants. Ainsi, dans la Figure 3, l'étudiant sélectionné d'un niveau global B1, présente certaines compétences du CECR en dessous de ce niveau, comme pour la précision grammaticale (A1), également en dessous de la moyenne de la classe. Accompagné des explications locales, l'enseignant peut alors personnaliser ses recommandations.

4.4 Moteur générique d'explication

À chaque visualisation, une explication générale présente brièvement les indicateurs prédicteurs de niveau, la manière dont ces derniers ont été obtenus et le principe général d'interprétation (ex. Figures 2 et 3). Dès lors qu'un ou plusieurs étudiants sont sélectionnés, un moteur générique (i.e. non lié à un type de modèle) calcule leurs explications locales. Le principe général est de s'appuyer sur la contribution des mesures aux modèles statistiques, prise en considération lorsqu'elles s'écartent significativement des valeurs rencontrées lors de l'entraînement (respectivement en dessous du 1er ou au dessus du 9ème décile. La mesure de l'importance des mesures dépend du modèle, comme son effet (positif ou négatif) sur la décision. Le moteur sélectionne les mesures dont la valeur de l'étudiant est significativement forte ou faible et attribue à chacune d'entre elle une "direction interprétative" :

— Influence positive, valeur élevée : pratique "correcte" ;
— Influence négative, valeur faible : pratique "correcte" ;
— Influence positive, valeur faible : pratique "à augmenter" ;
— Influence négative, valeur élevée : pratique "à augmenter".

Pour dépasser la difficulté de compréhension d'une mesure isolée et tendre vers une potentielle actionnabilité, le moteur s'appuie sur la structure hiérarchique des unités linguistiques (UL) auxquelles sont rattachées les mesures (Cf. 3.3). Hypothétiquement, plus le niveau d'abstraction de l'UL est faible, plus celle-ci est précise avec un potentiel d'actionnabilité élevé. Toutefois, plus une UL est partagée par plusieurs mesures de même direction explicative, plus son importance est également élevée. L'objectif est donc de trouver le compromis entre faible niveau d'abstraction d'UL et importance explicative. Un algorithme de regroupement successif est appliqué pour former des ensembles de mesures de même direction interprétative et d'UL commune de niveau d'abstraction de plus en plus haut. Lorsque plusieurs étudiants sont sélectionnés, une liste commune est établie en opérant l'intersection des listes de mesures de chaque étudiant, tout en additionnant leurs poids. Les poids sont ensuite ajustés pour pénaliser les séquences d'unités plus courtes (moins interprétables pour les utilisateurs) tout en favorisant les agrégats contenant davantage de variables

Le nombre maximal de groupes à retenir est paramétrique, défini en fonction de la place disponible pour afficher les explications, mais également de l'estimation de la charge cognitive requise pour

l'interprétation de plusieurs UL. Les explications sont retranscrites dans un tableau synthétique (**A** sur la Figure 4), mentionnant leur direction interprétative, et détaillées à droite (**B**) avec le nombre de mesures associées, la possibilité de les lister et l'interprétation proposée (**C**). Enfin, un bouton (**D**) permet de générer sur la base de cette UL une activité pédagogique sous la forme d'un prompt à copier/coller dans une IA générative. Cette fonctionnalité a pour ambition de faciliter l'actionnabilité. Une dernière visualisation (Figure 7, en annexe) projette les apprenants sur un plan en 2D et permet d'identifier des profils similaires ou distants, la recherche de traits caractéristiques d'un même niveau ou la composition de groupes.

