

# Vers des *learning analytics* linguistiques actionnables pour l'apprentissage de l'anglais en contexte universitaire

No Author Given

No Institute Given

## Abstract.

Cet article présente la conception d'un tableau de bord d'apprentissage de l'anglais de spécialité à l'Université exploité dans le cadre de l'évaluation formative des écrits. Depuis un LMS, celui-ci exploite différentes dimensions linguistiques pour la prédiction de niveau, des analyses multifactielles 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.

**Mots-clés :** Learning analytics · Explicabilité · Métriques linguistiques · Tableau de bord pour l'apprentissage · CECR · Anglais de spécialité.

## Abstract.

This paper presents the design of a dashboard for English for Specific Purposes (ESP) at university. It is used in the context of formative writing assessment. Embedded in an LMS, it exploits different linguistic dimensions for proficiency prediction, multi-factor 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 in the classroom.

Translated with DeepL.com (free version)

**Keywords:** Learning analytics · Explainability · Textual metrics · Learning Analytic Dashboard · CEFR · English for Specific Purposes.

## 1 Introduction

Le domaine de l'apprentissage des langues étrangères (L2) est saturé d'applications focalisées sur les apprenants pris dans leur dimension individuelle, proposant majoritairement des exercices interactifs auto-corrigés. Ceux-ci sont en général

fondés sur la détection et la correction d'erreurs, et reposent sur l'autonomie des apprenants pour les tâches cognitives de compréhension métalinguistique. Ce mode d'apprentissage autonome ne concerne pas les enseignants, pourtant essentiels pour le nécessaire guidage métalinguistique comprenant des phases explicites avec exercices. Ces processus sont notamment à l'œuvre en production écrite dans le contexte universitaire par les retours des enseignants. Cependant, les temps de correction réduisent la fréquence des productions écrites, renvoyant celles-ci à des évaluations sommatives de fin de semestre focalisées sur l'erreur, élément plus simple à inventorier manuellement. Les diagnostics sont donc trop tardifs et biaisés.

Pour sortir de cette ornière, le diagnostic automatique formatif représente une solution, en particulier pour les textes écrits en L2. Le cadre conceptuel *Complexity, Accuracy, FFluency* (CAF) offre une approche holistique de la compétence de production en langue [21]. Il repose sur des mesures permettant de modéliser les strates d'interlangue [51, 36, 40], 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 [9] 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 prêtérit ou le passif en anglais, pourtant essentiels dans l'enseignement des langues étrangères [37]. Il existe donc un hiatus entre la significativité des mesures et leur explicabilité. Construire des modèles combinant ces deux exigences nous a conduits à trois interrogations :

1. Comment sélectionner de mesures pertinentes statistiquement ?
2. Comment lier les mesures linguistiques à ces traits langagiers dans un cadre unifié ?
3. Comment créer un Tableau de Bord d'Apprentissage (TBA) pour visualiser et expliquer ces traits et mesures dans un objectif d'actionnabilité ?

Cet article répond aux questions 2 et 3, tandis que la question 1 a fait l'objet d'une étude spécifique, actuellement en cours de publication. Seuls quelques résultats succincts sont mentionnés dans la discussion. L'article présente un système d'analytics linguistiques en apprentissage des langues et la manière dont les visualisations répondent à la contrainte d'explicabilité. Les modèles qui le structurent sont fondés sur une taxonomie de caractéristiques linguistiques reliées à des mesures statistiques significatives. Ce système s'adresse aux enseignants de langue étrangère du Supérieur et ambitionne de visualiser les caractéristiques linguistiques d'une production écrite au regard d'une cohorte de référence (groupe-classe). La Section 2 propose un état de l'art des mesures exploitées dans les systèmes d'analyse L2 ainsi que sur la notion d'explicabilité et les TBA. Le dispositif est présenté en Section 3, son TBA en Section 4. La discussion en Section 5, précède la conclusion.

## 2 État de l'art

### 2.1 Mesures exploitées dans des systèmes d'évaluation automatique en L2

Les systèmes de prédiction de compétence en langue remontent aux années soixante [38] 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 [52]. 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 [52, 53, 47, 48, 39].

Au delà des fréquences d'unités, de nombreux indices de 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 [27, 28, 30]. Des indices de complexité grammaticale mesurent différentes unités syntaxiques comme les propositions en fonction du nombre de mots [26, 29, 12]. Les indices de cohésion textuelles prennent en compte l'usage de connecteurs ou les répétitions de mots entre phrases ou paragraphes [13, 11]. Par ailleurs, les erreurs ont aussi été intégrées dans des modèles de prédiction de niveau [3].

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 [44]. De manière similaire, Yannakoudakis et al. ont développé un outil d'analyse de traits linguistiques déterminants d'un niveau donné [52]. [4] avaient proposé un prototype de visualisation des productions individuelles reposant uniquement sur les mesures de complexité en comparaison du corpus de natifs ICE-GB [35]. Le système VizLing [6] 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. Les retours d'expériences auprès d'utilisateurs laissaient entrevoir des difficultés de compréhension des mesures et de leur portées. Un travail de classification des mesures avait été entrepris afin d'élaborer une taxonomie des mesures selon leur portée grammaticale [5].

### 2.2 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 [1], 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 [33], 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 [34]. 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 [42]. Elles s'appuient sur des comparaisons contrastives [25] 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 [19].

La restitution de l'ensemble de ces éléments (décisions des modèles, explications) se fait au travers d'un tableau de bord d'apprentissage (TBA), aujourd'hui présent dans la plupart des formations. Leurs effets sur l'apprentissage sont mitigés, des études montrant des effets bénéfiques, comme par exemple la motivation ou l'autorégulation [32], et d'autres négatifs, comme dans le cas de comparaison entre pairs [23]. Une critique régulière est l'exclusion de l'utilisateur final lors de la conception du TBA [20]. Pour répondre à cette problématique, [18, 22] ont proposé un kit tangible de conception participative de TBA, appelé PADDLE, dont l'objet est de permettre à des petits groupes de co-concevoir un TBA en identifiant son objectif (ex.: planification), son contexte d'utilisation, le choix des données et des visualisations, et d'élaborer plusieurs écrans de ce dernier<sup>1</sup>.

