Un prototype en ligne pour la prédiction du niveau de compétence en anglais des productions écrites

Thomas Gaillat, Nicolas Ballier, Annanda Sousa, Manon Bouyé, Andrew Simpkin, Bernardo Stearns, Manel Zarrouk

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition),
Traitement Automatique des Langues Naturelles (TALN, 27e édition),
Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition)
Nancy, France, 08-19 juin 2020

Un prototype en ligne pour la prédiction du niveau de compétence en anglais des productions écrites

Thomas Gaillat[1] Nicolas Ballier[2] Annanda Sousa [3] Manon Bouyé [2]
Andrew Simpkin [3] Bernardo Stearns [3] Manel Zarrouk [4]
(1) LIDILE, Université de Rennes 2, 35000 Rennes, France
(2) Insight Centre for Data analytics, NUI Galway, Irlande
(3) CLILLAC-ARP, Université de Paris, 75013 Paris, France
(4) LIPN, Université Sorbonne Paris Nord, 93430, France
thomas.gaillat@univ-rennes2.fr,nicolas.ballier@u-paris.fr,
a.defreitassousa@nuigalway.ie,mbouye@eila.univ-paris-diderot.fr,
andrew.simpkin@nuigalway.ie, zarrouk@lipn.univ-paris13.fr,
bernardo.stearns@nuigalway.ie

RÉSUMÉ

Cet article décrit un prototype axé sur la prédiction du niveau de compétence des apprenants de l'anglais. Le système repose sur un modèle d'apprentissage supervisé, couplé à une interface web.

ABSTRACT

A prototype for web-based prediction of English proficiency levels in writings.

This paper describes a proof-of-concept system focused on proficiency level prediction in learners of English. The systems relies on a supervised learning model coupled with a web interface for users.

MOTS-CLÉS : CECRL, Système d'évaluation automatique, anglais d'apprenant, complexité.

KEYWORDS: CEFR, Automatic Essay Scoring, learner English, linguistic complexity.

1 Introduction

Le système présenté [1] était destiné à servir de preuve de concept pour l'évaluation automatique du niveau de langue en anglais. Nos recherches visent à identifier les caractéristiques linguistiques et à les intégrer dans un système fondé sur l'intelligence artificielle (IA). L'objectif est de créer un système permettant d'analyser les essais des apprenants de l'anglais et de les mettre en correspondance avec des niveaux spécifiques des niveaux de langue du Cadre Européen Commun de Référence pour les langues (CECRL, European Council (2001)). Nous présentons un système fonctionnant par apprentissage supervisé. Le modèle s'appuie sur les caractéristiques des textes, indépendamment des erreurs, pour en construire une représentation multidimensionnelle sous la forme de traits critériés (Hawkins & Filipović, 2012). Les caractéristiques retenues comprennent la lisibilité et les mesures de complexité utilisées dans le domaine de l'exploration de textes et du TAL.

1. Cette démonstration synthétise en français des publications précédentes. Les détails concernant l'approche linguistique, la modélisation, les résultats en terme de précision et l'architecture ont été publiés dans (Sousa *et al.*, 2020). La chaîne de traitement est en python (3.6) et ses parties redistribuables le sont avec la licence Creative Commons.

Volume 4 : Démonstrations et résumés d'articles internationaux, pages 30–33.
HAL : hal-02768504.

2 Description du système

Le système a été entraîné sur une base de données [2] de plus de 40 000 textes (environ 3 298 343 tokens), qui ont déjà été étiquetés et annotés en niveaux (Geertzen *et al.*, 2013). A partir des 769 caractéristiques des différentes dimensions linguistiques, le meilleur modèle s'est avéré être une régression logistique. Le modèle repose sur les caractéristiques des productions, indépendamment des erreurs commises par les apprenants (par opposition à une analyse fondée sur les erreurs annotées, cf. Ballier *et al.* (2019)). Plusieurs outils (cf.Tableau 1) analysant les niveaux de complexité linguistique sont concaténés dans une chaîne de traitement pour construire une représentation multidimensionnelle des caractéristiques des essais écrits. L'évaluation du système a montré une précision de 82% (cf. Gaillat *et al.* (submitted)).