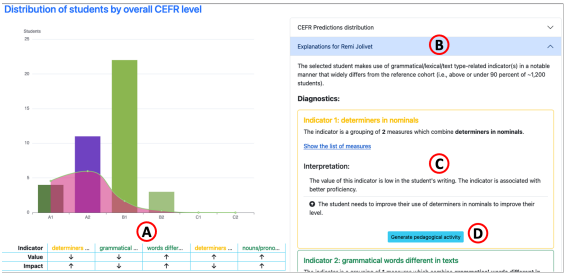

FIGURE 4 – Explications locales pour un étudiants pour son niveau global

5 Résultats

5.1 Performances de classification

En préliminaire, il est à noter que les modèles statistiques sur lesquels reposent les visualisations ont déjà fait l'objet d'évaluation. Leurs résultats sont en cours de publication. On peut toutefois indiquer que le modèle global de CECR obtient une précision globale (balanced accuracy) de 82.9% sur l'échantillon test du corpus EFCAMDAT (Shatz, 2020). A titre de comparaison d'autres systèmes de *scoring* fondés sur des approches d'apprentissage supervisé offrent des résultats similaires en anglais allant de 70% (Ballier *et al.*, 2019b) à 83.9% (Caines & Buttery, 2020). Ce modèle ayant été entraîné sur l'échantillon hybride EFCAMDAT/CELVA.Sp corpus (Cf. 3, nous avons aussi évalué le modèle sur l'échantillon test du CELVA.Sp représentatif des futurs usagers du logiciel. La précision globale obtenue est 63% sur 5 classes du CECR (les données de niveaux C2 ont été ignorées car insuffisantes). Les modèles par domaine linguistique ont aussi été évalués, donnant des précisions globales de l'ordre de 40 à 50%.

5.2 Evaluation qualititive du TBA

Le TBA vise à offrir aux enseignants une exploitation à la fois des prédictions du modèle et des explications associées, fondées sur les taxonomies présentées précédemment. Un *focus groupe* composé de 10 enseignantes d'anglais des universités Rennes 2 et de Rennes a évalué le TBA lors de trois sessions de 3 heures. Une démarche de *Design-Based Research*, reposant sur la méthode PADDLE, adaptée pour ce projet, a été retenue.

Lors de la première session, le projet et les mesures ont fait l'objet de premiers retours qualitatifs à partir de maquettes papier du TBA (ex. Figure 5). La plupart des enseignantes ont montré une certaine difficulté à se projeter dans un système d'analytiques qui ne soit pas centré sur les fautes, malgré la présentation initiale et les taxonomies de mesures proposées. Sur les trois TBA proposés, les enseignantes ont fait remonter les objectifs principaux suivants : identifier le niveau de l'étudiant, identifier des éléments de *feedback* positifs, permettre la montée en compétence selon les faiblesses identifiées en langue (fluidité, structure, aisance), permettre une remédiation automatique par la génération d'exercices et permettre la création d'ateliers sur mesure avec des ressources et des activités personnalisées.

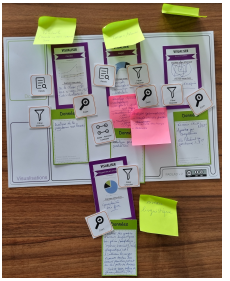

FIGURE 5 – Exemple de maquette de TBA

Il est à noter que les trois groupes ont identifié un scénario original de "parcours utilisateur". Cette navigation permet l'exploration de données, soit temporelles, soit du groupe vers l'individu, soit les deux, ce qui n'est pas formalisé par la méthode PADDLE. Les enseignantes ont souhaité pouvoir débuter sur une vue d'ensemble de leur classe ou leur groupe, puis approfondir leur diagnostic. Cet élément était exprimé à travers la manière d'utiliser leur TBA et non pas seulement par l'agencement et le choix des visualisation.

A partir de cette première session, une première maquette informatique a été développée proche des visualisations principales présentées en section 4. Celle-ci a été présentée lors d'un second *focus group*. Aprés une présentation synthétique de l'outil, les enseignantes devaient explorer et manipuler le TBA proposé et répondre à un questionnaire critique portant sur (**i**) les scénarios d'utilisation qu'elles avaient pu imaginer, (**ii**) la pertinence des informations fournies, (**iii**) les visualisations proposées et (**iv**) l'aisance de navigation du TBA. Les scénarios étaient cohérents avec ceux que nous avions proposés : le suivi global de la classe, l'établissement individuel du niveau, l'établissement des forces et faiblesses d'un étudiant et son positionnement par rapport à la classe. Deux axes exploratoires qui convenaient aux enseignantes ont été retenus : un sur les compétences de CECR et un second sur les dimensions lexico-grammaticales. Parmi les critiques négatives principales, ont été soulignés le manque d'explication locale (et, sans elle, la difficulté d'imaginer une action en classe d'après la lecture du TBA), la complexité de la troisième visualisation et enfin le choix des couleurs, trop orienté (une échelle catégorielle du rouge au vert par niveau avait été proposée).

