Prédire le niveau de langue d'apprenants d'anglais

Natalia Grabar, Thierry Hamon, Bert Cappelle, Cyril Grandin, Benoît Leclercq, Ilse Depraetere

▶ **To cite this version:**

Natalia Grabar, Thierry Hamon, Bert Cappelle, Cyril Grandin, Benoît Leclercq, et al.. Prédire le niveau de langue d'apprenants d'anglais. 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), 2020, Nancy, France. pp.223-231. hal-02784771v2

HAL Id: hal-02784771

https://hal.science/hal-02784771v2

Submitted on 18 Jun 2020 (v2), last revised 23 Jun 2020 (v3)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Actes de la 6e conférence conjointe Journées d'Études sur la Parole (JEP, 31e 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

Prédire le niveau de langue d'apprenants d'anglais

Natalia Grabar[1,2] Thierry Hamon[3,4] Bert Cappelle[2] Cyril Grandin[2]
Benoît Leclercq[2] Ilse Depraetere[2]
(1) CNRS, UMR 8163, F-59000 Lille, France
(2) Univ. Lille, UMR 8163 - STL - Savoirs Textes Langage, F-59000 Lille, France
(3) LIMSI, CNRS, Université Paris-Saclay, F-91405 Orsay, France
(4) Université Paris 13, Sorbonne Paris Cité, F-93430, Villetaneuse, France
{natalia.grabar;bert.cappelle;cyril.grandin}@univ-lille.fr
{benoit.leclercq;ilse.depraetere}@univ-lille.fr
thierry.hamon@limsi.fr

RÉSUMÉ

L'apprentissage de la deuxième langue (L2) est un processus progressif dans lequel l'apprenant améliore sa maîtrise au fur et à mesure de l'apprentissage. L'analyse de productions d'apprenants intéresse les chercheurs et les enseignants car cela permet d'avoir une meilleure idée des difficultés et les facilités d'apprentissage et de faire des programmes didactiques plus adaptés. Cela peut également donner des indications sur les difficultés cognitives à maîtriser les notions grammaticales abstraites dans une nouvelle langue. Nous proposons de travailler sur un corpus de productions langagières d'apprenants d'anglais provenant de différents pays et donc ayant différentes langues maternelles (L1). Notre objectif consiste à catégoriser ces productions langagières selon six niveaux de langue (A1, A2, B1, B2, C1, C2). Nous utilisons différents ensembles de descripteurs, y compris les verbes et expressions modaux. Nous obtenons des résultats intéressants pour cette catégorisation multiclasse, ce qui indique qu'il existe des différences linguistiques inhérentes entre les différents niveaux.

ABSTRACT

Predict the language level for English learners.

Second language (L2) learning is a progressive process during which the learner improves his/her language proficiency as the learning process progresses. The analysis of linguistic productions of learners attracts the attention of researchers and language teachers because this helps to get a better idea on learning difficulties and easiness, and to prepare more approriate didactic programs. This can also provide indications on cognitive difficulties to master grammatical and abstract notions in a new language. We propose to work with a corpus of language productions of English learners coming from different countries and having different mother tongues (L1). Our purpose is to categorize these language productions according to six language levels (A1, A2, B1, B2, C1, C2). We exploit different sets of descriptors, including modal verbs and expressions. We obtain interesting results for this multiclass categorization, which indicates that these language levels contain inherent linguistic features.

MOTS-CLÉS : Catégorisation supervisée, apprentissage L2, anglais, formules de lisibilité, n-grammes, verbes et expressions modaux.

KEYWORDS: Supervised categorization, L2 learning, English, readability scores, n-grams, modal verbs and expressions.

Volume 2 : Traitement Automatique des Langues Naturelles, pages 223–231.
HAL : hal-02784771.
Cette œuvre est mise à disposition sous licence Attribution 4.0 International.

