Actes de JEP-TALN-RECITAL 2024
Toulouse, France, 8 au 12 juillet 2024

TCFLE-8 : un corpus de productions écrites d'apprenants de français langue étrangère et son application à la correction automatisée de textes

Rodrigo Wilkens[1] Alice Pintard[1] David Alfter[2] Vincent Folny[3]
Thomas François[1]
(1) CENTAL, IL&C, Université catholique de Louvain, Belgique prenom.nom@uclouvain.be,
(2) University of Gothenburg david.alfter@gu.se
(3) France Éducation International Folny@france-education-international.fr
{rodrigo.wilkens, alice.pintard, thomas.francois}@uclouvain.be,
Folny@france-education-international.fr, david.alfter@gu.se

RÉSUMÉ

La correction automatisée de textes (CAT) vise à évaluer automatiquement la qualité de textes écrits. L'automatisation permet une évaluation à grande échelle ainsi qu'une amélioration de la cohérence, de la fiabilité et de la normalisation du processus. Ces caractéristiques sont particulièrement importantes dans le contexte des examens de certification linguistique. Cependant, un goulot d'étranglement majeur dans le développement des systèmes CAT est la disponibilité des corpus. Dans cet article, nous visons à encourager le développement de systèmes de correction automatique en fournissant le corpus TCFLE-8 [1], un corpus de 6 569 essais collectés dans le contexte de l'examen de certification *Test de Connaissance du Français* (TCF). Nous décrivons la procédure d'évaluation stricte qui a conduit à la notation de chaque essai par au moins deux évaluateurs selon l'échelle du Cadre européen commun de référence pour les langues (CECR) et à la création d'un corpus équilibré. Nous faisons également progresser les performances de l'état de l'art pour la tâche de CAT en français en expérimentant deux solides modèles de référence.

ABSTRACT

TCFLE-8 : a Corpus of Learner Written Productions for French as a Foreign Language and its Application to Automated Essay Scoring [2]

Automated Essay Scoring (AES) aims to automatically assess the quality of essays. Automation enables large-scale assessment, improvements in consistency, reliability, and standardization. Those characteristics are of particular relevance in the context of language certification exams. However, a major bottleneck in the development of AES systems is the availability of corpora. In this paper, we aim to foster the development of AES by providing the TCFLE-8 corpus, a corpus of 6.5k essays collected in the context of the French Knowledge Test (TCF) certification exam. We report the strict quality procedure that led to the scoring of each essay by at least two raters according to the levels of the Common European Framework of Reference for Languages (CEFR) and to the creation of a

1. TCFLE-8 est disponible à l'adresse https://www.france-education-international.fr/corpus
2. Cet article est une adaptation d'une publication en anglais : Wilkens, R., Pintard, A., Alfter, D., Folny, V., & François, T. (2023). TCFLE-8 : a Corpus of Learner Written Productions for French as a Foreign Language and its Application to Automated Essay Scoring. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 3447-3465).

31ème Conférence sur le Traitement Automatique des Langues Naturelles,
volume 1 : articles longs et prises de position, pages 677–685.

balanced corpus. We also advance the state-of-the-art performance for the AES task in French by experimenting with two strong baselines.

MOTS-CLÉS : correction automatisée de textes, français langue étrangère, TCF, corpus d'apprenants.

KEYWORDS: automated essay scoring, French as a foreign language, TCF, learner corpus.

1 Introduction

La correction automatisée de textes (CAT – AES pour *automatic essay scoring* en anglais) vise à développer des algorithmes capables d'évaluer la qualité de productions écrites de la même manière que des évaluateurs humains. Les origines de ce domaine remontent aux recherches de Page (1966). Depuis lors, de nombreux chercheurs se sont penchés sur cette question et plusieurs synthèses du domaine ont été publiées récemment (Ramesh & Sanampudi, 2022; Lagakis & Demetriadis, 2021; Klebanov & Madnani, 2021; Uto, 2021; Klebanov & Madnani, 2020; Ke & Ng, 2019; Shermis *et al.*, 2013). En bref, le début du 21e siècle fut marqué par des systèmes reposant soit sur des corpus annotés et sur l'extraction automatisée de caractéristiques linguistiques qui étaient ensuite combinées via des algorithmes d'apprentissage automatisé (Burstein *et al.*, 1998; Rudner & Liang, 2002; Dikli, 2006), soit sur des approches non supervisées recourant souvent à l'analyse sémantique latente (Landauer & Dumais, 1997). Les récentes avancées en CAT ont été rendues possibles par les algorithmes d'apprentissage profond et les grands modèles de langage (Ramesh & Sanampudi, 2022). Néanmoins, ces approches ont également exacerbé le besoin de grands corpus annotés.