5.3 Evaluation de l'expérience utilisateur

En prenant en compte l'ensemble des remarques de la deuxième session, et avec l'intégration des travaux parallèles sur les modèles prédictifs, nous avons expérimenté une première version exploitable

du TBA (présentée en section 4). Le *focus group* réduit à huit utilisatrices a permis de tester l'interprétabilité de ces recommandations, soit de vérifier la pertinence et la cohérence des commentaires métalinguistiques produits. L'objectif était de comprendre la capacité des enseignants à prendre des décisions sur la base de ces visualisations. Du point de vue de l'évaluation de l'inférence d'utilité, celle-ci se rapproche de la notion d'expérience utilisateur, qui se définit comme un ensemble de critères qualitatifs distincts les uns des autres. Ceux-ci qui incluent l'efficacité, le contrôle, l'aptitude à l'apprentissage ou encore des critères hédoniques (Hassenzahl, 2001) tels que la stimulation our l'esthétique. L'outil UEQ (User Experience Questionnaire) (Schrepp *et al.*, 2014) opérationalise cette définition à partir d'une liste de dimensions sémantiques qualifiant l'utilisation d'une inferface graphique. Les utilisateurs proposent une note entre deux termes sémantiquement opposés. Les valeurs moyennes de ces notes permettent de discerner des pôles d'attractions qualitatifs comme l'efficacité ou la clarté. Les résultats obtenus (Cf. Tableau 1) permettent d'évaluer les retours d'utilisateurs en fonction de différentes mesures. Les valeurs sont comprises sur une échelle théorique [-3 ;+3], sachant qu'en pratique elles ne dépassent pas [-2 ;+2]. Les auteurs indiquent que des valeurs [-0.8 and 0.8] représentent une évaluation neutre et que des valeurs > 0,8 représentent une évaluation positive. (L'inverse étant vrai pour une évaluation négative). Les résultats révèlent une expérience utilisateur positive pour ce qui concerne les dimensions d'attractivité, d'efficacité, de stimulation et d'originalité. Cependant, la notion de clarté (*perspicuity*) semble poser problème, renforçant ainsi les retours qualitatifs indiquant des difficultés d'interprétation de certains indicateurs et de leur portée en termes de pédagogie.

item	Mean	Std. Dev.	N	Confidence	Confidence interval	
Attractivité	1.083	0.519	8	0.360	0.723	1.443
Clarté	0.344	1.395	8	0.966	-0.623	1.310
Efficacité	1.281	0.687	8	0.476	0.805	1.757
Fiabilité	0.969	0.687	8	0.476	0.493	1.445
Stimulation	1.438	0.609	8	0.422	1.016	1.859
Originalité	1.000	0.756	8	0.524	0.476	1.524

TABLE 1 – Retours des 8 experts selon le questionnaire d'utilisateurs UEQ, CI (p=0.05) par item

6 Discussion et perspectives

Les résultats préliminaires concernant l'utilité du tableau de bord d'analytiques d'apprentissage (TBA) présenté dans cet article révèlent un potentiel prometteur, mais encore fragile. Il convient de noter le nombre restreint de participantes au focus groupe, ce qui impacte la taille des intervals de confiance (Tableau 1 et limite la généralisation. Néanmoins, les retours des enseignantes, recueillis via le questionnaire UEQ, mettent en lumière une expérience utilisateur globalement positive, notamment en termes d'attractivité, d'efficacité et de stimulation. Cependant, la dimension de la clarté des indicateurs linguistiques reste un point d'amélioration majeur, comme en témoignent les scores modérés obtenus pour cette catégorie. Cela souligne la nécessité de renforcer l'interprétabilité des mesures et des visualisations proposées, afin de garantir leur actionnabilité pédagogique. Pour consolider ces résultats, il est prévu d'élargir la base d'utilisateurs afin de recueillir davantage de retours via le questionnaire UEQ, ce qui permettra de distinguer les niveaux d'acceptabilité, d'utilisabilité et d'utilité pédagogique (Tricot *et al.*, 2003), afin d'affiner l'évaluation de l'inférence d'utilité.

La taxonomie axée sur les unités linguistiques a permis l'interfaçage entre les mesures utilisées par les modèles et les explications données aux enseignants. À partir de mesures regroupées en fonction de leur influence positive ou négative sur le niveau, le système permet un tri en fonction du domaine linguistique ou du descripteur CECR ciblé. Il permet alors la sélection personnalisée d'unités linguistiques remarquables propres au filtre employé. L'enseignant peut alors analyser les textes au regard de l'unité linguistique signalée et générer un prompt d'IA pour générer des activités grammaticale personnalisées et adaptées au contexte.