Les productions des apprenants sont souvent évaluées en fonction des trois dimensions que sont la complexité, la précision et la fluidité ("fluency") (Housen *et al.*, 2012). Les métriques proposées opérationalisent la dimension de la complexité linguistique et celle de la précision des écrits. Les métriques de complexité portent sur les dimensions syntaxiques, lexicales, discursives et psycholinguistiques sous la forme d'indices de lisibilité et d'îlots de fiabilité ("*reliability islands*" unités phraséologiques recurrentes, Dechert, Hans-Wilhelm (1983)). Les métriques de complexité syntaxique traditionnelles abordent généralement les textes de manière syntagmatique. Nous proposons de nouvelles métriques fondées sur les rapports paradigmatiques qu'entretiennent des formes linguistiques entre elles. Nous formalisons des microsystèmes linguistiques sous la forme de ratios d'usage de ces formes les unes par rapport aux autres et indépendamment des autres formes des textes. Le principe est de saisir les variations d'usage des formes par les apprenants (Gaillat *et al.*, submitted). L'outil L2SCA a été modifié en ce sens et est disponible à l'adresse du projet.

Outils	Dimension	Exemples de métriques
L2SCA modified (Lu, 2014)	Complexité syntaxique	Mean Length Sentence (MLS), Microsystèmes
LCA (Lu, 2014)	Diversité lexicale	Type Token Ratio (TTR)
TAALES (Kyle *et al.*, 2018)	Diversité et Sophistication lexicale	Fréquences de mots et écarts types dans un corpus
TAASC (Kyle *et al.*, 2018)	Sophistication syntaxique	Nombre de prépositions par groupe nominal
TAACO (Kyle *et al.*, 2018)	Complexité discursive : Cohésion textuelle	Chevauchements lexicaux, répétitions de pronoms
Pyenchant 2.0.0 (Kelly, 2016)	Erreurs d'orthographe : dictionnaire Aspell	Fréquences d'erreurs
Textstat (Bansal, 2018)	Lisibilité, Phraséologie	Indices de type Dale_Chall, fréquences de N-grams

TABLE 1 – Récapitulatif des principaux outils de la chaîne de traitement

Ce prototype est composé de deux modules. Le premier est une interface web utilisateur [3] en accès libre permettant la prédiction de niveau CECRL (Cf. Figure 1). Les apprenants saisissent leurs textes avant de recevoir un feedback sur la classe CECRL estimée et sa probabilité. Cette infrastructure est

2. Les données tirées du corpus EFCAMDAT ont été annotées indépendamment des équipes de recherche de Cambridge et d'EF Education.

3. Cf. l'URL du projet : www.clillac-arp.univ-paris-diderot.fr/projets/ulysse2019 .

composée de modules Docker (Merkel, 2014) interconnectés permettant la production des métriques et la classification par le modèle.

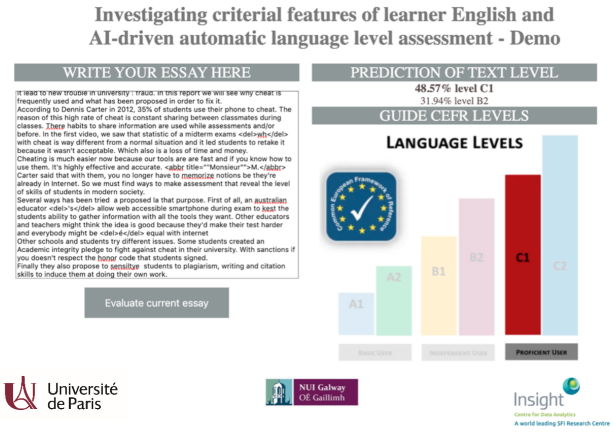

FIGURE 1 – L'interface utilisateur pour les apprenants de l'anglais

Le second module est une chaîne de traitement en ligne de commande. Au contraire de l'interface utilisateur, ce module permet de traiter des fichiers textes par lots et ainsi de créer des jeux de données représentant les valeurs des métriques pour les textes. Du fait de la mise en oeuvre d'un outil dépendant d'un corpus propriétaire, seule une version plus restreinte de la chaîne de traitement, telle que décrite dans (Sousa *et al.*, 2020) est disponible pour la communauté.

