COMPTE-RENDU

DeepFLE : L'INTELLIGENCE ARTIFICIELLE POUR PREDIRE ET DECRIRE LE(S) NIVEAU(X) DU CECRL D'UN TEXTE

Simona Ruggia, Université Côte d'Azur, CNRS, Bases Corpus Langage (UMR 7320)

Résumé

Cette contribution se penche sur les atouts de l'Intelligence Artificielle en didactique du FLE en présentant les fonctionnalités de DeepFLE : une plateforme capable d'évaluer et de décrire le(s) niveau(x) d'un texte en français.

Mots-clés

Didactique du FLE, intelligence artificielle, deep learning, niveaux de langue, CECRL

Abstract

This contribution examines the strengths of Artificial Intelligence in teaching French as a foreign language by presenting the features of DeepFLE: a platform capable of evaluating and describing the level (s) of a text in French.

Key-words

Didactics of French as a foreign language, Artificial intelligence, Deep learning, language levels, CEFR

INTRODUCTION

L'intelligence artificielle[1] (I.A.), nouvel atout de nombreuses disciplines, offre :

> aux chercheurs en analyse de corpus [des possibilités nouvelles] en donnant à voir des représentations du texte originales, en objectivant des parcours de lecture heuristiques, en faisant émerger de nouveaux observables linguistiques. (Mayaffre et Vanni, 2021, p.10)

Les potentialités de l'I.A. nous ont permis d'envisager de nouvelles pistes de recherche en didactique du français langue étrangère (FLE) qui ont mené à la

[1] L'Intelligence artificielle est une branche de l'informatique fondamentale.

création d'un outil d'analyse de textes innovant et performant : la plateforme DeepFLE[2] qui est capable de prédire et de décrire les spécificités du ou des niveau(x) d'un texte[3] oral en français selon les échelles du *Cadre Européen Commun de Référence pour les langues* (CERCL) (Conseil de l'Europe, 2001 ; 2018). Dans cette contribution, nous présenterons la méthodologie adoptée ainsi que les fonctionnalités de la plateforme DeepFLE.

UNE METHODOLOGIE INTERDISCIPLINAIRE

Le caractère innovant et interdisciplinaire de nos recherches[4] réside dans la méthodologie adoptée qui fait dialoguer la didactique du FLE, l'Intelligence Artificielle (IA) et l'analyse des données textuelles (ADT). Pour ce qui est de l'ADT, la méthode exploitée est la lecture contrôlée et assistée par l'analyse statistique des données textuelles, que nous appelons à l'instar de Mayaffre, la *logométrie*, une méthode qui prend « une valeur heuristique plus que probatoire : interroger plutôt que prouver, interpréter autant qu'établir » (2010, p.12).

Ainsi, la didactique du FLE et notamment les ouvrages de référence tels que le CECRL (Conseil de l'Europe, 2001, 2018), les *Référentiels pour le français* (Beacco et al., 2004, 2008, 2011 ; Beacco, Porquier, 2007 ; Riba, 2016) et les manuels de FLE constituent le point de départ pour l'étude de la description des caractéristiques des textes en fonction des six niveaux de langue, allant de A1 à C2. Le *deep learning* et plus particulièrement le modèle de *deep learning* : *Text Deconvolution Saliency* (Vanni et al., 2018 ; 2020) et la logométrie enrichissent et complètent la description des niveaux des textes, mais surtout en permettent l'évaluation. D'une part, le modèle TDS « implémente l'analyse prédictive du *deep learning* à l'analyse descriptive grâce à une extraction des passages-clés » (Ruggia 2019, p.83) en fournissant « une évaluation de leur pertinence interprétative » (Vanni et al., 2018, p.460). D'autre part, la logométrie grâce à l'analyse statistique met au jour des observables linguistiques complexes susceptibles de caractériser un locuteur ou un discours.