Les développements futurs prévoient l'intégration de fonctionnalités supplémentaires, telles que la génération automatisée d'activités pédagogiques ciblées pour une meilleure adéquation du TBA aux besoins concrets des enseignants. Les traces numériques clavier seront collectées pour la modélisation CECR des écrits L2. Dans un second temps, ces traces permettront de visualiser le déroulement de la construction du texte. Les pauses et *bursts* seront identifiés en fonction des catégories grammaticales sur lesquelles elles adviennent, permettant de mieux apprécier les raisons éventuelles d'hésitations. La collecte de données supplémentaires permettra, à terme, d'affiner les modèles. Le système a vocation à s'inscrire dans un cercle vertueux qui permet d'affiner les modèles statistiques à partir des données collectées. Ces développements futurs visent à transformer ce prototype en un outil robuste et pleinement exploitable en contexte éducatif, tout en s'inscrivant dans le cadre conceptuel de la validité.

7 Conclusion

A l'heure des premières expériences, pas toujours concluantes (Wang & Demszky, 2023), de l'utilisation de chatGPT pour la production de *feedback* pour les écrits d'apprenants en anglais, cette méthode d'analyse des données d'apprenants garantit la traçabilité des décisions prises et inscrit notre système dans l'IA de confiance pour les données éducatives. Même si le système nous paraît utilisable en autonomie, moyennant une formation initiale aux principales notions fondatrices (mesures, plans d'analyse), cette démarche s'inscrit dans une approche collaborative qui intègre l'apport de l'expertise humaine dans l'analyse automatique des données d'apprenants. Une démo du prototype peut être consultée[3] et le code est disponible en ligne[4]. Conçu au départ pour l'anglais, mais également testé pour le suédois et l'espagnol, l'ensemble du dispositif A4LL a vocation a devenir un analytique de l'apprentissage des langues, puisqu'un suivi longitudinal des apprenants est possible, ainsi, à moyen terme, qu'une analyse des traces numériques clavier.

Remerciements

Nous remercions les enseignantes de langues des deux centres de langues des deux universités rennaises (CdL et SCELVA) qui ont participé au *focus group*. Le système A4LL a été financé dans le cadre du projet ANR-22-CE38-0015-01.

3. URL : https://linguisticdataprocessing.huma-num.fr/a4ll-tool/, login : teacher1@mail.com, mdp : teacher1
4. URL : https://gitlab.huma-num.fr/lidile/a4ll_mlpipeline

Références

AL SAWAR A., PACQUETET E., MALLART C., SIMPKIN A. & BALLIER N. (2025). Analyse exploratoire des traces numériques clavier pour la prédiction des niveaux d'apprenants. In F. BECHET, A.-G. CHIFU, K. PINEL-SAUVAGNAT, B. FAVRE, E. MAES & D. NURBAKOVA, Éds., *Actes de l'atelier Traitement de données langagières dynamiques par les outils et méthodes du TAL 2025 (DYN-TAL)*, p. 7–11, Marseille, France : ATALA.

AMRO F. & BORUP J. (2019). Exploring Blended Teacher Roles and Obstacles to Success When Using Personalized Learning Software. *Journal of Online Learning Research*, **5**(3), 229–250. Publisher : Association for the Advancement of Computing in Education ERIC Number : EJ1241760.

ATTALI Y. & BURSTEIN J. (2006). Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment*, **4**(3), 3–29.

BALLIER N., GAILLAT T. & PACQUETET E. (2019a). Prototype de feedback visuel des productions écrites d'apprenants francophones de l'anglais sous Moodle. In *Actes EIAH 2019*, Actes EIAH 2019, Paris, France : University of Paris Sorbonne. HAL : hal-02496651.

BALLIER N., GAILLAT T., SIMPKIN A., STEARNS B., BOUYÉ M. & ZARROUK M. (2019b). A Supervised Learning Model for the Automatic Assessment of Language Levels Based on Learner Errors. In M. SCHEFFEL, J. BROISIN, V. PAMMER-SCHINDLER, A. IOANNOU & J. SCHNEIDER, Éds., *Transforming Learning with Meaningful Technologies*, Lecture Notes in Computer Science, p. 308–320, Switzerland : Springer International Publishing.

BENAMARA F., HATOUT N., MULLER P. & OZDOWSKA S., Éds. (2007). *Actes de TALN 2007 (Traitement automatique des langues naturelles)*, Toulouse. ATALA, IRIT.

BIBER D., GRAY B., STAPLES S. & EGBERT J. (2020). Investigating grammatical complexity in L2 English writing research : Linguistic description versus predictive measurement. *Journal of English for Academic Purposes*, **46**, 100869.

CAINES A. & BUTTERY P. (2020). REPROLANG 2020 : Automatic Proficiency Scoring of Czech, English, German, Italian, and Spanish Learner Essays. In N. CALZOLARI, F. BÉCHET, P. BLACHE, K. CHOUKRI, C. CIERI, T. DECLERCK, S. GOGGI, H. ISAHARA, B. MAEGAARD, J. MARIANI, H. MAZO, A. MORENO, J. ODIJK & S. PIPERIDIS, Éds., *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 5614–5623, Marseille, France : European Language Resources Association.