Dans le contexte des TBA appliqués à l'apprentissage d'une langue étrangère, la plupart des approches sont centrées sur les apprenants [44, 2]. 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* [ANONYME]

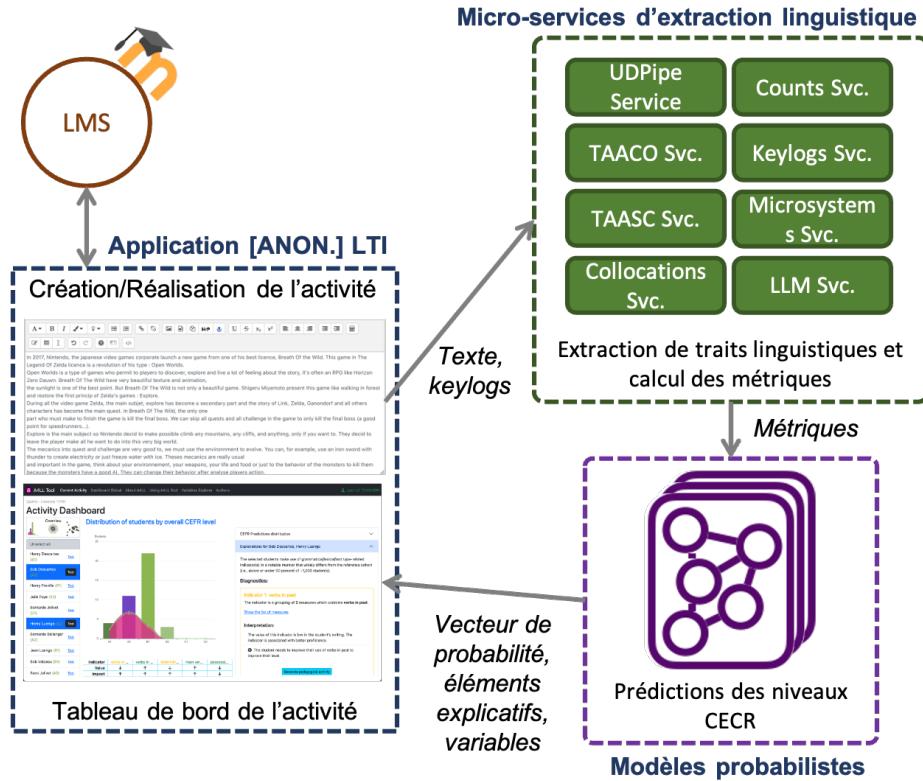
#### 3.1 Contexte du projet

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 [ANONYME] 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

---

<sup>1</sup> [https://padlad.github.io/epaddle/co/3\\_PhasesPADDLE.html](https://padlad.github.io/epaddle/co/3_PhasesPADDLE.html)

peut détourner d'une pratique régulière de l'écrit et poser problèmes pour les évaluations sommatives. Pour cela, [ANONYME] 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<sup>2</sup>.



**Fig. 1.** Vue d'ensemble du système [ANONYME]

Dans son ensemble, le système illustré en figure 1 s'apparente à un pipeline de traitement en quatre blocs :

<sup>2</sup> 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 [49].

- une application LTI<sup>3</sup> intégrable dans un système de 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);
- différentes mesures linguistiques calculées sur la base de ce texte ;
- un modèle prédictif général des niveau CECR et des sous-modèles par dimension linguistique ;
- un TBA, intégré dans l'application LTI.

Dans cette 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 prédictif dépassent le cadre de cet article.

### 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 [46, 41] d'annotation morphosyntaxique en dépendance (et l'extraction de traits correspondants) en suivant le schéma de *Universal Dependencies* [31]. 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 d'exactitude. D'autres outils s'appuient sur l'axe paradigmatic pour des calculs relatifs aux alternances potentielles entre formes de même fonction. 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 [15]. 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. Au total, le système repose sur 591 mesures. Le Tableau 1 présente 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 mettant en correspondance les mesures avec des notions conceptuelles propres au métier d'enseignant de langues étrangères.

### 3.3 Taxonomie des mesures

Cette taxonomie répond à deux besoins qui correspondent à deux paradigmes conceptuels qu'utilisent les enseignants. Le premier correspond aux catégories

---

<sup>3</sup> <https://www.1edtech.org/standards/lti>

	Var description mesure bas niveau	Type	Outil	Domaine linguistique
1	Number of collocations used nb backspace sequence longer than 3	Collocations	Collocation_tool lexical	
2		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 1.** Illustration de quelques-unes des 591 mesures du dispositif

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) [16] en langue. Des descripteurs 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 2). 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 descripteurs CECR (Cf. Tableau 3) de l’expression écrite [16, p.184] principalement concernés. Par exemple, le nombre de collocations relève de la catégorie *étendue du vocabulaire* (*vocabulary range*).

## 4 Le tableau de bord

### 4.1 Méthode participative de conception

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. Une démarche de *Design-Based Research*, reposant sur la méthode PADDLE, adaptée pour ce projet, a été retenue. Cette adaptation a été

Var	Description mesure bas niveau	Type	Unité première	Lien logique	Unité secondeaire	Lien logique	Unité 2 tertiaire
1	Number of collocations used Nb backspace	Collocations	collocations include verb		with	noun	
2	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
	Probability of actual						
7	learner adverb being predicted by learner model	BERT_EF	adverb	NA	NA	NA	NA

**Table 2.** Extrait illustratif de 14 mesures en fonction de leur catégorisation lexico-grammaticale et discursive

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
	Probability of the actual learner		
7	adverb tokens being predicted by the learner model	User simulator	vocabulary range: diversity

**Table 3.** Extrait (à titre d'exemple) de 14 mesures en fonction de leur catégorisation CECR

effectuée en raison des mesures déjà connues. Deux sessions de 3 heures chacune ont été organisées avec 10 enseignantes d'anglais de l'Université ANONYME. Lors de la première, le projet et les mesures ont été présentés, suivis d'ateliers de 3 groupes pour concevoir des maquettes papier de TBA, fondées sur les taxonomies pour répondre à des exemples choisis par les enseignantes (ex. Figure 2).