Cette méthodologie a permis de vérifier notre hypothèse de recherche, à savoir « le TDS est capable d'extraire les caractéristiques de textes en français et, plus précisément, il est capable d'extraire les saillances qui marquent un changement de niveau selon le CECRL » (Ruggia, 2019, p.82). Pour ce faire, nous avons d'abord

[2] http://deeptext.unice.fr/FLE. Cette plateforme est en libre accès.

[3] A l'instar du CECRL, nous considérons un texte comme une « séquence discursive (orale et/ou écrite) » (Conseil de l'Europe, 2001, p.15). Actuellement, la plateforme reconnaît le niveau des textes oraux, prochainement elle reconnaîtra aussi celui des textes écrits.

[4] Ce travail a bénéficié d'une aide du gouvernement français, gérée par l'Agence Nationale de la Recherche au titre du projet Investissements d'Avenir UCAJEDI portant la référence n° ANR-15-IDEX-01.

constitué un corpus d'entraînement[5] indispensable pour l'apprentissage profond. Ce corpus constitué de six classes[6], soit 100 000 occurrences[7] minimum pour chaque classe, correspondant aux six niveaux du CECRL, comprend des textes oraux (monologues et interactions) extraits de nombreux manuels de FLE qui s'inscrivent dans l'approche actionnelle[8]. Ensuite, nous avons analysé la véridicité des résultats du TDS, en comparant les passages-clés[9] détectés pour la reconnaissance d'un ou des niveaux d'un texte, avec les inventaires des *Référentiels pour le français* (Beacco et al., 2004, 2008, 2011 ; Beacco, Porquier, 2007 ; Riba, 2016). Enfin, grâce à la logométrie nous avons cherché la distribution statistique de ces passages-clés, ce qui a permis non seulement de prouver les résultats du TDS mais aussi d'attribuer des observables linguistiques aux diverses classes de niveau (Ruggia, 2020).

LA PLATEFORME DEEPFLE

DeepFLE, premier résultat d'une recherche en cours dont nous avons brièvement illustré supra le protocole méthodologique, a été créée pour tous les acteurs du FLE, aussi bien pour les chercheurs en didactique que pour les enseignants, évaluateurs, concepteurs de manuels et apprenants. L'utilisateur peut obtenir en quelques secondes la prédiction et la description du ou des niveau(x) d'un texte oral en français.

Concrètement, il suffit de copier-coller dans la fenêtre « entrez votre texte » le texte que l'on souhaite faire analyser et de cliquer sur « détection du niveau », comme l'illustre la figure 1.

[5] Le corpus nécessaire pour le *deep learning* se nomme corpus d'entraînement ou d'apprentissage.

[6] En *deep learning,* on appelle « classes » les différentes parties du corpus à identifier.

[7] Les occurrences correspondent aux formes graphiques (mots) et aux ponctèmes (signes de ponctuation).

[8] Ces ouvrages ont été publiés entre 2005 et 2018. Pour une description détaillée du corpus, voir Ruggia et Vanni (Sous presse) et Ruggia (2020).

[9] Un passage-clé est « une unité de surcroit textométrique ; c'est-à-dire une unité dont la pertinence est calculable et l'extraction automatique » (Vanni et al. 2018, p.461).