CHAPELLE C. A., COTOS E. & LEE J. (2015). Validity arguments for diagnostic assessment using automated writing evaluation. *Language Testing*, **32**(3), 385–405. DOI : 10.1177/0265532214565386.

CHAPELLE C. A. & LEE H.-W. (2021). Conceptions of validity. In *The Routledge Handbook of Language Testing*, p. 17–31. London and New York : Routledge, 2 édition. Num Pages : 15.

COUNCIL OF EUROPE . (2001). *Common European framework of reference for languages : Learning, teaching, assessment*. Cambridge University Press.

CROSSLEY S. A., KYLE K. & DASCALU M. (2019). The Tool for the Automatic Analysis of Cohesion 2.0 : Integrating semantic similarity and text overlap. *Behavior Research Methods*, **51**(1), 14–27. DOI : 10.3758/s13428-018-1142-4.

CROSSLEY S. A., KYLE K. & MCNAMARA D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO) : Automatic assessment of local, global, and text cohesion. *Behavior Research Methods*, **48**(4), 1227–1237. DOI : 10.3758/s13428-015-0651-7.

CROSSLEY S. A. & MCNAMARA D. S. (2012). Predicting second language writing proficiency : the roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, **35**(2), 115–135. DOI : 10.1111/j.1467-9817.2010.01449.x.

DAVIES M. (2009). The 385+ million word Corpus of Contemporary American English (1990–2008+) : Design, architecture, and linguistic insights. *International Journal of Corpus Linguistics*, **14**(2), 159–190. DOI : 10.1075/ijcl.14.2.02dav.

DE MARNEFFE M.-C., MANNING C. D., NIVRE J. & ZEMAN D. (2021). Universal Dependencies. *Computational Linguistics*, **47**(2), 255–308. Place : Cambridge, MA Publisher : MIT Press, DOI : 10.1162/coli_a_00402.

DIAS G., Éd. (2015). *Actes de TALN 2015 (Traitement automatique des langues naturelles)*, Caen. ATALA, HULTECH.

FRANK E. & HALL M. (2001). A simple approach to ordinal classification. In *Machine Learning : ECML 2001 : 12th European Conference on Machine Learning Freiburg, Germany, September 5–7, 2001 Proceedings 12*, p. 145–156 : Springer.

GAILLAT T., LAFONTAINE A. & KNEFATI A. (2023). Visualizing Linguistic Complexity and Proficiency in Learner English Writings. *CALICO Journal*, **40**(2), 178–197. DOI : 10.1558/cj.19487.

GUIDOTTI R. (2024). Counterfactual explanations and how to find them : literature review and benchmarking. *Data Mining and Knowledge Discovery*, **38**(5), 2770–2824.

HASSENZAHL M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction*, **13**(4), 479–497. DOI : 10.1207/S15327590IJHC1304_07.

HOUSEN A. (2014). Difficulty and Complexity of Language Features and Second Language Instruction. In *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Inc.

KACZYNSKI D., WOOD L. & HARDING A. (2008). Using radar charts with qualitative evaluation : Techniques to assess change in blended learning. *Active Learning in Higher Education*, **9**(1), 23–41.

KANE M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, **50**(1), 1–73. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/jedm.12000, DOI : 10.1111/jedm.12000.

KIM B., KHANNA R. & KOYEJO O. O. (2016). Examples are not enough, learn to criticize ! criticism for interpretability. *Advances in neural information processing systems*, **29**.

KYLE K. (2016). *Measuring syntactic development in L2 writing : Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication*. Dissertation, Georgia State University, Georgia.

KYLE K., CROSSLEY S. & BERGER C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES) : version 2.0. *Behavior Research Methods*, **50**(3), 1030–1046. DOI : 10.3758/s13428-017-0924-4.

KYLE K. & CROSSLEY S. A. (2015). Automatically Assessing Lexical Sophistication : Indices, Tools, Findings, and Application. *TESOL Quarterly*, **49**(4), 757–786. DOI : 10.1002/tesq.194.

LAIGNELET M. & RIOULT F. (2009). Repérer automatiquement les segments obsolescents à l'aide d'indices sémantiques et discursifs. In A. NAZARENKO & T. POIBEAU, Éds., *Actes de TALN 2009 (Traitement automatique des langues naturelles)*, Senlis : ATALA LIPN.