1 Introduction

Les chercheurs distinguent en général deux catégories en acquisition de langues (Robertson & Ford, 2009) : l'acquisition de la première langue (L1) et de la deuxième langue (L2). L'acquisition de la première langue est un processus universel, inconscient et indépendant de la langue. Ainsi, de jeunes enfants commencent très rapidement à imiter les productions langagières de leurs parents et de l'entourage. Cependant, l'acquisition de L2 suppose la connaissance et la maîtrise de la première langue. Ce processus suppose également que la personne apprenne consciemment les éléments d'une nouvelle langue, comme le vocabulaire, les composants phonologiques, les structures grammaticales et l'écriture. L'apprentissage de L2 est donc un processus progressif dans lequel l'apprenant améliore sa maîtrise au fur et à mesure de l'apprentissage. Les productions d'apprenants L2 intéressent les chercheurs qui veulent comprendre les difficultés d'apprentissage pour une langue donnée, faire des programmes d'apprentissage plus appropriés ou pour étudier les capacités cognitives des élèves à maîtriser les notions plus ou moins abstraites, par exemple.

Les productions langagières en L2 sont étudiées de différents points de vue : étudier un aspect langagier donné (Gibbs, 1990; Moloi, 1998; Watanabe & Iwasaki, 2009; Mortelmans & Anthonissen, 2016; Murakami *et al.*, 2016; Ayoun & Gilbert, 2017; Römer, 2019), faire le parallèle entre l'apprentissage de L1 et de L2 (Laufer & Eliasson, 1993; Chenu & Jisa, 2009; Ipek, 2009; Rabinovich *et al.*, 2016), identifier automatiquement la L1 des apprenants en L2 (Jiang *et al.*, 2014; Malmasi & Dras, 2015; Nisioi, 2015) ou définir le niveau de maîtrise d'apprenants de L2 (Granfeldt & Nugues, 2007; Pilan *et al.*, 2016; Arnold *et al.*, 2018; Balikas, 2018). Les deux premières tâches sont en général étudiées manuellement par les linguistes et didacticiens, alors que les deux autres tâches attirent l'attention des chercheurs en TAL. Les travaux effectués manuellement concernent typiquement l'étude de catégories abstraites, comme la notion de modalité et l'usage de modaux. Tout d'abord, notons qu'il a été observé que, chez les apprenants d'anglais L1, l'usage de modaux apparaît à partir de 2 ans avec des modaux déontiques (comme *can*) et se diversifie progressivement avec l'apparition de modaux épistémiques (comme *must* ou *might*) vers 3 ans (Shatz & Wilcox, 1991; Papafragou, 1998; Cournane, 2015). En ce qui concerne l'apprentissage de modaux d'anglais L2, dans une étude (Gibbs, 1990), les chercheurs ont analysé la compréhension de la valeur des modaux principaux (*can, could, may, might*) par les apprenants d'anglais parlant Panjabi, une langue indienne. Les apprenants devaient définir la valeur sémantique des modaux parmi quatre valeurs possibles : capacité, permission, possibilité et possibilité hypothétique. Dans un autre travail, les chercheurs analysaient la maîtrise des fonctions grammaticale et modale des modaux chez des enfants apprenants d'anglais (Moloi, 1998). L'auteur indique entre autre qu'il existe des similarités dans l'apprentissage de l'anglais comme L1 ou L2. De rares travaux ont porté sur l'acquisition de la modalité dans d'autres langues, comme par exemple en japonais (Watanabe & Iwasaki, 2009) ou en allemand (Mortelmans & Anthonissen, 2016).

Concernant la prédiction du niveau d'apprenants, dans un travail, les productions d'apprenants de français d'origine suédoise ont été analysées du point de vue syntaxique (Granfeldt & Nugues, 2007), en mettant l'accent sur leur étiquetage et analyse syntaxique automatique. 142 descripteurs ont ainsi pu être exploités, comme par exemple le pourcentage de séquences déterminant-nom avec accord, de mots inconnus, de GNs avec accord en genre, de prépositions, de séquences nom-adjectif avec accord, d'accord sujet-verbe avec des verbes modaux et la longueur moyenne des phrases. Les productions ont ensuite été classées automatiquement en cinq stades d'apprentissage. L'utilisation des 10 meilleurs descripteurs avec C4.5 montre une F-mesure entre 0,46 et 0,53 selon les stades. Dans un autre travail, les productions d'apprenants de suédois comme L2 sont analysées de différents points de vue (lexique, syntaxe, morphologie, sémantique) grâce à leur comparaison avec des manuels de