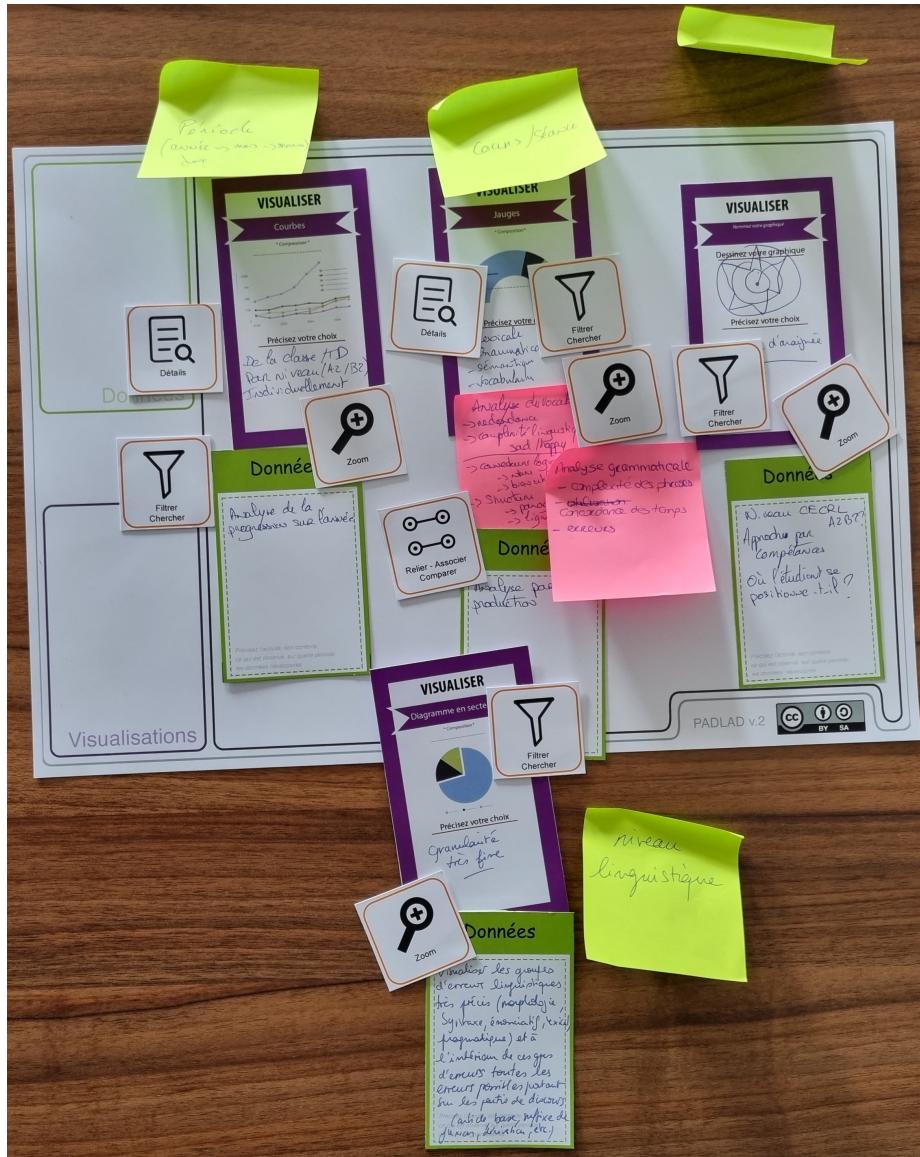
La plupart des enseignantes ont montré une certaine difficulté à se projeter dans un système d'analytiques qui ne soit pas centré sur les fautes, malgré la présentation initiale et les taxonomies de mesures proposées. Sur les trois TBA proposés, les enseignantes ont fait remonter les objectifs principaux 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 ;
- permettre la création d'ateliers sur mesure avec des ressources et des activités personnalisées.

Il est à noter que les trois groupes ont identifié un scénario original de "parcours utilisateur". Cette navigation permet l'exploration données, soit temporelle, 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 présentant trois vues principales : (i) un histogramme des différents niveaux CECR estimés des étudiants, (ii) un diagramme en radar permettant d'affiner la prédiction des étudiants selon un axe d'analyse choisi (paradigmatique/syntagmatique, domaine linguistique, ou compétences du CECR), et (iii) une représentation des proximités de modèles d'apprenants. Cette maquette était alors pourvue uniquement d'explications globales.