LANGLAIS P. & PATRY A. (2007). Enrichissement d'un lexique bilingue par analogie. In (Benamara *et al.*, 2007), p. 101–110.

LU X. (2012). The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives. *The Modern Language Journal*, **96**(2), 190–208. DOI : 10.1111/j.1540-4781.2011.01232_1.x.

LU X. (2014). *Computational Methods for Corpus Annotation and Analysis*. Dordrecht : Springer.

MALLART C., SIMPKIN A., BALLIER N., LISSÓN P., VENANT R., STEARNS B., LI J.-Y. & GAILLAT T. (2025). Assessing the validity of syntactic alternations as criterial features of proficiency in L2 writings in English. *Research Methods in Applied Linguistics*, **4**(3), 100238. DOI : 10.1016/j.rmal.2025.100238.

MALLART C., SIMPKIN A., VENANT R., BALLIER N., STEARNS B., LI J. Y. & GAILLAT T. (2023). A new learner language data set for the study of English for Specific Purposes at university level. In *Proceedings of the 4th Conference on Language, Data and Knowledge - LDK 2023*, volume 1, p. 281–287, Vienna, Austria.

MILLER T. (2019). Explanation in artificial intelligence : Insights from the social sciences. *Artificial Intelligence*, **267**, 1–38. DOI : 10.1016/j.artint.2018.07.007.

MOLNAR C. (2020). *Interpretable machine learning*. Lulu. com.

NELSON G., WALLIS S. & AARTS B. (1998). The British Component of the International Corpus of English (ICE-GB) and ICECUP software (CD-ROM).

NORRIS J. M. & ORTEGA L. (2009). Towards an Organic Approach to Investigating CAF in Instructed SLA : The Case of Complexity. *Applied Linguistics*, **30**(4), 555–578. Publisher : Oxford Academic, DOI : 10.1093/applin/amp044.

NORRIS J. M., ROSS S. J. & SCHOONEN R. (2015). Improving Second Language Quantitative Research. *Language Learning*, **65**(S1), 1–8. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/lang.12110, DOI : 10.1111/lang.12110.

PAGE E. B. (1968). The Use of the Computer in Analyzing Student Essays. *International Review of Education*, **14**(2), 210–225.

PILÁN I., VOLODINA E. & ZESCH T. (2016). Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics : Technical Papers*, p. 2101–2111, Osaka, Japan : The COLING 2016 Organizing Committee.

PLONSKY L. & GONULAL T. (2015). Methodological Synthesis in Quantitative L2 Research : A Review of Reviews and a Case Study of Exploratory Factor Analysis. *Language Learning*, **65**(S1), 9–36. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/lang.12111, DOI : 10.1111/lang.12111.

QI P., ZHANG Y., ZHANG Y., BOLTON J. & MANNING C. D. (2020). Stanza : A Python Natural Language Processing Toolkit for Many Human Languages. arXiv :2003.07082 [cs], DOI : 10.48550/arXiv.2003.07082.

RACHHA A. & SEYAM M. (2023). *Explainable AI In Education : Current Trends, Challenges, And Opportunities*. Pages : 239, DOI : 10.1109/SoutheastCon51012.2023.10115140.

RUDZEWITZ B., ZIAI R., NUXOLL F., KUTHY K. D. & MEURERS W. D. (2019). Enhancing a Web-based Language Tutoring System with Learning Analytics. In LUC PAQUETTE & C. ROMERO, Éds., *Joint Proceedings of the Workshops of the 12th International Conference on Educational Data Mining co-located with the 12th International Conference on Educational Data Mining, EDM 2019 Workshops*, volume 2592, p. 1–7, Montréal, Canada : CEUR-WS.

SCHREPP M., HINDERKS A. & THOMASCHEWSKI J. (2014). Applying the User Experience Questionnaire (UEQ) in Different Evaluation Scenarios. In A. MARCUS, Éd., *Design, User Experience, and Usability. Theories, Methods, and Tools for Designing the User Experience*, p. 383–392, Cham : Springer International Publishing. DOI : 10.1007/978-3-319-07668-3_37.

SERETAN V. & WEHRLI E. (2007). Collocation translation based on sentence alignment and parsing. In (Benamara *et al.*, 2007), p. 401–410.

SHATZ I. (2020). Refining and modifying the EFCAMDAT : Lessons from creating a new corpus from an existing large-scale English learner language database. *International Journal of Learner Corpus Research*, **6**(2), 220–236. Publisher : John Benjamins, DOI : 10.1075/ijlcr.20009.sha.