langue (Pilan *et al.*, 2016). Il s'avère que les descripteurs lexicaux apportent le plus de gain dans la définition de la maîtrise du suédois. Actuellement, la plupart des travaux qui cherchent à définir le niveau de maîtrise de L2 sont effectués sur le corpus EFCAMDAT avec six niveaux d'apprenants (Geertzen *et al.*, 2013; Huang *et al.*, 2018). Ce corpus est construit et maintenu à l'université de Cambridge. Il contient des productions d'apprenants adultes d'anglais comme L2 et de différentes langues maternelles L1 (portugais brésilien, chinois, russe, espagnol mexicain, allemand, français, italien, arabe saoudien, taiwanais et japonais). Plusieurs travaux qui cherchent à définir le niveau de maîtrise de L2 ont été effectués dans le cadre de la campagne d'évaluation de la conférence CAp en 2018 [1], avec une mesure d'évaluation spécifique basée sur l'erreur et une matrice de coût spécifique (Ballier *et al.*, 2018). L'erreur est calculée comme $E = \frac{1}{n}\sum_{i=1}^{6}\sum_{j=1}^{6} C_{ij}N_{ij}$, où N est la matrice de confusion (N_{ij} compte le nombre de fois où un exemple de la classe i a été classé j), n le nombre d'exemples classés et C la matrice de coût, calculée comme une mesure d'entropie croisée pondérée entre les classes. Les probabilités prises en compte sont les probabilités d'apparition des classes telles qu'elles apparaissent dans les échantillons d'apprentissage et de test. Les poids des classes et des erreurs ont été donnés par les experts du domaine pour prendre en compte l'importance de chacune des classes. Sur les 14 équipes participantes de cette compétition, nous avons pu trouver la description de deux systèmes seulement. Le système gagnant (Balikas, 2018) met en place une large palette de descripteurs (formules de lisibilité, modèles probabilistes de langues, plongements lexicaux, topic model, étiquettes morpho-syntaxiques, n-grammes de mots). Avec Gradient Boosted trees et tous les descripteurs, le système obtient 98,2 d'*accuracy* avec l'*erreur* de 4,97. Un autre système (Arnold *et al.*, 2018) utilise une palette plus restreinte de descripteuts (diversité lexicale, complexité syntaxique, formules de lisibilité). Gradient Boosted trees est utilisé pour effectuer la classification binaire entre deux niveaux de langues. Selon les paires de niveaux, l'*AUC*, qui est la seule métrique présentée dans ce travail, varie entre 0,587 et 0,916.

Nous travaillons avec les productions langagières d'apprenants d'anglais comme L2 provenant de différents pays et de différentes langues L1. Notre objectif consiste également à prédire le niveau de ces apprenants sur la base de leurs productions écrites. Nous exploitons différents types de descripteurs, y compris les verbes et expressions modaux. Comme les modaux sont des catégories abstraites de la langue, ils peuvent être plus difficiles à acquérir et maîtriser par les apprenants. Selon les observations des chercheurs, lors des étapes initiales d'apprentissage d'anglais L2, l'usage de modaux est pauvre qu'il s'agisse de leur fréquence et diversité, les modaux déontiques sont dominants, certains modaux peuvent être sous-utilisés ou sur-utilisés (Biewer, 2011). Par ailleurs, il existe peu de travaux qui s'intéressent à cette catégorie d'unités linguistiques (Jabbari & Sedghi, 2015) chez les apprenants d'anglais L2. Nous proposons donc d'étudier le rôle que les modaux pourraient jouer dans la détection automatique du niveau de langue, à côté d'autres descripteurs.

2 Approche et données

Nous travaillons avec le corpus EFCAMDAT [2] (Geertzen *et al.*, 2013; Huang *et al.*, 2018). La partie exploitée du corpus contient 27 287 productions contenant presque 2M occurrences de mots. Une production correspond au texte écrit par un apprenant d'anglais pour répondre à une question donnée. Une production peut contenir une ou plusieurs phrases et véhiculer des idées plus ou moins complexes, en fonction de la maîtrise de la langue. La figure 1 présente des exemples de productions.