Celle-ci a été présentée lors d'un second *focus group* de 3h où une synthèse était présentée dans un premier temps. Les enseignantes devaient ensuite découvrir 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

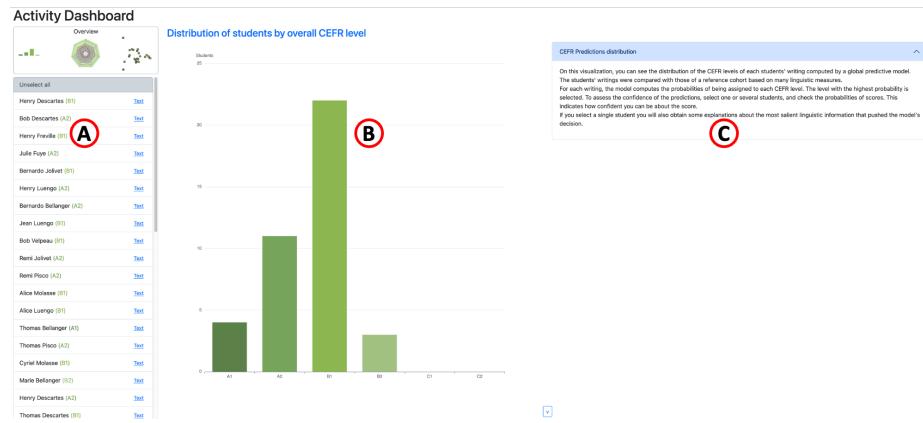


**Fig. 2.** Exemple de maquette de TBA

catégorielle du rouge au vert par niveau avait été proposée). En prenant en compte l'ensemble des remarques, et avec l'intégration des travaux parallèles sur les modèles prédictifs, nous avons implanté une première version exploitable du TBA.

## 4.2 Vue générale du tableau de bord

Le TBA actuel (Figure 3) est contextuel à une activité de rédaction de texte particulière. Cette photographie instantanée des réalisations des apprenants permet de les situer la production des apprenants sur l'échelle CEFR 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 second raffinement sur les concepts linguistiques mobilisés (par axe et dimensions).



**Fig. 3.** Le TBA (vue d'ensemble)

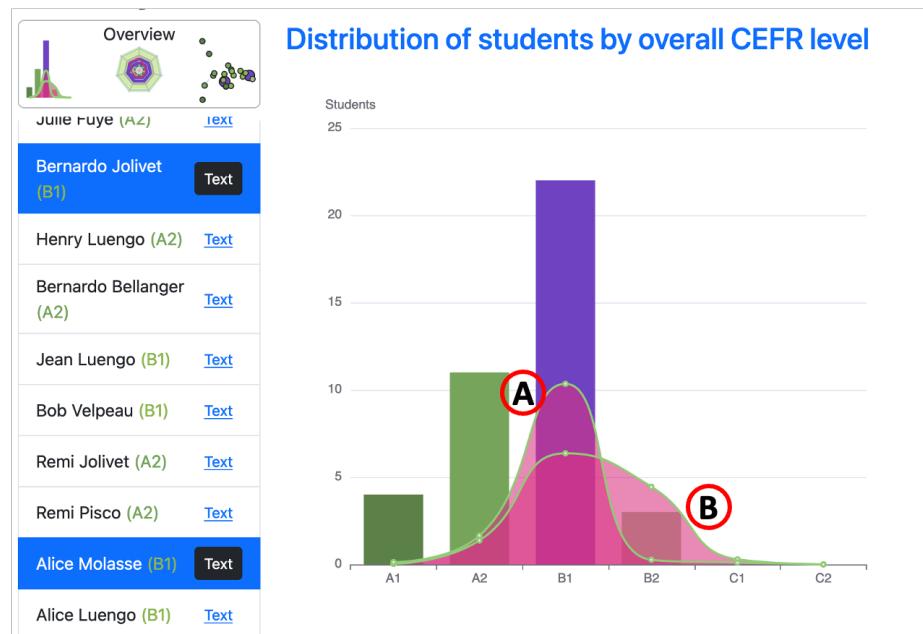
Le TBA se divise en 3 colonnes : à gauche (**A** sur la Figure 3) est présentée la liste des apprenants, qui permet leur sélection et la consultation éventuelle de leur texte ; les visualisations sont affichées au centre (**B**). Un panneau escamotable à droite (**C**) présente les explications accompagnant ces visualisations. Le système offre trois visualisations :

- l'histogramme de distribution du groupe classe selon les niveaux CEFR ;
- un diagramme en radar des niveaux CEFR estimés des apprenants par dimension selon l'axe d'analyse choisi (Compétences CEFR ou domaines linguistiques) ;
- 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.

### 4.3 Distribution globale du groupe

À l'échelle du groupe, cet histogramme reflète la répartition des niveaux CEFR estimés par le modèle global. Lorsqu'un ou plusieurs étudiants sont sélectionnés, leurs courbes de densité de probabilité apparaissent en surimpression (Figure 4), 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.

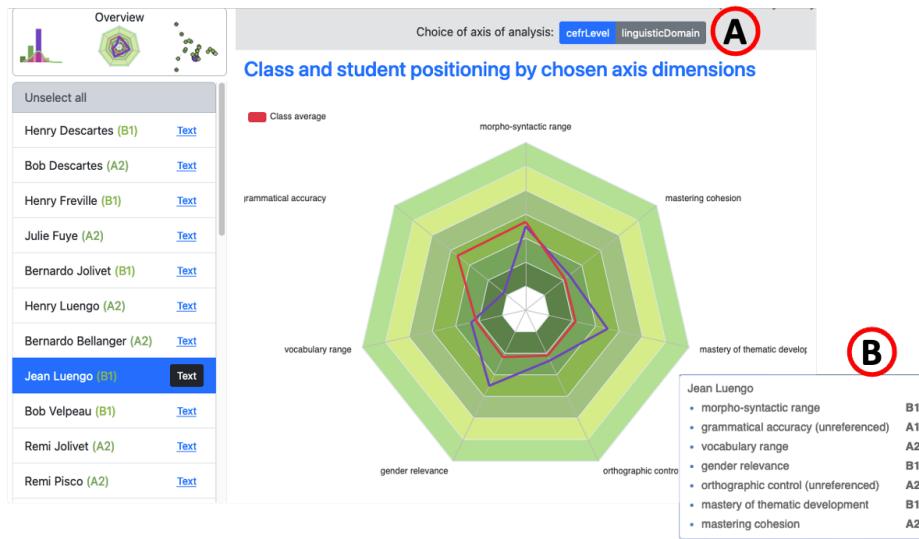


**Fig. 4.** Distribution globale des niveaux CEFR estimés

### 4.4 Niveaux CECR par dimensions d'analyse

Comme indiqué précédemment (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 *Random Forest*, selon [17], 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 justesse moyenne supérieure à 0,4 (le seuil de décision aléatoire sur 6 classes étant de 0,16) ont été retenus.