STRAKA M., HAJIČ J. & STRAKOVÁ J. (2016). UDPipe : Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, p. 4290–4297, Portorož, Slovenia : European Language Resources Association (ELRA).

TRICOT A., PLÉGAT-SOUTJIS F., CAMPS J.-F., AMIEL A., LUTZ G. & MORCILLO A. (2003). Utilité, utilisabilité, acceptabilité : interpréter les relations entre trois dimensions de l'évaluation des EIAH. In DESMOULINS, C., MARQUET, P., BOUHINEAU & D., Éds., *Environnements Informatiques pour l'Apprentissage Humain 2003*, p. 391–402, Strasbourg, France : ATIEF ; INRP.

VAJJALA S. & LÕO K. (2014). Automatic CEFR level prediction for Estonian learner text. In E. VOLODINA & L. BORIN, Éds., *Proceedings of the third workshop on NLP for computer-assisted language learning*, p. 113–127, Uppsala, Sweden : LiU Electronic Press.

VAJJALA S. & RAMA T. (2018). Experiments with Universal CEFR Classification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 147–153, New Orleans, Louisiana : Association for Computational Linguistics. DOI : 10.18653/v1/W18-0515.

VELENTZAS G., CAINES A., BORGO R., PACQUETET E., HAMILTON C., ARNOLD T., NICHOLLS D., BUTTERY P., GAILLAT T., BALLIER N. & YANNAKOUDAKIS H. (2024). Logging keystrokes in writing by English learners. In N. CALZOLARI, M.-Y. KAN, V. HOSTE, A. LENCI, S. SAKTI & N. XUE, Éds., *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, p. 10725–10746, Torino, Italia : ELRA and ICCL.

WANG R. & DEMSZKY D. (2023). Is ChatGPT a good teacher coach ? measuring zero-shot performance for scoring and providing actionable insights on classroom instruction. In E. KOCHMAR, J. BURSTEIN, A. HORBACH, R. LAARMANN-QUANTE, N. MADNANI, A. TACK, V. YANEVA, Z. YUAN & T. ZESCH, Éds., *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, p. 626–667, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.bea-1.53.

WILLIAMSON D. M., XI X. & BREYER F. J. (2012). A Framework for Evaluation and Use of Automated Scoring. *Educational Measurement : Issues and Practice*, **31**(1), 2–13. _eprint : https ://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1745-3992.2011.00223.x, DOI : 10.1111/j.1745-3992.2011.00223.x.

WOLFE-QUINTERO K., INAGAKI S. & KIM H.-Y. (1998). *Second language development in writing : measures of fluency, accuracy, & complexity*. Honolulu : Second Language Teaching & Curriculum Center, University of Hawaii at Manoa. OCLC : 40664312.

YANNAKOUDAKIS H., ANDERSEN E., GERANPAYEH A., BRISCOE T. & NICHOLLS D. (2018). Developing an automated writing placement system for ESL learners. *Applied Measurement in Education*, **31**(3). DOI : 10.1080/08957347.2018.1464447.

YANNAKOUDAKIS H., BRISCOE T. & ALEXOPOULOU T. (2012). Automating Second Language Acquisition Research : Integrating Information Visualisation and Machine Learning. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, p. 35–43, Avignon, France : Association for Computational Linguistics.

A Annexes

A.1 Exemples de métriques (mesures) implémentées

Var	description mesure bas niveau	Type	Outil	Domaine linguistique
1	Number of collocations used	Collocations	Collocation_tool	lexical
2	nb backspace sequence longer than 3	Keylogs	Keylogger	behavioural
3	Passive voice without past participle aspect on verb	Syntagmatic microsystem	MSAnalyzer	grammatical
4	Negative logical connectives	Connectives	TAACO	semantic
5	dependents per nominal subject (no pronouns, standard deviation)	Noun Phrase Variety	TAASSC v1.3.8	grammatical
6	Ratio of _ING per text	Universal Dependencies	UD_feat	grammatical
7	disagreement between token probabilities of pairs of models	BERT_EF	user_simulator	semantic

TABLE 2 – Illustration de quelques-unes des 591 mesures du dispositif

A.2 Exemples de catégorisation lexico-grammaticale et discursive des métriques (mesures) implémentées