1. http://cap2018.litislab.fr/competition.html
2. https://corpus.mml.cam.ac.uk/efcamdat2/

C2	*In France, robots are commonly used in industrial fields. They replace humans for heavy duties, dangerous or repetitive tasks. Also, we find more and more robots in houses, like vacums that move by themselves, same idea for lawn mawers. But in my opinion, these kind of robots are making people lazy. I admit that we can save time by using them, but on the other hand, do we really use this spare time doing good stuff ? I am not sure.*
C1	*In general I admire succesfull persons who keep their humility and stay simple. This kind of person are not self oriented, they have all they could wish and they are not selfish. Being focused on the others and attentive to his/her friends, relatives and colleagues give to succesful something more, making them more human and noble.*
B2	*I don't practise any extreme sport. I am definitely not a thrill-seeker. I'm scared of deep water, I have the vertigo... However, I am awestruck by people who pratise this kind of activities. They still search to overtook their limits. I can understand them. When I was younger, I was passionate about pacey carousels. It was fantastic to fly in the air even though I was just in a seat with a harness. I always got in them alone because my friends were so afraid. I could feel such a rush. One time, I tried one carousel who was going very fast. I felt like I was going to pass out. Now I always wimp out.*
B1	*I saw this girl at the swimming pool. I found her very attractive. She had a necklace with her first name. I introduced me and I asked her few questions to know her better. After we met, I invited her for a date the next day. Unfortunately, I was late. She left when I arrived. The next day, I tried my luck again and we had a new date for a tennis party. Of course I was careful to be on time. We played tennis but the next time she left on holiday for a week. We phoned us during this week and I welcomed her for her back at the station. It was 27 years ago. We got married since.*
A2	*Dear friends, excuse me but I couldn't come to a mariage. I feel awful and I think that I'm sick. I have a cold, a headache and a fever. I went to the doctor and I should stay in bed a few days. The pharmacist gave me some medecine. I'm sorry, let' go and have fun.*
A1	*Hi Sue, Sorry, I'm busy. Right now I'm working in my office. Then, I have to clean my house. And after, cook for my parents. See you another day. xoxo.*

FIGURE 1 – Exemples de productions d'apprenants d'anglais pour chaque niveau

Ces productions sont catégorisées selon six niveaux allant de A1 (le plus bas niveau de maîtrise) à C2 (le plus haut niveau de maîtrise) selon CECRL, un cadre européen de référence pour les langues. Le tableau 1 indique le nombre de productions par niveau, le nombre d'occurrences de mots que cela représente et la longueur moyenne des productions pour un niveau donné. Nous pouvons voir que les niveaux C, et surtout le niveau C2, contiennent très peu de productions. La longueur moyenne des productions tous niveaux confondus est de 68 mots. Cependant, la longueur moyenne des productions par niveau montre une tendance à augmenter avec l'amélioration de la maîtrise de l'anglais.

Pour la prédiction du niveau d'anglais, nous exploitons les algorithmes d'apprentissage de la bibliothèque Scikit-learn [3] (Pedregosa *et al.*, 2011). Plusieurs descripteurs sont exploités :

— un ensemble de 59 formules et indicateurs de lisibilité fournis avec le corpus EFCAMDAT. Il s'agit de formules classiques de lisibilité, comme celles proposées dans les travaux existants (Flesch, 1948; McLaughlin, 1969; Kincaid *et al.*, 1975). Leur calcul est basé le plus souvent sur des indicateurs de surface des textes (nombre de mots, nombre de phrases, nombre de

3. https://scikit-learn.org/stable/

Niveau	*# productions*	*# occurrences*	*Moyenne d'occurrences*
A1	11 346	432 442	38
A2	7 680	503 246	66
B1	5 383	511 356	95
B2	2 337	308 433	132
C1	491	80 786	165
C2	50	8 317	166
Total	27 287	1 844 580	68

TABLE 1 – Nombre de productions selon les six niveaux d'anglais et leurs taille

syllabes, longueur moyenne de phrases, taille moyenne de mots, etc.). Ces formules associent les productions avec les niveaux scolaires et universitaires et reflètent la complexité des textes. Les formules de lisibilité fournissent des scores associés aux productions ;