**Fig. 5.** Niveaux CECR par dimensions d'un axe d'analyse

Tous les axes étant exprimés sur la même échelle (niveaux CECR), un diagramme radar a été choisi pour la visualisation (Figure 5), outil fréquent en *learning analytics* [24] 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 radar affiche en rouge les moyennes du groupe classe. Lorsque 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 sur une ligne détaille les niveaux estimés des dimensions (**B**).

Cette visualisation permet ainsi d'identifier les disparités entre étudiants. Ainsi, dans la Figure 5, 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.

#### 4.5 Moteur générique d'explication

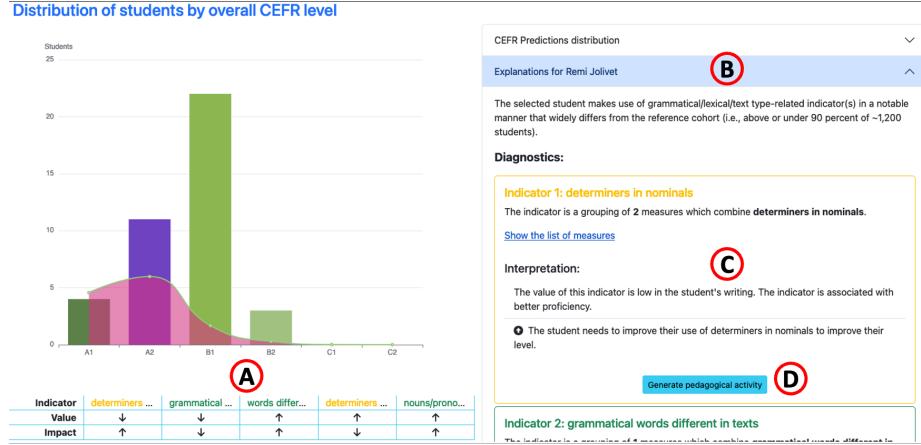
À chaque visualisation, une explication générale explique brièvement l'indicateur affiché sur la visualisation, la manière dont ce dernier a été obtenu et le principe général d'interprétation (ex. Figures 3 et 5). Dès lors qu'un ou plusieurs étudiants sont sélectionnés, un moteur générique (i.e.: non 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 interprétable" :

- 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".

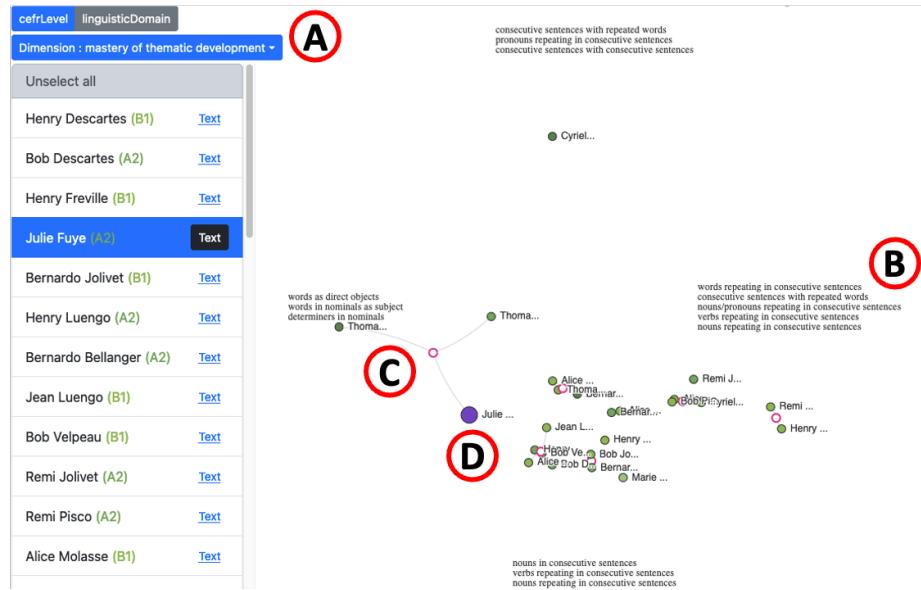
Pour dépasser la difficulté de compréhension d'une mesure isolée et tendre vers une potentielle actionnabilité, le moteur s'appuie sur la structure hiérarchique des unités linguistiques (UL) auxquelles sont rattachées les mesures (Cf. 3.3). Hypothétiquement, plus le niveau d'abstraction de l'UL est faible, 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 interprétable 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 étudiants. Les groupes sont triés par ordre de poids, dépendant positivement avec l'importance des mesures pour le modèle et de l'écart de leur valeur pour l'étudiant à la médiane, et négativement avec le niveau d'abstraction de l'UL.

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 6), mentionnant leur direction interprétable, 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.



**Fig. 6.** Explications locales pour un étudiant pour son niveau global



**Fig. 7.** Projection des étudiants dans le plan

#### 4.6 Comparaison des modèles d'apprenant

Une dernière visualisation (Figure 7) projette les apprenants sur un plan en 2D, 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**). Cette dernière visualisation permet d'identifier des profils similaires ou distants, la recherche de traits caractéristiques d'un même niveau ou la composition de groupes.

## 5 Discussion et perspectives

Les modèles statistiques sur lesquels reposent ces 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 un échantillon test du corpus EFCAMDAT [45]. 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% [3] à 83.9% [10]. Ce modèle ayant été entraîné sur un autre échantillon du même corpus, nous avons aussi évalué le modèle sur un corpus externe 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%.