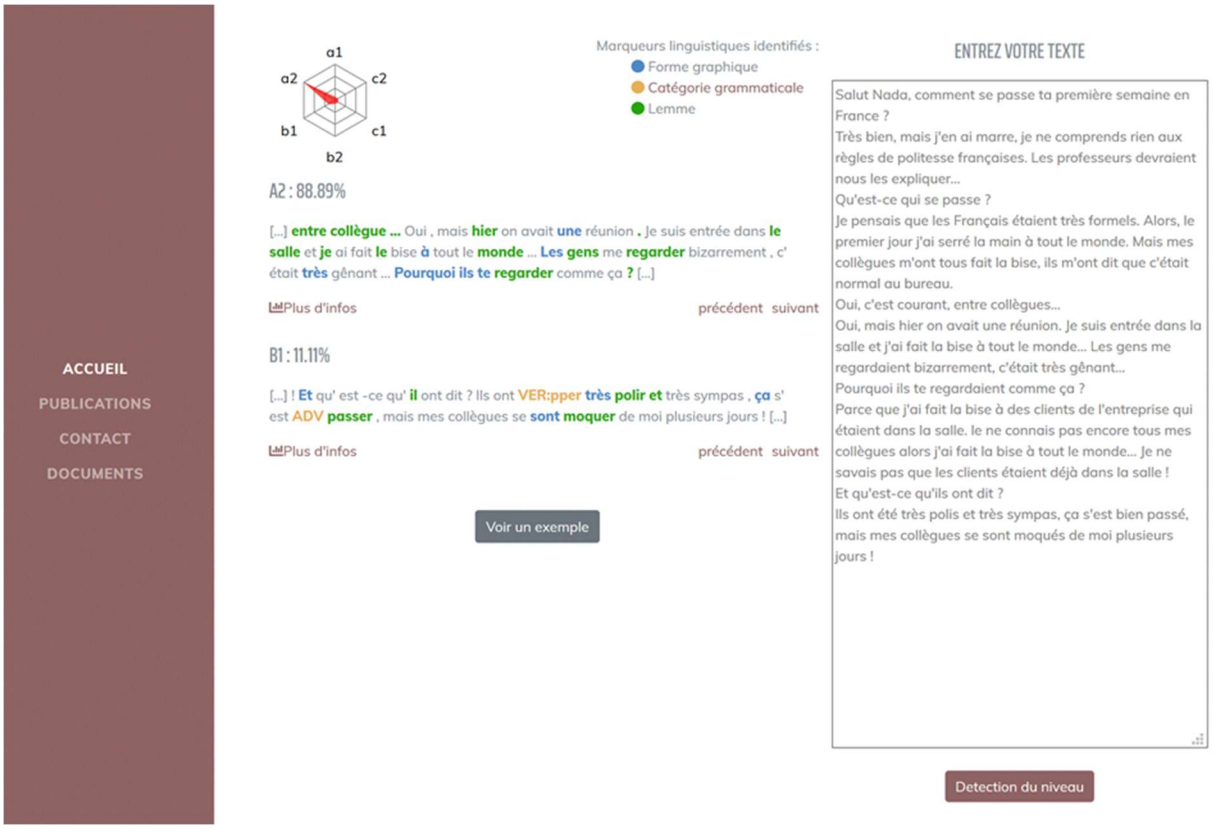

Figure 1 : Prédiction et description des niveaux d'un texte avec DeepFLE

Grâce au modèle de *deep learning* intégré, la plateforme détecte les passages-clés du texte soumis qui correspondent à un ou plusieurs niveaux. Les résultats de la prédiction s'affichent aussi bien sous forme de diagramme type radar que de score attribué. Dans le cas de la figure 1, le texte soumis est reconnu à 88.89% de niveau A2 et à 11.11% de niveau B1[10]. La description des spécificités lexicales, grammaticales et morphosyntaxiques est visible grâce aux couleurs attribuées à certains marqueurs des passages-clés :

Figure 2 : Prédiction et description d'un passage-clé de niveau A1 avec DeepFLE

Cette analyse descriptive[11] (figure 2) met en évidence la nature des marqueurs qui ont fortement contribué à la prédiction du niveau. Dans cet exemple, « et » (en

[10] Ces résultats montrent la finesse de l'analyse. Un texte a un niveau global mais sauf pour les textes de niveau A1 il comporte toujours un ou plusieurs passages d'un niveau inférieur ou supérieur.

[11] L'analyse descriptive est possible grâce à la lemmatisation préalable des textes qui a été effectuée avec TREE TAGGER.

bleu) a été détecté en tant que mot, donc pour sa forme graphique, et « il » (en vert) en tant que lemme. En orange sont indiquées les catégories grammaticales sous forme de codes, ici « VER :pper »[12] correspondant au verbe au participé passé « été ». En cliquant sur « précédent » et « suivant », on peut naviguer dans le texte, en visualisant les autres passages-clés analysés.