— les n-grammes de 2, 3 et 4 mots (formes) provenant de deux corpus de référence : COCA (Corpus of Contemporary American English) (Davies, 2010) et BNC (British National Corpus) (Burnard, 2000). COCA contient plus de 560M mots et couvre la période entre 1990 et 2017. BNC contient 100M mots provenant de productions écrites et orales produites entre les années 1980 et 1990. La motivation d'utiliser des n-grammes vient du fait que les cooccurrences de mots peuvent également indiquer le niveau de connaissance d'une langue. Par exemple, on dit *commit atrocities* plutôt que *do atrocities* ou *perform atrocities*. Ainsi, plus on maîtrise une langue et plus on a tendance à utiliser des expressions standards, comme par exemple l'usage d'expressions plus ou moins figées, de prépositions ou de temps verbaux. L'utilisation de ces descripteus va dans le même sens que l'exploitation de manuels de langue effectué dans un travail existant (Pilan *et al.*, 2016). Les n-grammes sont extraits de la même manière à partir de corpus et à partir de chaque production, ce qui permet de calculer ensuite les n-grammes communs (en nombre et pourcentage) ;

— un ensemble de 17 verbes modaux (*may, might, can, could, shall, should, will, would, must, have to, got to, need to, be supposed to, had better, be allowed to, be able to*), qui véhiculent les caractéristiques modales principales en anglais. Il s'agit de verbes modaux utilisés le plus souvent par les locuteurs natifs de la langue. Les travaux existants se focalisent le plus souvent sur l'emploi de ces verbes modaux (Gibbs, 1990; Moloi, 1998; Saeed, 2009; Elturki & Salsbury, 2016). Comme déjà indiqué, nous pensons que, comme les valeurs modales sont des notions abstraites, leur maîtrise et utilisation peuvent être indicatives du niveau d'avancement dans l'apprentissage de la langue ;

— un ensemble d'autres expressions modales (verbes, adjectifs, noms, adverbes) qui véhiculent une sémantique similaire, comme par exemple *possible, probably* ou *seem*.

Pour les modaux, nous calculons la fréquence de leur utilisation dans les productions.

Nous exploitons ces ensembles de descripteurs séparément pour voir leur pertinence pour la tâche mais aussi en combinaison car nous pensons que les niveaux de langue correspondent aux catégories complexes et reposent sur différents aspects liés à la maîtrise de la langue.

Notre tâche consiste donc à effectuer une catégorisation multiclasse et à assigner les productions écrites d'apprenants à l'un des six niveaux de CECRL. Nous utilisons plusieurs algorithmes d'apprentissage supervisé avec leurs paramètres par défaut en validation croisée à 10 plis. Les résultats sont évalués avec trois mesures standards : précision P, rappel R et F-mesure F dans leur version macro au niveau des catégories.

3 Détection du niveau d'apprenants

	DT			*RF*			*SVM*		
Descripteurs	*P*	*R*	*F*	*P*	*R*	*F*	*P*	*R*	*F*
Lisibilité (Lis)	0,63	0,63	0,63	0,67	0,59	0,61	0,67	0,65	0,66
BNC	0,53	0,53	0,53	0,61	0,57	0,59	0,47	0,43	0,44
COCA	0,53	0,54	0,54	0,78	0,58	0,59	0,49	0,44	0,45
17 modaux	0,35	0,28	0,28	0,35	0,29	0,29	0,32	0,26	0,25
Autres modaux	0,25	0,19	0,15	0,24	0,19	0,15	0,20	0,19	0,14
Lis+17 modaux	0,63	0,64	0,64	0,63	0,59	0,60	0,69	0,67	0,68
Lis+autres modaux	0,64	0,63	0,63	0,62	0,58	0,59	0,68	0,66	0,67
Lis+tous les modaux	0,63	0,63	0,63	0,62	0,57	0,59	0,70	0,67	0,69
Tous+BNC	0,71	0,71	0,71	0,81	0,66	0,69	**0,74**	**0,70**	**0,72**
Tous+COCA	0,71	0,71	0,71	0,86	0,66	0,69	0,73	0,69	0,71