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 d'intérêt. Il permet alors la sélection personnalisée d'unités linguistiques remarquables propres au filtre employé. L'enseignant (et ses étudiants) peuvent alors analyser les textes au regard de l'unité linguistique signalée. Cela permet une démarche métalinguistique essentielle dans le processus d'apprentissage. Les apprenants interrogent les formes en contexte, ce qui les aide à diagnostiquer les points à poursuivre et amplifier ou consolider. Ces points peuvent être renforcés grâce à la génération d'un prompt LLM permettant la création d'exercices spécifiques.

Cette taxonomie permet d'unifier les mesures en fonction d'une liste exhaustive de concepts linguistiques. Cependant, la correspondance entre les mesures et les concepts reste à évaluer. Une expérience en cours interroge l'accord inter-juge que des linguistes peuvent avoir sur cette classification en unités linguistiques. Les unités linguistiques contextualisées en recommandations restent aussi à évaluer du point de vue de leur actionnabilité. Un nouveau *focus group* va permettre de tester l'interprétabilité de ces recommandations, soit de vérifier la pertinence et cohérence des commentaires métalinguistiques produits. Il s'agira de comprendre la capacité des enseignants à prendre des décisions sur la base de ces visualisations. Plus généralement, les prochains travaux porteront sur la validité de la démarche [14, p.92]. Au delà de l'évaluation des scores sur des données historiques annotées, de l'évaluation de la généralisabilité des modèles sur différents types de données (tâches, sujets, type de textes), il reste des questions concernant **i)** l'extrapolation potentielle des scores indicatifs de niveau et **ii)** l'impact

positif ou négatif sur les apprenants du système. Une étude pilote sur cohorte sera mise en place en milieu écologique pour étudier cet impact en analysant les évolutions de mesures répétées chez les apprenants.

En terme de développements ultérieurs, un module est prévu pour visualiser les propriétés décelables par les résultats des mesures au niveau du texte. Cela concerne particulièrement, l'écart aux mesures d'association des collocations, ce qui permettra de signaler les erreurs de collocation produites par les apprenants. Les traces numériques clavier ont vocation à servir de modélisation, dans un premier temps de classification des essais écrits par des humains ou recopiés à partir d'une IA générative. 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. Les LLM pourront également permettre de prédire le niveau des textes rédigés. 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.

L'ensemble des dimensions explorées de la compétence ne porte que sur les productions écrites. Il est tout à fait envisageable de chercher à étendre pour l'oral des représentations sous forme de TBA qui permettent aux apprenants de se situer dans une cohorte, mais également de situer leurs productions orales en rapport avec les normes de prononciation en vigueur. Le recours à une version modifiée de Whisper [43] permet d'établir un alignement entre transcription et structures formantiques pour une analyse plus fine des compétences phonologiques [7], et de générer un score global de la qualité de l'anglais oral [8], qu'il conviendra d'étonner sur les valeurs du CEGR.

## 6 Conclusion

A l'heure des premières expériences, pas toujours concluantes [50], 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. Le prototype peut être consulté sur [ANONYME]. Conçu au départ pour l'anglais, mais également testé pour le suédois et l'espagnol, l'ensemble du dispositif [ANONYME] a vocation à 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.

## References

- [1] Falah Amro and Jered Borup. "Exploring Blended Teacher Roles and Obstacles to Success When Using Personalized Learning Software". en. In: *Journal of Online Learning Research* 5.3 (2019). Publisher: Association for the Advancement of Computing in Education ERIC Number: EJ1241760, pp. 229–250. URL: <https://eric.ed.gov/?id=EJ1241760> (visited on 10/03/2024).
- [2] Yigal Attali and Jill Burstein. "Automated Essay Scoring With e-rater® V.2". en-US. In: *The Journal of Technology, Learning and Assessment* 4.3 (2006), pp. 3–29. ISSN: 1540-2525. URL: <https://ejournals.bc.edu/ojs/index.php/jtla/article/view/1650> (visited on 04/04/2019).
- [3] Anon. authors. "Anon. title". en. In: 2019. ISBN: Anon. isbn.
- [4] Anon. authors. "Anon. title". In: June 2019.
- [5] Anon. authors. "Anon. title". In: 2022.
- [6] Anon. authors. "Anon. title". en. In: (May 2023).
- [7] Anon. authors. "Anon. title". In: 2023.
- [8] Anon. authros. "Anon. title". In: (2024).
- [9] Douglas Biber et al. "Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement". en. In: *Journal of English for Academic Purposes* 46 (2020), p. 100869. ISSN: 1475-1585. URL: <http://www.sciencedirect.com/science/article/pii/S1475158519305909> (visited on 01/07/2021).
- [10] Andrew Caines and Paula Butterly. "REPROLANG 2020: Automatic Proficiency Scoring of Czech, English, German, Italian, and Spanish Learner Essays". eng. In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, May 2020, pp. 5614–5623. ISBN: 979-10-95546-34-4. URL: <https://aclanthology.org/2020.lrec-1.689/> (visited on 01/10/2025).
- [11] Scott A. Crossley, Kristopher Kyle, and Mihai Dascalu. "The Tool for the Automatic Analysis of Cohesion 2.0: Integrating semantic similarity and text overlap". en. In: *Behavior Research Methods* 51.1 (2019), pp. 14–27. ISSN: 1554-3528. DOI: 10.3758/s13428-018-1142-4. URL: <https://doi.org/10.3758/s13428-018-1142-4> (visited on 05/03/2019).
- [12] Scott A. Crossley, Kristopher Kyle, and Danielle S. McNamara. "The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion". en. In: *Behavior Research Methods* 48.4 (Dec. 2016), pp. 1227–1237. ISSN: 1554-3528. DOI: 10.3758/s13428-015-0651-7. URL: <https://doi.org/10.3758/s13428-015-0651-7> (visited on 01/25/2024).
- [13] Scott A. Crossley and Danielle S. McNamara. "Predicting second language writing proficiency: the roles of cohesion and linguistic sophistication". en. In: *Journal of Research in Reading* 35.2 (2012), pp. 115–135. ISSN: 1467-9817. DOI: 10.1111/j.1467-9817.2010.01449.x. (Visited on 06/23/2021).