DeepFLE exploite la dernière version du TDS, à savoir le TDS pondéré qui « attribue un score à chaque mot (chaque *token*) pour chaque classe » (Vanni et al. 2020, p.7). Ainsi, le TDS de chaque mot ou *token* « peut être soit positif soit négatif selon la classe observée en sortie et en fonction du fait que le *token* a servi ou au contraire desservi cette classe » (*ibid.*).

Cette fonctionnalité, accessible en cliquant sur le lien « plus d'infos » au-dessous de chaque passage-clé, comme le montre l'exemple de la figure 3,

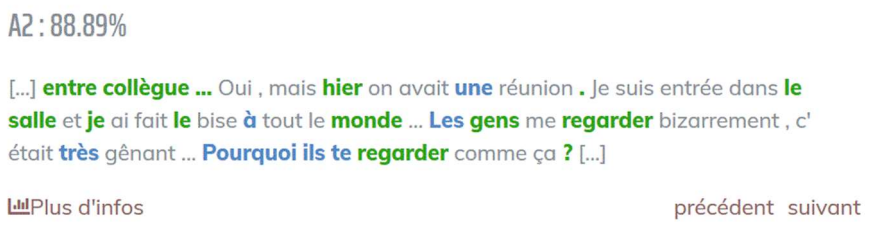

Figure 3 : Prédiction et description d'un passage-clé de niveau A2

est illustrée par un tableau (figure 4) du taux d'activation des marqueurs du passage sélectionné.

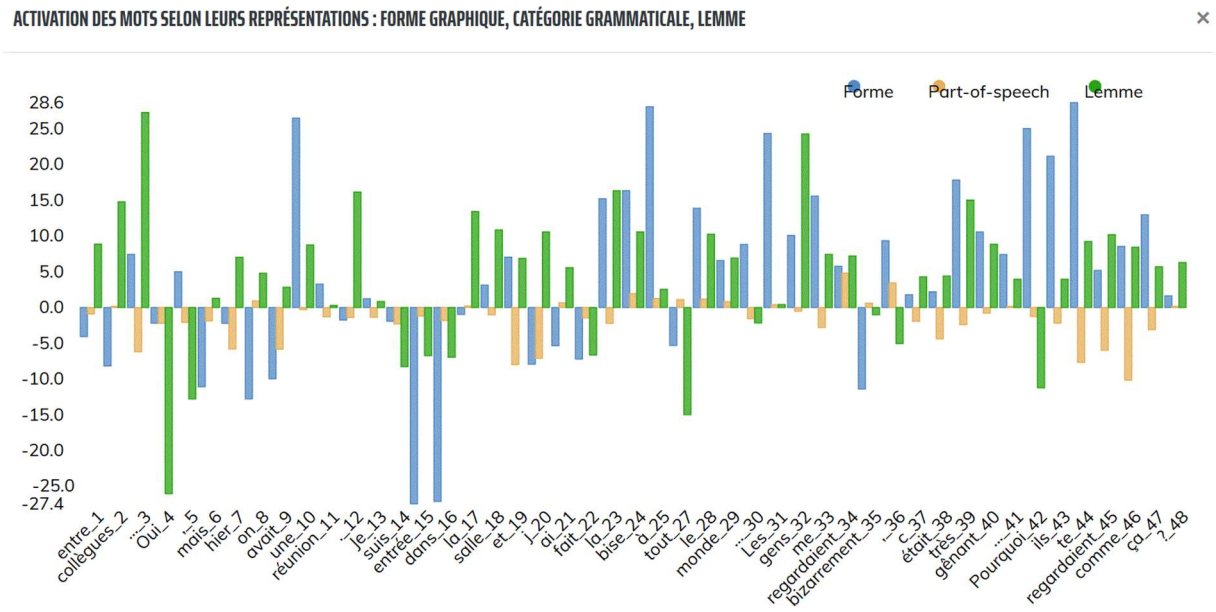

Figure 4 : Tableau du taux d'activation des marqueurs d'un passage-clé de niveau A2

[12] La liste des codes grammaticaux utilisés est consultable sur la plateforme.