Par conséquent, les équipes de recherche en CAT se sont intéressées aux travaux menées par ailleurs sur les corpus d'apprenants, une branche de la linguistique de corpus qui collecte des productions authentiques et informatisées d'apprenants à grande échelle. Des travaux pionniers tels que le *International Corpus of Learner English* (ICLE) (Granger, 1993) et le *European Science Foundation L2 Database* (Perdue, 1993) ont démontré le potentiel de ces collections de données d'apprenants pour la recherche sur l'acquisition d'une langue seconde, mais ce n'est que récemment que davantage de corpus d'apprenants ont été exploités en CAT.

Malheureusement, il n'existe pas de corpus suffisamment important pour le français, ce qui rend la situation de la CAT en français loin d'être encourageante. Les premiers systèmes ont, par conséquent, utilisé des approches non supervisées : Lemaire & Dessus (2001) a utilisé l'analyse sémantique latente pour comparer des essais en français langue maternelle (FLM) avec des passages de référence tirés de manuels, tandis que Zaghouani (2002) capture automatiquement plusieurs caractéristiques d'essais en FLM, lesquelles sont combinées de manière heuristique. Plus récemment, Parslow (2015) a entraîné un classificateur « Naive Bayes » sur un corpus de 200 essais écrits en français langue étrangère (FLE). Enfin, Ranković *et al.* (2020) ont été les premiers à affiner BERT pour la CAT en FLE, mais ils n'ont pas publié leur jeu de données où, de plus, une seule langue maternelle est représentée.

Par conséquent, afin de soutenir le développement de solutions en CAT pour le français, le besoin d'un large corpus fiablement annoté en matière de compétence est patent. Dans cet article, nous apportons deux contributions principales. Premièrement, nous fournissons à la communauté le corpus TCFLE-8 [3], composé de 6 569 productions d'apprenants. Ces textes ont été collectés dans le cadre

3. Ce nom combine le nom de l'examen d'où proviennent les textes, le *Test de connaissance du français*, et l'acronyme FLE. Le 8 fait référence aux huit différentes langues usuelles représentées dans le corpus.

de l'examen officiel du Test de connaissance du français (TCF), l'un des principaux examens de certification pour le FLE. La section 2 décrit en détails les propriétés du TCF et de son évaluation par FEI ainsi que le processus de nettoyage et d'échantillonnage du corpus. Le corpus résultant de ces opérations est présenté à la Section 3, qui décrit les différentes informations disponibles, parmi lesquelles la langue usuelle des candidats, l'évaluation du niveau de compétence à l'écrit des candidats, assignés par au moins 2 évaluateurs sur l'échelle du Cadre européen commun de référence pour les langues (CECR) (Council of Europe, 2001), des informations sur la tâche à accomplir et une annotation automatisée de nombreuses variables linguistiques réalisées par FABRA (Wilkens *et al.*, 2022). La seconde contribution consiste en une série d'expériences en correction automatique de textes sur la base de ce corpus, qui visent à proposer un solide point de référence en vue de futures recherches. Ces expériences sont décrites à la section 4.

2 Méthodologie de conception du corpus

2.1 Collecte des données

TCFLE-8 étant un corpus de productions écrites de candidats au TCF. Il a été collecté par l'un des opérateurs effectuant la certification officielle en FLE : *France Education International* (FEI). FEI est un opérateur français placé sous la tutelle du ministère de l'Éducation nationale et de la Jeunesse. Avec un effectif de plus de 250 salariés et un réseau de plus de 1 000 experts, FEI intervient dans différents domaines de la coopération en matière d'éducation et de formation et contribue à la promotion de la langue française et de la francophonie. FEI propose une large gamme de certifications en français alignées sur les six niveaux du CECR : diplôme initial de langue française (DILF), diplôme d'études en langue française (DELF), diplôme d'études approfondies en langue française (DALF) et test de connaissance du français (TCF). Environ 650 000 candidats se présentent chaque année à l'un de ces examens dans plus de 180 pays.