- [14] Sara Cushing Weigle. "English language learners and automated scoring of essays: Critical considerations". en. In: *Assessing Writing*. Automated Assessment of Writing 18.1 (Jan. 2013), pp. 85–99. ISSN: 1075-2935. DOI: 10.1016/j.asw.2012.10.006. URL: <https://www.sciencedirect.com/science/article/pii/S1075293512000499> (visited on 11/16/2022).
- [15] Mark Davies. "The 385+ million word Corpus of Contemporary American English (1990–2008+): Design, architecture, and linguistic insights". en. In: *International Journal of Corpus Linguistics* 14.2 (2009), pp. 159–190. ISSN: 1384-6655, 1569-9811. DOI: 10.1075/ijcl.14.2.02dav. URL: <http://www.jbe-platform.com/content/journals/10.1075/ijcl.14.2.02dav> (visited on 10/03/2015).
- [16] Council of Europe. Council for Cultural Co-operation. Education Committee. Modern Languages Division. *Common European framework of reference for languages: Learning, teaching, assessment*. Cambridge University Press, 2001.
- [17] Eibe Frank and Mark Hall. "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. Springer, 2001, pp. 145–156.
- [18] Jean-Marie Gilliot et al. "Conception participative de tableaux de bord d'apprentissage". In: *IHM'18: 30e Conférence Francophone sur l'Interaction Homme-Machine*. 2018, pp–119.
- [19] Riccardo Guidotti. "Counterfactual explanations and how to find them: literature review and benchmarking". In: *Data Mining and Knowledge Discovery* 38.5 (2024), pp. 2770–2824.
- [20] Carl C Haynes. "The role of self-regulated learning in the design, implementation, and evaluation of learning analytics dashboards". In: *Proceedings of the seventh ACM conference on learning@ scale*. 2020, pp. 297–300.
- [21] Alex Housen. "Difficulty and Complexity of Language Features and Second Language Instruction". en. In: *The Encyclopedia of Applied Linguistics*. John Wiley & Sons, Inc., 2014. ISBN: 978-1-4051-9843-1. URL: <http://onlinelibrary.wiley.com.passerelle.univ-rennes1.fr/doi/10.1002/9781405198431.wbeal1443/abstract> (visited on 06/27/2016).
- [22] Dabbebi Ines et al. "Towards adaptive dashboards for learning analytician approach for conceptual design and implementation". In: *International Conference on Computer Supported Education*. Vol. 2. SCITEPRESS. 2017, pp. 120–131.
- [23] Ioana Jivet et al. "License to evaluate: Preparing learning analytics dashboards for educational practice". In: *Proceedings of the 8th international conference on learning analytics and knowledge*. 2018, pp. 31–40.
- [24] Dan Kaczynski, Leigh Wood, and Ansie Harding. "Using radar charts with qualitative evaluation: Techniques to assess change in blended learning". In: *Active Learning in Higher Education* 9.1 (2008), pp. 23–41.

- [25] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. "Examples are not enough, learn to criticize! criticism for interpretability". In: *Advances in neural information processing systems* 29 (2016).
- [26] Kristopher Kyle. "Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication". Dissertation. Georgia: Georgia State University, 2016. URL: [https://scholarworks.gsu.edu/alesl\\_diss/35](https://scholarworks.gsu.edu/alesl_diss/35).
- [27] Kristopher Kyle, Scott Crossley, and Cynthia Berger. "The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0". eng. In: *Behavior Research Methods* 50.3 (2018), pp. 1030–1046. ISSN: 1554-3528. DOI: 10.3758/s13428-017-0924-4.
- [28] Kristopher Kyle and Scott A. Crossley. "Automatically Assessing Lexical Sophistication: Indices, Tools, Findings, and Application". en. In: *TESOL Quarterly* 49.4 (2015), pp. 757–786. ISSN: 1545-7249. DOI: 10.1002/tesq.194. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1002/tesq.194> (visited on 04/02/2024).
- [29] Xiaofei Lu. *Computational Methods for Corpus Annotation and Analysis*. en. Dordrecht: Springer, 2014. (Visited on 03/30/2016).
- [30] Xiaofei Lu. "The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives". en. In: *The Modern Language Journal* 96.2 (2012), pp. 190–208. ISSN: 1540-4781. DOI: 10.1111/j.1540-4781.2011.01232\_1.x. URL: [http://onlinelibrary.wiley.com.passerelle.univ-rennes1.fr/doi/10.1111/j.1540-4781.2011.01232\\_1.x/abstract](http://onlinelibrary.wiley.com.passerelle.univ-rennes1.fr/doi/10.1111/j.1540-4781.2011.01232_1.x/abstract) (visited on 06/27/2016).
- [31] Marie-Catherine de Marneffe et al. "Universal Dependencies". In: *Computational Linguistics* 47.2 (June 2021). Place: Cambridge, MA Publisher: MIT Press, pp. 255–308. DOI: 10.1162/coli\_a\_00402. URL: <https://aclanthology.org/2021.cl-2.11> (visited on 04/05/2023).
- [32] Wannisa Matcha et al. "A Systematic Review of Empirical Studies on Learning Analytics Dashboards: A Self-Regulated Learning Perspective". In: *IEEE Transactions on Learning Technologies* 13.2 (2020), pp. 226–245. DOI: 10.1109/TLT.2019.2916802.
- [33] Tim Miller. "Explanation in artificial intelligence: Insights from the social sciences". In: *Artificial Intelligence* 267 (Feb. 2019), pp. 1–38. ISSN: 0004-3702. DOI: 10.1016/j.artint.2018.07.007. (Visited on 02/23/2024).
- [34] Christoph Molnar. *Interpretable machine learning*. Lulu. com, 2020.
- [35] Gerald Nelson, Sean Wallis, and Bas Aarts. *The British Component of the International Corpus of English (ICE-GB) and ICECUP software (CD-ROM)*. London, 1998. URL: <http://www.ucl.ac.uk/english-usage/projects/ice-gb/> (visited on 10/25/2011).
- [36] John M. Norris and Lourdes Ortega. "Towards an Organic Approach to Investigating CAF in Instructed SLA: The Case of Complexity". en. In: *Applied Linguistics* 30.4 (2009). Publisher: Oxford Academic, pp. 555–578. ISSN: 0142-6001. DOI: 10.1093/applin/amp044. URL: <https://academic.oup.com/applij/article/30/4/555/225652> (visited on 09/07/2020).