3 Développements et perspectives

Nous envisageons deux voies d'amélioration du prototype. Une première extension concerne l'ajout du traitement de la parole en ayant recours à des outils existants de retranscription automatique tels que la librairie Python SpeechRecognition (Zhang, 2017). La seconde extension concernera la variété des métriques. On explorera d'autres métriques de mesure des microsystèmes linguistiques, notamment pour l'analyse de la morphologie (Brezina & Pallotti, 2019), afin de mieux analyser les variations d'usages propres aux apprenants. En outre, les métriques seront exploitées dans le cadre de visualisations permettant de comparer des individidus avec des profils types d'apprenants. Cela nous mènera à étudier l'impact de ce type d'outil en situation pédagogique d'enseignement. Les feedbacks sur les propriétés positives (Hawkins & Filipović, 2012) de la langue peuvent avoir des conséquences sur la précision dans l'expression (*accuracy*), ce qui est important à évaluer. Ce prototype est à considérer comme un moyen d'inventer de nouveaux outils pour aider les enseignants dans leur pratique. Dans le cadre des diagnostics d'apprentissage, ils/elles bénéficieraient d'outils d'analyse faciles à utiliser, et qui objectivent les progrès de leurs apprenants.

Remerciements

Nous tenons à remercier l'Irish Council et les Ministères français des Affaires étrangères et de la Recherche pour leur soutien dans le cadre du programme de financement Ulysse PHC 2019 (projet 43121RJ). Nous sommes reconnaissants à Kristopher Kyle et Scott A. Crossley de nous avoir fourni le code source des outils TAALES, TAACO et TAASC. Nos remerciements vont également à Xiaofei Lu pour avoir fourni L2SCA et LSA.

Références

BALLIER N., GAILLAT T., SIMPKIN A., STEARNS B., BOUYÉ M. & ZARROUK M. (2019). A supervised learning model for the automatic assessment of language levels based on learner errors. In *European Conference on Technology Enhanced Learning*, p. 308–320 : Springer.

BANSAL S. (2018). Textstat Python package. Accessible sur https://github.com/shivam5992/textstat.

BREZINA V. & PALLOTTI G. (2019). Morphological complexity in written l2 texts. *Second language research*, **35**(1), 99–119.

DECHERT, HANS-WILHELM (1983). How a story is done in a second language. In C. FÆRCH & G. KASPER, Éds., *Strategies in interlanguage communication*, p. 175–195. London ; New York : Longman. OCLC : 644977107.

EUROPEAN COUNCIL (2001). *Common European Framework of Reference for Languages : Learning, teaching, assessment*. Cambridge : Cambridge University Press.

GAILLAT T., SIMPKIN A., BALLIER N., STEARNS B., SOUSA A., BOUYÉ M. & ZARROUK M. (submitted). Predicting CEFR levels in learners of English : the use of microsystem criterial features in a machine learning approach. *Journal With Anonymous Submission*.

GEERTZEN J., ALEXOPOULOU T. & KORHONEN A. (2013). Automatic Linguistic Annotation of Large Scale L2 Databases : The EF-Cambridge Open Language Database (EFCamDat). In R. T. MILLER, K. I. MARTIN, C. M. EDDINGTON, A. HENERY, N. MIGUEL, A. TSENG, A. TUNINETTI & D. WALTER, Éds., *Proceeedings of the 31st Second Language Research Forum*, Carnegie Mellon : Cascadilla Press.

HAWKINS J. A. & FILIPOVIĆ L. (2012). *Criterial Features in L2 English : Specifying the Reference Levels of the Common European Framework*. United Kingdom : Cambridge University Press.

HOUSEN A., KUIKEN F. & VEDDER I., Éds. (2012). *Dimensions of L2 performance and proficiency : complexity, accuracy and fluency in SLA*, volume 32 de *Language Learning & Language Teaching (LL&LT)*. Amsterdam, Pays-Bas, Etats-Unis d'Amérique : John Benjamins Publishing Company.

KELLY R. (2016). PyEnchant a spellchecking library for Python. Accessible sur https://pythonhosted.org/pyenchant.

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

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

MERKEL D. (2014). Docker : lightweight linux containers for consistent development and deployment. *Linux journal*, **2014**(239), 2.

SOUSA A., BALLIER N., GAILLAT T., STEARNS B., ZARROUK M., SIMPKIN A. & BOUYÉ M. (2020). From Linguistic Research Projects to Language Technology Platforms : A Case Study in learner data. In *Proceedings of the 1st International Workshop on Language Technology Platforms IWLTP 2020*, Marseille : LREC / ACL.

ZHANG A. (2017). Speech recognition (version 3.8) python library. Accessible sur https://pypi.org/project/SpeechRecognition/.