Comme son nom l'indique, TCFLE-8 est basé sur le TCF, un test linéaire aligné sur les six niveaux du CECR. Le TCF est principalement utilisé dans les contextes d'admission à des études universitaires, de migration et d'accès à la citoyenneté. Sa composante écrite, composée de trois tâches indépendantes, est passée chaque année par 120 000 candidats, dont 60 % passent l'examen sur ordinateur.

Les trois tâches visent à tester les capacités des candidats à s'exprimer en français à l'écrit et nécessitent de rédiger, par exemple, un message, un article, un courrier ou un texte comparant deux points de vue [4]. Ces tâches sont corrigées par des évaluateurs experts. FEI dispose d'un panel d'une centaine de correcteurs, recrutés sur la base de leur profil professionnel (enseignants expérimentés ayant une expérience préalable de l'évaluation en français). Les candidats évaluateurs passent un test psychométrique validant leurs compétences en évaluation de l'écrit et suivent une formation de deux jours. À l'issue de cette procédure, le recrutement est confirmé ou non. Pour garantir la fidélité des corrections à long terme, les indices de fidélité des évaluateurs sont évalués périodiquement et une décision est prise quant à leur maintien dans le panel. En outre, pour garantir la fidélité au niveau des candidats, FEI adopte une approche de double notation indépendante. En cas de désaccord, un troisième évaluateur vient en renfort pour évaluer indépendamment les trois productions. Le niveau final du candidat est établi sur la base de la fréquence des niveaux du CECR attribués aux trois

4. Plus de détails sur la nature des tâches sont disponibles sur le site du TCF : https://www.france-education-international.fr/test/tcf-tout-public?langue=fr

productions du candidat.

Malgré cette stricte procédure, la compétence langagière étant multidimensionnelle (Bachman, 1990; Bachman & Palmer, 2010; Oller & Hinofotis, 1980; Vollmer & Sang, 1983) et mesurable (Vollmer & Caroll, 1983), mesurer les compétences rédactionnelles implique de prendre en compte différentes facettes : les compétences du candidat, l'indulgence ou la sévérité de l'évaluateur et la difficulté de la tâche. À cette fin, « l'utilisation de modèles de Rasch à multi-facettes (MRMF) est une approche psychométrique qui établit un cadre cohérent pour tirer des conclusions fiables, valides et justes des évaluations effectuées par les évaluateurs, répondant ainsi au problème des évaluations humaines faillibles » (Eckes, 2009). Nous avons donc appliqué le MRMF à l'ensemble de la base de données des examens TCF afin d'identifier les évaluations humaines faillibles et d'éviter de les intégrer dans le corpus.

2.2 Nettoyage des données

Les données collectées par FEI ont dû être nettoyées à différents égards. Tout d'abord, la détection des valeurs aberrantes nous a conduit à supprimer les réponses des candidats qui n'atteignaient pas le niveau A1, étaient des copies de la question, ou étaient trop courtes, trop longues ou encore hors sujet. Ensuite, nous avons exploité les informations du modèle multi-facettes de Rasch afin de détecter les textes pour lesquels les évaluateurs humains semblaient ne pas avoir fourni un jugement fiable. À cette fin, nous avons comparé les scores CECR originaux des évaluateurs FEI et les scores ajustés par la méthode MRMF et avons supprimé tous les essais dont la valeur des résidus standardisés était supérieure à 4. En outre, nous avons également supprimé les productions dont l'évaluation semblait peu fiable (par exemple, pour les candidats qui se situent à la limite entre deux niveaux). Pour ce faire, nous avons supprimé tous les cas où les deux évaluateurs n'étaient pas d'accord entre eux ni avec la note finale du candidat, et nous avons également supprimé les cas où il y avait une distance de trois niveaux du CECR entre la note la plus basse et la note la plus élevée attribuée à l'une des trois tâches.

Après ce processus, nous avons attribué à chaque production le niveau CECR du candidat, lorsqu'au moins un des évaluateurs avait également donné ce niveau à la production. Par ailleurs, si les deux évaluateurs avaient attribué le même niveau à la production, nous lui avons attribué ce niveau (même si ce niveau n'était pas identique au niveau global du candidat). Toute production ne répondant pas à l'un de ces deux critères a été supprimée.