- [37] John M. Norris, Steven J. Ross, and Rob Schoonen. “Improving Second Language Quantitative Research”. en. In: *Language Learning* 65.S1 (2015). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/lang.12110>, pp. 1–8. ISSN: 1467-9922. DOI: 10.1111/lang.12110. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12110> (visited on 11/16/2023).
- [38] Ellis B. Page. “The Use of the Computer in Analyzing Student Essays”. In: *International Review of Education* 14.2 (1968), pp. 210–225. ISSN: 0020-8566. URL: <http://www.jstor.org/stable/3442515> (visited on 04/27/2018).
- [39] Ildikó Pilán, Elena Volodina, and Torsten Zesch. “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*. Osaka, Japan: The COLING 2016 Organizing Committee, Dec. 2016, pp. 2101–2111. URL: <https://aclanthology.org/C16-1198> (visited on 10/15/2021).
- [40] Luke Plonsky and Talip Gonulal. “Methodological Synthesis in Quantitative L2 Research: A Review of Reviews and a Case Study of Exploratory Factor Analysis”. en. In: *Language Learning* 65.S1 (2015). \_eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/lang.12111>, pp. 9–36. ISSN: 1467-9922. DOI: 10.1111/lang.12111. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/lang.12111> (visited on 11/16/2023).
- [41] Peng Qi et al. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. arXiv:2003.07082 [cs]. Apr. 2020. DOI: 10.48550/arXiv.2003.07082. URL: <http://arxiv.org/abs/2003.07082> (visited on 06/01/2023).
- [42] Ashwin Rachha and Mohammed Seyam. *Explainable AI In Education : Current Trends, Challenges, And Opportunities*. Pages: 239. Apr. 2023. DOI: 10.1109/SoutheastCon51012.2023.10115140.
- [43] Alec Radford et al. “Robust speech recognition via large-scale weak supervision”. In: *International Conference on Machine Learning*. PMLR. 2023, pp. 28492–28518.
- [44] Björn Rudzewitz et al. “Enhancing a Web-based Language Tutoring System with Learning Analytics”. In: *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*. Ed. by Luc Paquette and Cristóbal Romero. Vol. 2592. Montréal, Canada: CEUR-WS, 2019, pp. 1–7.
- [45] Itamar Shatz. “Refining and modifying the EFCAMDAT: Lessons from creating a new corpus from an existing large-scale English learner language database”. en. In: *International Journal of Learner Corpus Research* 6.2 (2020). Publisher: John Benjamins, pp. 220–236. ISSN: 2215-1478, 2215-1486. DOI: 10.1075/ijlcr.20009.sha. URL: <https://www.jbe-platform.com/content/journals/10.1075/ijlcr.20009.sha> (visited on 10/25/2021).

- [46] Milan Straka, Jan Hajič, and Jana Straková. “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)*. Portorož, Slovenia: European Language Resources Association (ELRA), May 2016, pp. 4290–4297. URL: <https://aclanthology.org/L16-1680> (visited on 12/01/2021).
- [47] Sowmya Vajjala and Kaidi Lõo. “Automatic CEFR level prediction for Estonian learner text”. In: *Proceedings of the third workshop on NLP for computer-assisted language learning*. Ed. by Elena Volodina and Lars Borin. Uppsala, Sweden: LiU Electronic Press, 2014, pp. 113–127. URL: <https://aclanthology.org/W14-3509> (visited on 12/01/2021).
- [48] Sowmya Vajjala and Taraka Rama. “Experiments with Universal CEFR Classification”. In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*. New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 147–153. DOI: 10.18653/v1/W18-0515. URL: <https://www.aclweb.org/anthology/W18-0515> (visited on 12/11/2019).
- [49] Georgios Velentzas et al. “Logging Keystrokes in Writing by English Learners”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Ed. by Nicoletta Calzolari et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 10725–10746. URL: <https://aclanthology.org/2024.lrec-main.938>.
- [50] Rose Wang and Dorottya Demszky. “Is ChatGPT a Good Teacher Coach? Measuring Zero-Shot Performance For Scoring and Providing Actionable Insights on Classroom Instruction”. In: *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Ed. by Ekaterina Kochmar et al. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 626–667. DOI: 10.18653/v1/2023.bea-1.53. URL: <https://aclanthology.org/2023.bea-1.53>.
- [51] Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. *Second language development in writing: measures of fluency, accuracy, & complexity*. English. OCLC: 40664312. Honolulu: Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, 1998. ISBN: 978-0-8248-2069-5.
- [52] Helen Yannakoudakis, Ted Briscoe, and Theodora Alexopoulou. “Automating Second Language Acquisition Research: Integrating Information Visualisation and Machine Learning”. In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 35–43. URL: <https://aclanthology.org/W12-0206> (visited on 03/25/2022).
- [53] Helen Yannakoudakis et al. “Developing an automated writing placement system for ESL learners”. en. In: *Applied Measurement in Education* 31.3

(2018). DOI: 10.1080/08957347.2018.1464447. URL: <https://doi.org/10.1080/08957347.2018.1464447> (visited on 01/09/2025).