TABLE 2 – Résultats de catégorisation selon les descripteurs exploités (version macro des mesures)

Les résultats globaux de quelques expériences obtenus avec trois algorithmes (arbres de décision DT, RandomForest RF et SVM linéaire SVM) sont présentés dans le tableau 2. Nous avons plusieurs ensembles de descripteurs : différents descripteurs utilisés séparément (*lisibilité, BNC, COCA, 17 modaux, autres modaux*) et leurs combinaisons (formules de lisibilité avec 3 ensembles de modaux (*17 modaux, autres modaux* et *tous les modaux*) et la combinaison de tous les descripteurs (*tous+BNC* et *tous+COCA*)). Nous voyons que tous les algorithmes montrent de meilleurs résultats avec la combinaison de tous les descripteurs. SVM se détache des autres algorithmes lorsqu'il est exploité avec les scores de lisibilité et les combinaisons de descripteurs, alors que DT et RF montrent aussi de bons résultats avec les n-grammes de mots. Les 17 modaux principaux permettent de catégoriser correctement un peu moins d'un tiers des productions (F-mesure entre 0,25 et 0,29). Les autres expressions modales, sans doute parce qu'elles sont utilisées moins fréquemment par les apprenants, montrent les performances les plus faibles (F-mesure entre 0,14 et 0,15). L'exploitation de modaux avec les formules de lisibilité améliore la F-mesure obtenue avec les formules de lisibilité seules de 0,20 points avec DT et de 0,25 points avec SVM. Cela indique donc que les modaux apportent des informations importantes sur le niveau de maîtrise de la langue.

Le tableau 3 indique les résultats par niveau de langue obtenues avec SVM et tous les descripteurs combinés avec les n-grammes BNC : 0,72 de F-mesure macro (pour information, cela correspond à 0,82 de F-mesure micro). Le niveau A1 montre les résultats les plus élevés (F-mesure de 0,89), ce qui peut être dû aux facteurs quantitatifs (grand nombre de productions) et qualitatifs (les apprenants ont

Niveau	*P*	*R*	*F*
A1	0,89	0,90	0,89
A2	0,76	0,76	0,76
B1	0,78	0,79	0,79
B2	0,79	0,75	0,77
C1	0,72	0,66	0,69
C2	0,47	0,36	0,40

TABLE 3 – Catégorisation par niveau (SVM et tous les descripteurs avec les n-grammes de BNC)

Ref/Predit	*A1*	*A2*	*B1*	*B2*	*C1*	*C2*	
A1	**10 154**	1 045	131	13	3	0	11 346
A2	1 238	**5 827**	576	36	1	2	7 680
B1	80	713	**4 266**	307	17	0	5 383
B2	12	31	438	**1 763**	85	8	2 337
C1	4	17	29	106	**325**	10	491
C2	1	2	2	8	19	**18**	50
tous	11 489	7 635	5 442	2 233	450	38	27 287

TABLE 4 – Matrice de confusion (SVM et tous les descripteurs avec les n-grammes de BNC)

des productions très éloignées de la langue standard et des productions d'autres niveaux plus avancés). Le niveau C2 montre les performances les plus pauvres, sans doute à cause du très faible nombre de productions. Le niveau C1 a une F-mesure de 0,69. Les trois autres niveaux (A2, B1 et B2) montrent une F-mesure entre 0,76 et 0,79. Le tableau 4 présente la matrice de confusion de la même expérience. Nous voyons en diagonal le nombre de productions associées correctement aux niveaux de langue.

Il est difficile de comparer nos résultats avec le système de Arnold *et al.* (2018) car nous n'effectuons pas le même type de catégorisation (bi-classe *vs* multiclasse). Nos résultats (valeurs macro) sont inférieurs à ceux de (Balikas, 2018) (98,2 d'*accuracy* mais il n'est pas indiqué s'il s'agit de valeurs micro ou macro). Ce travail utilise les descripteurs différents des nôtres. Nous préférons cependant exploiter les descripteurs observés directement dans les productions d'apprenants (n-grammes, modaux...) car nous pensons que les modèles probabilistes ou inductifs, comme les plongements lexicaux, ne reflettent pas forcément la compétence d'un apprenant donné. À notre avis, ces modèles correspondent à la performance collective des apprenants. Par exemple, les plongements lexicaux peuvent grouper ensemble les verbes comme *commit, perform* et *do*. Cependant, si un apprenant n'utilise pas le bon verbe dans une expression comme *commit atrocities*, les n-grammes de mots sont plus susceptibles de refléter correctement son niveau de langue que les plongements lexicaux.