Après l'élimination des valeurs aberrantes, l'étape suivante a consisté à obtenir un échantillon représentatif de l'ensemble des tâches du TCF disponibles. Pour une représentation équilibrée, le niveau CECR du texte est une variable évidente à contrôler. En outre, nous avons contrôlé la langue usuelle [5], dans le but d'obtenir une représentativité des langues usuelles les plus fréquentes. Comme les cinq premières étaient toutes européennes et que la sixième était le kabyle, une langue afro-asiatique, nous avons également inclus le chinois et le japonais afin d'obtenir une meilleure représentativité des différentes familles typologiques de langues. Nous avons donc lancé une procédure d'échantillonnage aléatoire stratifié en contrôlant les 6 niveaux du CECR et la langue usuelle du candidat. La Table 1 décrit le corpus résultant de cette procédure. Pour finir, le corpus a été anonymisé et pseudo-anonymisé à l'aide de l'outil MAPA (Gianola *et al.*, 2020), afin de préserver l'identité des auteurs des textes qui contiennent parfois des informations personnelles.

5. La langue usuelle est la langue que le candidat a indiqué dans le formulaire d'inscription comme étant celle qu'il utilise habituellement.

Langue	A1	A2	B1	B2	C1	C2	Total
JPN	8	135	171	170	48	2	534
CHI	34	165	244	189	45	4	681
SPA	124	187	175	182	178	58	904
ARA	135	160	163	153	160	135	906
POR	102	187	182	191	172	38	872
ENG	125	163	167	165	169	128	917
RUS	103	198	183	196	180	29	889
KAB	58	180	181	181	175	91	866
Total	689	1375	1466	1427	1127	485	6569

TABLE 1 – Nombre de textes en fonction de la langue usuelle et du niveau CECR (les codes des langues suivent la norme ISO639-2)

3 Présentation du corpus

À la fin du processus de compilation, le corpus TCFLE-8 comprend 6 569 essais (581 333 mots). La Table 1 présente des chiffres plus précis sur la proportion de textes par niveau du CECR et par langue usuelle. Les niveaux extrêmes (A1 et C2) sont moins représentés dans le corpus. Cela s'explique par deux facteurs : (1) peu d'apprenants de niveau A1 cherchent à passer une certification en langue, car ce niveau est rarement suffisant à des fins officielles (par exemple, pour l'obtention d'un emploi ou d'un visa), et (2) il est extrêmement difficile d'atteindre le niveau C2 dans une langue étrangère.

Dans la version anglaise originale de cet article (publié à EMNLP), le lecteur trouvera davantage de détails sur le corpus, notamment en matière de représentation de genres, de distribution des tâches et de longueur des productions. En résumé, on note que 58% des textes ont été écrits par des femmes et que cette proportion varie selon les niveaux CECR. Au niveau des tâches, les trois types de tâches sont relativement uniformément représentés, ce qui était espéré au vu de la procédure d'échantillonnage. Par ailleurs, une comparaison systématique entre TCFLE-8 et les autres corpus existants révèle qu'il s'agit du plus grand corpus d'apprenants de FLE adapté à la CAT – à la fois en termes de taille et de représentativité des L1 (ici, langue usuelle) –, du troisième plus large corpus de productions écrites de candidats à notre connaissance, toute langue confondue, et que ses couches d'annotation fournissent les informations les plus riches. En effet, non seulement, il couvre les 6 niveaux du CECR, mais a également fait l'objet d'une annotation linguistique visant à décrire les compétences des apprenants avec plus de 400 variables (chacune associées à 18 agrégateurs statistiques, comme la moyenne, la médiane, l'écart-type, etc. ce qui donne plus de 5 000 caractéristiques). Cette annotation a été réalisée automatiquement à l'aide de la boîte à outils FABRA (Wilkens *et al.*, 2022).

Par ailleurs, en complément du niveau CECR consolidé, issu de la procédure décrite plus haut, TCFLE-8 comprend également le niveau CECR atteint par chaque candidat au terme des trois productions écrites. Ce score correspond au niveau officiel du CECR attribué au candidat pour la partie écrite de l'examen du TCF. Le Kappa quadratique de Cohen pondéré (KQP) entre ces deux scores (niveau CECR du texte et niveau CECR du candidat) atteint 0,98. On s'attend à ce que cette valeur soit élevée, mais pas égale à 1, en raison des cas où les candidats ne peuvent pas maintenir un niveau constant de qualité durant l'épreuve. En outre, les notes attribuées par les deux évaluateurs de FEI dans le cadre de la procédure de double évaluation sont également disponibles. Elles ont un KQP de 0,71 entre elles et de 0,84 avec le niveau CECR consolidé de l'essai. Enfin, la nature de la tâche et sa position

dans la séquence des trois tâches du TCF sont également rapportées. Ces informations permettent de contextualiser la réponse du candidat.