BILAN ET PERSPECTIVES

La puissance de l'Intelligence Artificielle ainsi que les nombreuses recherches sur ses exploitations possibles sont aujourd'hui un atout incontournable. En didactique du FLE, son utilisation pour la prédiction et la description automatique de(s) niveau(x) d'un texte selon les échelles du CECRL a déjà fourni des résultats très satisfaisants, comme le prouve la plateforme DeepFLE, dont nous souhaitons développer les fonctionnalités et optimiser l'analyse en poursuivant nos recherches et en constituant de nouveaux corpus d'entraînement.

REFERENCES

Beacco, J.C. et al. (2004). *Niveau B2 pour le français, un référentiel*. Didier.

Beacco, J.C. et al. (dir.). (2008). *Niveau A2 pour le français, un référentiel*. Didier.

Beacco, J.C. et al. (dir.). (2011). *Niveau B1 pour le français, un référentiel*. Didier.

Beacco, J.C. et Porquier R. (2007). *Niveau A1 pour le français, un référentiel*. Didier.

Conseil de l'Europe. (2001). *Cadre Européen Commun de Référence pour les langues : apprendre, enseigner, évaluer.* Didier.

Conseil de l'Europe. (2018). *Cadre Européen Commun de Référence pour les Langues : volume complémentaire avec des nouveaux descripteurs*. https://rm.coe.int/cecr-volume-complementaire-avec-de-nouveaux-descripteurs/16807875d5

Mayaffre, D. (2010). *Vers une herméneutique matérielle numérique. Corpus textuels, logométrie et langage politique*, [Habilitation à Diriger des Recherches]. Université Nice Sophia Antipolis.

Mayaffre, D. et Vanni, L. (2021). (dir.). *L'intelligence artificielle des textes. Des algorithmes à l'interprétation.* Honoré Champion.

Riba, P. (2016). *Niveaux C1 / C2 pour le français. Eléments pour un référentiel*. Didier.

Ruggia, S. (2019). Le deep learning : un outil pour la didactique du FLE ?. *Dialettica pedagogica*. 1, 79-106.

Ruggia, S. (2020). Caractériser un texte en français : les passages-clés des niveaux A1 et A2 du CECRL. *Actes des 15èmes Journées internationales d'Analyse statistique des Données Textuelles*, 1-11. http://lexicometrica.univ-paris3.fr/jadt/JADT2020/jadt2020_pdf/RUGGIA_JADT2020.pdf.

Ruggia, S. et Vanni, L. (Sous presse). DeepFLE : la plateforme pour évaluer le niveau d'un texte selon le CECRL. *Dialogues et Cultures*.

Vanni, L. et al. (2018). Text Deconvolution Saliency (TDS): a deep tool box for linguistic analysis. Actes du 56th Annual Meeting of the Association for Computational Linguistics, 548–557. https://doi 10.18653/v1/P18-1051

Vanni, L. et al. (2020). Hyperdeep : deep learning descriptif pour l'analyse de données textuelles. *Actes des 15èmes Journées internationales d'Analyse statistique des Données Textuelles*, 1-12. http://lexicometrica.univ-paris3.fr/jadt/JADT2020/jadt2020_pdf/VANNI_CORNELI_LONGREE_MAYAFFRE_PRECIOSO_JADT2020.pdf.

L'auteure

Simona Ruggia est maître de conférences / HDR en didactique du FLE à l'Université Côte d'Azur. Elle est responsable de l'axe 3 : « Corpus et didactique des langues » de l'équipe « Logométrie. Corpus, traitements, modèles » au sein du laboratoire « Bases, Corpus, Langage » UMR 7320/CNRS/UCA. Ses travaux de recherche portent sur la didactique du FLE en l'adossant à l'étude outillée de corpus numériques et en faisant appel aux atouts de l'Intelligence Artificielle.

Courriel

simona.ruggia@univ-cotedazur.fr