Var	Desc. mesure bas niveau	Type	Unité première	Lien logique	Unité secondaire	Lien logique 2	Unité tertiaire
1	Number of collocations used	Collocations	collocations	include	verb	with	noun
2	Nb backspace seq shorter than or equal 3 for revision	Keylogs	reversal	in	character	per	revision
3	Duration : ago	Paradigmatic microsystems	temporals	vs	temporals	NA	NA
4	Lda divergence (adjacent paragraphs)	Cohesion	consecutive paragraphs	with	consecutive paragraphs	NA	NA
5	Dependents per direct object (no pronouns)	Complexity	words	as	direct objects	NA	NA
6	Adjectival modifier	UD features	adjectives	NA	NA	NA	NA
7	Probability of actual learner adverb being predicted by learner model	BERT_EF	adverb	NA	NA	NA	NA

TABLE 3 – Extrait illustratif de 14 mesures en fonction de leur catégorisation lexico-grammaticale et discursive

A.3 Exemple de 14 mesures en fonction de leur catégorisation CECR

Var	Description mesure bas niveau	Outil	Compétence langage CECR
1	Number of collocations used	Collocation tool	vocabulary range : diversity
2	Nb of backspace sequences shorter than or equal 3 for revision	Keylogger	accuracy
3	Duration : ago	MSAnalyzer	morpho-syntactic range
4	Lda divergence (adjacent paragraphs)	TAACO	pragmatic competence : thematic development
5	Dependents per direct object (no pronouns)	TAASSC v1.3.8	morpho-syntactic range
6	Adjectival modifier	UD feat extractor	morpho-syntactic range
7	Probability of the actual learner adverb tokens being predicted by the learner model	User simulator	vocabulary range : diversity

TABLE 4 – Extrait (à titre d'exemple) de 14 mesures en fonction de leur catégorisation CECR

A.4 Visualisation globale des niveaux CECR estimés

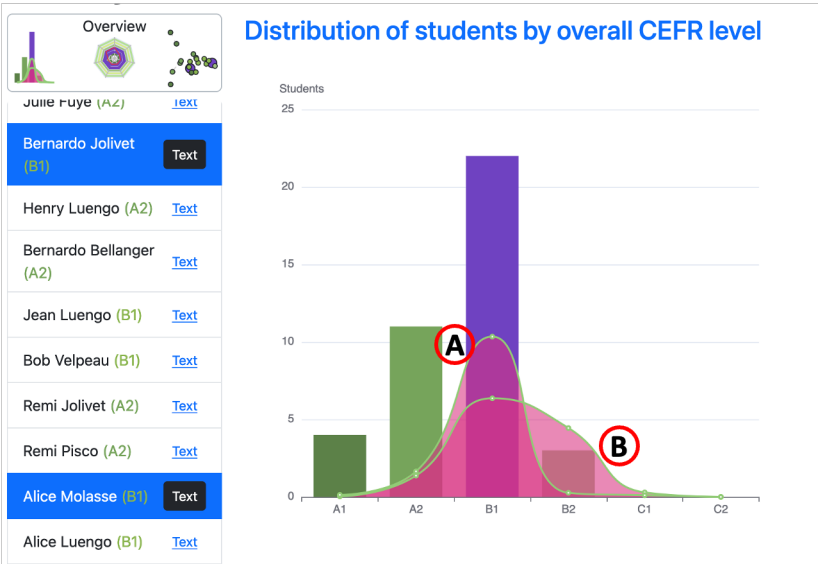

FIGURE 6 – Distribution globale des niveaux CECR estimés

A.5 Comparaison des modèles d'apprenant

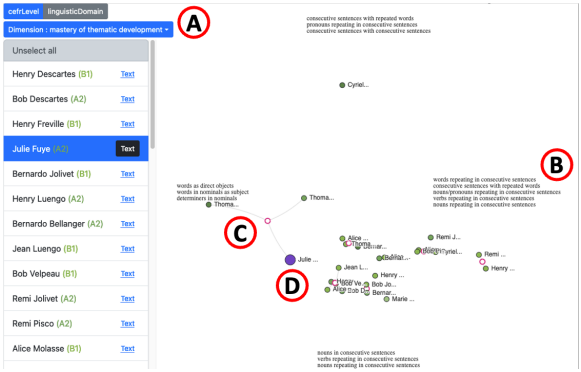

FIGURE 7 – Projection des étudiants dans le plan, via une analyse par composantes principales faite sur l'ensemble des dimensions de l'axe d'analyse, ou sur une dimension particulière (**A** sur la figure). Les UL les plus importantes expliquant la signification d'une position particulière (**B**) sont visualisées sur le plan et dans le panneau des explications. Le lien possible entre étudiants représente un cluster identifié (K-Means utilisé) (**C**). Les étudiants sélectionnés sont colorés en violet (**D**).