4 Résultats pour la CAT en FLE

Dans cette section, nous rapportons rapidement nos expériences visant à évaluer l'utilité du corpus TCFLE-8 pour l'entraînement des systèmes de CAT. En d'autres termes, prédire le niveau CECR des textes rédigés par les candidats au TCF revient donc à rendre possible l'automatisation de la correction des productions écrites du TCF, mais nous espérons que le corpus puisse soutenir plus largement la correction de textes automatisée pour le français.

À cette fin, nous explorons deux approches : l'apprentissage profond, étant donné que la plupart des systèmes AES s'appuient sur des réseaux neuronaux (Ramesh & Sanampudi, 2022), et l'apprentissage automatique basé sur des variables. Pour le modèle d'apprentissage profond, nous avons utilisé CamemBERT (Martin *et al.*, 2020). Pour l'apprentissage non-neuronal, nous utilisons XGBoost, d'une part, et un simple modèle de régression logistique, d'autre part. Les performances de ces modèles sont présentées à la Table 2. Il apparaît clairement que le modèle basé sur CamemBERT obtient une meilleure exactitude et un meilleur score F1 que les deux autres modèles. Malgré tout, on peut constater qu'il y a encore une marge de progression lorsque l'on compare les résultats de CamemBERT avec l'évaluation des experts de FEI (colonne « Évaluateurs »). Néanmoins, ce modèle est proche de la performance des évaluateurs lorsque l'on considère la relation d'ordinalité entre les niveaux. Celle-ci est capturée à l'aide de la métrique κ quadratique pondéré (KQP), qui évalue l'accord entre le système et les annotateurs, en tenant compte de la distance entre les 6 niveaux du CECR, ainsi que de la mesure d'exactitude contiguë, calculée de la même manière que l'exactitude contiguë, mais où les erreurs d'un niveau de différence ne sont pas prises en compte.

	CamemBERT	**XGBoost**	**Logistique**	**Évaluateurs**
KQP	**0,88** (0,01)	0,79 (0,02)	0,69 (0,02)	0,93 (0,01)
Exactitude	**0,57** (0,01)	0,46 (0,01)	0,37 (0,01)	0,76 (0,01)
$\text{Exactitude}_{\text{Contiguë}}$	**0,98** (0,01)	0,92 (0,02)	0,80 (0,01)	0,99 (0,01)
$\text{F1}_{\text{pondérée}}$	**0,56** (0,01)	0,46 (0,02)	0,36 (0,02)	0,76 (0,01)
A1_{F1}	**0,63** (0,01)	0,59 (0,04)	0,54 (0,06)	0,76 (0,02)
A2_{F1}	**0,57** (0,04)	0,53 (0,01)	0,40 (0,05)	0,76 (0,03)
B1_{F1}	**0,56** (0,04)	0,45 (0,05)	0,32 (0,02)	0,75 (0,01)
B2_{F1}	**0,56** (0,04)	0,43 (0,03)	0,34 (0,03)	0,76 (0,02)
C1_{F1}	**0,56** (0,04)	0,42 (0,03)	0,30 (0,05)	0,77 (0,02)
C2_{F1}	**0,48** (0,09)	0,19 (0,07)	0,31 (0,02)	0,80 (0,04)

TABLE 2 – Moyenne et écart-type des performances des 3 modèles ainsi que la performance des évaluateurs humains sur TCFLE-8.

TCFLE-8 étant un nouveau corpus pour la langue française, nous ne pouvons pas comparer nos résultats avec les travaux précédents, en raison d'une différence considérable au niveau de la taille de ces corpus. Dans la littérature en CAT pour le français nous n'avons identifié que deux articles portant sur l'identification de la compétence écrite en FLE (cf. Section 1). Tout d'abord, Parslow (2015) a rapporté des scores F1 allant de $0,51$ à $0,74$ pour les niveaux A1 à B2. Deuxièmement, Ranković

et al. (2020) ont utilisé les couches intermédiaires de CamemBERT comme caractéristiques pour prédire le niveau dans un corpus de 100 essais et ont rapporté des MSE allant de $0,35$ à $0,55$.