4 Conclusion

Nous avons présenté quelques expériences de prédiction du niveau de langue des apprenants d'anglais sur la base de leurs productions écrites. Nous exploitons pour ceci plusieurs ensembles de descripteurs : formules de lisibilité classiques, n-grammes de mots et expressions et verbes modaux. La catégorisation montre des résultats intéressants et souligne l'importance des modaux pour cette tâche. Ces résultats peuvent être améliorés par d'autres expériences (ajout d'autres descripteurs et de leurs combinaisons, exploitation d'autres algorithmes). Par ailleurs, le rôle de notions plus abstraites, comme les valeurs modales, pourra être étudié encore plus en détail dans les travaux futurs.

Remerciements

Cette publication s'inscrit dans le projet *REM (Re-thinking English Modal Constructions)* financé par l'ANR franco-suisse sous la référence ANR-16-CE93-0009. Nous remercions les relecteurs pour leurs remarques constructives.

Références

ARNOLD T., BALLIER N., GAILLAT T. & LISSÓN P. (2018). Predicting CEFRL levels in learner english on the basis of metrics and full texts. In *Conférence sur l'Apprentissage Automatique (CAp)*, p. 31–38.

AYOUN D. & GILBERT C. (2017). *The acquisition of modal auxiliaries in English by advanced Francophone learners*, In M. HOWARD & P. LECLERCQ, Éds., *Tense-Aspect-Modality in a Second Language : Contemporary perspectives*, p. 183–212.

BALIKAS G. (2018). Lexical bias in essay level prediction. In *CAp*, p. 1–5.

BALLIER N., CANU S., GAILLAT T., GASSO G., PETITJEAN C. & RAKOTOMAMONJY A. (2018). *Appel à participation à la compétition « my taylor is rich » de CAp 2018. Prédiction du niveau en anglais à partir de production écrite d'apprenants*. Rapport interne, CAP 2018.

BIEWER C. (2011). *Modal auxiliaries in second language varieties of English : A learner's perspective*, In J. MUKHERJEE & M. HUNDT, Éds., *Exploring second-language varieties of English and learner Englishes : Bridging a paradigm gap*, p. 7–33. John Benjamins : Amsterdam.

BURNARD L. (2000). *The British National Corpus Users Reference Guide*. Rapport interne, Oxfort university. http://www.natcorp.ox.ac.uk/docs/userManual/.

CHENU F. & JISA H. (2009). Reviewing some similarities and differences in L1 and L2 lexical development. *Acquisition et interaction en langue étrangère*, (1), 1–22.

COURNANE A. (2015). *Modal development : Input-divergent L1 acquisition in the direction of diachronic reanalysis*. Thèse de doctorat, University of Toronto, Toronto, Canada.

DAVIES M. (2010). The corpus of contemporary american english as the first reliable monitor corpus of English. *Literary and Linguistic Computing*, **25**(4), 447–65.

ELTURKI E. & SALSBURY T. (2016). A cross-sectional investigation of the development of modality in English language learners' written narratives : A corpus-driven study. *Issues in Applied Linguistics*, **20**(1), 51–72.

FLESCH R. (1948). A new readability yardstick. *Journ Appl Psychol*, **23**, 221–233.

GEERTZEN J., ALEXOPOULOU T. & KORHONEN A. (2013). Automatic linguistic annotation of large scale L2 databases : The EF-Cambridge open language database (EFCAMDAT). In *31st Second Language Research Forum (SLRF)*.

GIBBS D. A. (1990). Second language acquisition of the English modal auxiliaries can, could, may, and might. *Applied Linguistics*, **11**(3), 297–314.