5 Conclusion

Dans ce travail, nous avons présenté TCFLE-8, un corpus de 6 569 essais de candidats écrits pendant le test de connaissance du français (TCF), incluant 8 langues usuelles différentes. Cet article a décrit la collecte des données par France Education International (FEI) et les différentes étapes de nettoyage des données. Au final, nous obtenons le plus grand corpus en français ciblant le FLE pour la CAT. Ce corpus, ainsi que ses métadonnées (essais, métadonnées et annotations) sont à la disposition de la communauté. En explorant l'utilité de TCFLE-8 pour la tâche de CAT en FLE, nous avons appliqué différents algorithmes d'apprentissage automatique. CamemBERT apparaît comme le plus précis des trois.

Enfin, l'intérêt du corpus TCFLE-8 dépasse les frontières de l'AES, car le fait qu'il s'agisse d'un grand corpus d'apprenants, annoté avec les niveaux du CECR, ouvre de nombreuses pistes de recherches en TAL, mais aussi en acquisition du langage et en linguistique de corpus. Ainsi, il pourrait soutenir des recherches pour le développement de matériel pédagogique, qu'il s'agisse de dictionnaires (Longman, 2002), d'activités axées sur les difficultés et les erreurs courantes des apprenants (Kaszubski, 1998; Reppen, 2010), de logiciels d'apprentissage des langues assisté par ordinateur (Granger, 2003) ou d'aides à l'écriture en L2 (Link *et al.*, 2014). Avec 8 langues usuelles différentes, ce corpus pourrait également être utile pour des études interlinguistiques ciblant les mécanismes de transfert et l'influence de la L1 sur la production de la L2 (Golden *et al.*, 2017; Werner *et al.*, 2020) ou pour l'identification automatique de la langue maternelle (Tetreault *et al.*, 2013). Enfin, une autre application possible de ce nouveau corpus est la détection et la correction d'erreurs (Dahlmeier *et al.*, 2013), que nous étudions actuellement dans le cadre d'un travail futur sur TCFLE-8.

Références

BACHMAN L. & PALMER A. (2010). *Language assessment in practice : developing language assessments and justifying their use in the real world*. Oxford applied linguistics. Oxford : Oxford Univ. Press.

BACHMAN L. F. (1990). *Fundamental considerations in language testing*. Oxford University Press.

BURSTEIN J., KUKICH K., WOLFF S., LU C. & CHODOROW M. (1998). Computer analysis of essays. In *NCME Symposium on automated Scoring*.

COUNCIL OF EUROPE (2001). *Common European Framework of Reference for Lan- guages : Learning, Teaching, Assessment*. Press Syndicate of the University of Cambridge.

DAHLMEIER D., NG H. T. & WU S. M. (2013). Building a Large Annotated Corpus of Learner English : The NUS Corpus of Learner English. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 22–31, Atlanta, Georgia : Association for Computational Linguistics.

DIKLI S. (2006). An overview of automated scoring of essays. *The Journal of Technology, Learning and Assessment*, **5**(1).

ECKES T. (2009). *Quantitative Data Analysis for Language Assessment Volume I : Fundamental Techniques*. Routledge, 1 édition. DOI : 10.4324/9781315187815.

GIANOLA L., AJAUSKS Ē., ARRANZ V., GIBERT O. D. & MELERO M. (2020). Automatic removal of identifying information in official eu languages for public administrations : The mapa project. In *33rd International Conference on Legal Knowledge and Information Systems (JURIX 2020) : proceedings, Dec 2020, Brno, Prague, Czech Republic*, volume 334, p. 223–226 : IOS Press.

GOLDEN A., JARVIS S. & TENFJORD K. (2017). *Crosslinguistic Influence and Distinctive Patterns of Language Learning : Findings and Insights from a Learner Corpus*. Multilingual Matters.

GRANGER S. (1993). The International Corpus of Learner English. In *The European English Messenger*, p.3̃4.

GRANGER S. (2003). Error-tagged Learner Corpora and CALL : A Promising Synergy. *CALICO Journal*, **20**(3), 465–480.

KASZUBSKI P. (1998). Learner corpora : The cross-roads of linguistic norm. *TALC98 Proceedings*, p. 24–27.

KE Z. & NG V. (2019). Automated essay scoring : A survey of the state of the art. In *IJCAI*, volume 19, p. 6300–6308.

KLEBANOV B. B. & MADNANI N. (2020). Automated evaluation of writing–50 years and counting. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, p. 7796–7810.

KLEBANOV B. B. & MADNANI N. (2021). Automated Essay Scoring. *Synthesis Lectures on Human Language Technologies*, **14**(5), 1–314.

LAGAKIS P. & DEMETRIADIS S. (2021). Automated essay scoring : A review of the field. In *2021 International Conference on Computer, Information and Telecommunication Systems (CITS)*, p. 1–6 : IEEE.

LANDAUER T. K. & DUMAIS S. T. (1997). A solution to Plato's problem : The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, **104**(2), 211.

LEMAIRE B. & DESSUS P. (2001). A System to Assess the Semantic Content of Student Essays. *Journal of Educational Computing Research*, **24**(3), 305–320. DOI : 10.2190/G649-0R9C-C021-P6X3.

LINK S., DURSUN A., KARAKAYA K. & HEGELHEIMER V. (2014). Towards Better ESL Practices for Implementing Automated Writing Evaluation. *Calico Journal*, **31**(3).

LONGMAN (2002). *Longman Essential Activator*. Harlow : Pearson ESL.

MARTIN L., MULLER B., SUÁREZ P. J. O., DUPONT Y., ROMARY L., DE LA CLERGERIE É. V., SEDDAH D. & SAGOT B. (2020). Camembert : a tasty french language model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

OLLER, JR J. W. & HINOFOTIS F. B. (1980). Two mutually exclusive hypotheses about second language ability : factor analytic studies of a variety of language subtests.

PAGE E. B. (1966). The imminence of... grading essays by computer. *The Phi Delta Kappan*, **47**(5), 238–243.

PARSLOW N. (2015). Automated Analysis of L2 French Writing : a preliminary study. Mémoire de master. Publisher : Unpublished.

PERDUE C. (1993). Comment rendre compte de la "logique" de l'acquisition d'une langue étrangère par l'adulte. *Études de Linguistique Appliquée*, **92**(1), 8–23.

RAMESH D. & SANAMPUDI S. K. (2022). An automated essay scoring systems : a systematic literature review. *Artificial Intelligence Review*, **55**(3), 2495–2527.

RANKOVIĆ B., SMIRNOW S., JAGGI M. & TOMASIK M. J. (2020). Automated Essay Scoring in Foreign Language Students Based on Deep Contextualised Word Representations. In *LAK20-10th International Conference on Learning Analytics & Knowledge*. Issue : CONF.

REPPEN R. (2010). *Using Corpora in the Language Classroom*. Cambridge University Press.

RUDNER L. M. & LIANG T. (2002). Automated essay scoring using Bayes' theorem. *The Journal of Technology, Learning and Assessment*, **1**(2).

SHERMIS M. D., BURSTEIN J. & BURSKY S. A. (2013). Introduction to automated essay evaluation. In *Handbook of automated essay evaluation*, p. 23–37. Routledge.

TETREAULT J., BLANCHARD D. & CAHILL A. (2013). A report on the first native language identification shared task. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 48–57, Atlanta, Georgia : Association for Computational Linguistics.

UTO M. (2021). A review of deep-neural automated essay scoring models. *Behaviormetrika*, **48**(2), 459–484.

VOLLMER, J H. & CAROLL J. B. (1983). Psychometric theory and language testing. In J. W. OLLER, Éd., *Issues in language testing research*, p. 29–79. Rowley, Mass. : Newbury House. 00000.

VOLLMER, J H. & SANG F. (1983). Competing hypotheses about second language ability : a plea of caution. In J. W. OLLER, Éd., *Issues in language testing research*. Rowley, Mass. : Newbury House. 00000.

WERNER V., FUCHS R. & GÖTZ S. (2020). L1 influence vs. universal mechanisms : An SLA-driven corpus study on temporal expression. In *Learner Corpus Research Meets Second Language Acquisition*, p. 39–66. Cambridge University Press.

WILKENS R., ALFTER D., WANG X., PINTARD A., TACK A., YANCEY K. P. & FRANÇOIS T. (2022). Fabra : French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 1217–1233.

ZAGHOUANI W. (2002). AUTO-ÉVAL : vers un modèle d'évaluation automatique des textes. In *Actes du colloque des étudiants en sciences du langage*, p.1̃6, Montréal, Canada : Université du Québec à Montréal.