GRANFELDT J. & NUGUES P. (2007). Évaluation des stades de développement en français langue étrangère. In *TALN*, p. 1–10.

HUANG Y., MURAKAMI A., ALEXOPOULOU T. & KORHONEN A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, **23**(1), 28–54.

IPEK H. (2009). Comparing and contrasting first and second language acquisition : Implications for language teachers. *English Language Teaching*, **2**(2), 155–163.

JABBARI A. & SEDGHI M. (2015). Acquisition of English modality by Persian EFL learners. *International Journal of Educational Investigations*, **2**(5), 23–45.

JIANG X., GUO Y., GEERTZEN J., DORA ALEXOPOULOU AND L. S. & KORHONEN A. (2014). Native language identification using large, longitudinal data. In *LREC*, p. 1–4.

KINCAID J., FISHBURNE R., ROGERS R. & CHISSOM B. (1975). *Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel*. Rapport interne, Naval Technical Training, U. S. Naval Air Station, Memphis, TN.

LAUFER B. & ELIASSON S. (1993). What causes avoidance in L2 learning. L1-L2 difference, L1-L2 similarity or L2 complexity ? *Studies in Second Language Acquisition*, (15), 35–48.

MALMASI S. & DRAS M. (2015). Large-scale native language identification with cross-corpus evaluation. In *Annual Conference of the North American Chapter of the ACL*, p. 1403–1409.

MCLAUGHLIN G. H. (1969). SMOG grading – a new readability formula. *Journal of reading*, **12**(8), 639–646.

MOLOI F. (1998). Acquisition of modal auxiliaries in English L2. *Southern African Journal of Applied Language Studies*, **6**(2), 1–22.

MORTELMANS T. & ANTHONISSEN L. (2016). *German modals in second language acquisition : A constructionist approach*, In A. STEFANOWITSCH & T. HERBST, Éds., *Yearbook of the German Cognitive Linguistics Association*, p. 9–30.

MURAKAMI A., MICHEL M., ALEXOPOULOU T. & MEURERS D. (2016). Analyzing learner language in task contexts : A study case of linguistic complexity and accuracy in EFCAMDAT. In *European Second Language Association Conference*.

NISIOI S. (2015). Feature analysis for native language identification. In *CICLING*, p. 1–15.

PAPAFRAGOU A. (1998). The acquisition of modality : Implications for theories of semantic representation. *Mind and Language*, **13**(3), 370–399. DOI : 10.1111/1468-0017.00082.

PEDREGOSA F., VAROQUAUX G., GRAMFORT A., MICHEL V., THIRION B., GRISEL O., BLONDEL M., PRETTENHOFER P., WEISS R., DUBOURG V., VANDERPLAS J., PASSOS A., COURNAPEAU D., BRUCHER M., PERROT M. & DUCHESNAY E. (2011). Scikit-learn : Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830.

PILAN I., VOLODINA E. & ZESCH T. (2016). Predicting proficiency levels in learner writings by transferring a linguistic complexity model from expert-written coursebooks. In *Int Conf on Computational Linguistics*, p. 2101–2111.

RABINOVICH E., NISIOI S., ORDAN N. & WINTNER S. (2016). On the similarities between native, non-native and translated texts. In *Annual Meeting of the Association for Computational Linguistics*, p. 1870–1881.

ROBERTSON K. & FORD K. (2009). *Language Acquisition : An Overview*. Rapport interne, Colorin Colorado.

RÖMER U. (2019). A corpus perspective on the development of verb constructions in second language learners. *International Journal of Corpus Linguistics*, **24**(3), 270–292.

SAEED A. T. (2009). Arab EFL learners' acquisition of modals. *Research in Language*, **7**, 75–98.

SHATZ M. & WILCOX S. (1991). *Constraints on the acquisition of English modals*, In S. GELMAN & J. BYRNES, Éds., *Perspectives on language and thought*, p. 319–353. Cambridge University Press : Cambridge.

WATANABE S. & IWASAKI N. (2009). *The Acquisition of Japanese Modality during Study Abroad*, In B. PIZZICONI & M. KIZU, Éds., *Japanese Modality*, p. 231–258.