Validation interne et externe d'indices phraséologiques pour l'évaluation automatique de textes rédigés en anglais langue étrangère

Yves Bestgen

Centre for English Corpus Linguistics, Université catholique de Louvain
B-1348 Louvain-la-Neuve Belgique yves.bestgen@uclouvain.be

RÉSUMÉ. *Cette étude s'inscrit dans le champ de l'évaluation automatique de textes rédigés en langue étrangère, un domaine d'application du TALP en didactique très actif. Elle vise à évaluer d'une manière approfondie l'utilité d'y prendre en compte des mesures automatiques de la compétence phraséologique en raison de son importance lors de l'apprentissage d'une langue étrangère. Les indices phraséologiques employés sont obtenus par l'analyse de tous les bigrammes de mots présents dans un texte d'apprenants auxquels des scores d'association sont attribués sur la base d'un grand corpus natif. Les analyses effectuées sur trois ensembles de textes indiquent que ces indices apportent une information utile pour estimer la qualité des textes et que celle-ci est différente de celle apportée par des mesures lexicales plus simples. Une validation externe, réalisée en apprenant et en testant le modèle prédictif sur des ensembles de textes très différents, montre que ces indices présentent un degré de généralisation important.*

ABSTRACT. *This study belongs to the field of automatic scoring of texts written in a foreign language. It aims to evaluate in depth the utility to take into account automatic measures of the phraseological competence. The phraseological features were obtained by assigning to each bigram in a text several association scores computed on the basis of a large native corpus. The analyzes performed on three datasets indicate that the phraseological features are useful for estimating text quality and that they are complementary to simpler lexical measures. External validation, carried out by learning and testing the predictive model on very different sets of texts, shows that these indices have a high degree of generalizability.*

MOTS-CLÉS : *phraséologie*$_1$, *mesure d'association*$_2$, *expression conventionnelle*$_3$.

KEYWORDS: *phraseology*$_1$, *association measure*$_2$, *formulaic sequence*$_3$.

TAL. Volume 57 – n° 3/2016, pages 93 à 117

1. Introduction

Parmi les nombreuses applications du traitement automatique des langues dans le domaine de l'éducation et de l'enseignement, l'évaluation automatique de textes occupe une place importante comme l'attestent les systèmes commerciaux disponibles sur le marché depuis de nombreuses années tels que l'Intelligent Essay Assessor, e-Rater ou encore IntelliMetric (Ramineni et Williamson, 2013). Utilisés le plus souvent comme deuxième évaluation en supplément d'un évaluateur humain, ces systèmes facilitent le déploiement de tests standardisés comme le *Graduate Record Examination* et permettent l'évaluation de la compétence rédactionnelle en réduisant les moyens humains nécessaires (Williamson *et al.*, 2012). De tels systèmes sont particulièrement utiles dans le domaine de l'apprentissage d'une langue étrangère où être capable de rédiger un texte de qualité est un des objectifs principaux de la formation (Weigle, 2013). Dans ce champ, l'approche classique s'appuie sur des indices linguistiques plus ou moins fortement corrélés avec la qualité du texte. Le prototype de ce type d'approche est certainement e-Rater (et Criterion), développé par l'Educational Testing Service (ETS), qui emploie des caractéristiques lexicales (longueur des mots), stylistiques (phrases très courtes ou emploi du passif) et structurelles (présence d'une introduction et d'une conclusion) ainsi que la présence d'erreurs syntaxiques (mauvaise préposition, problème d'accord sujet-verbe...), typographiques (ponctuation, emploi de majuscules) et orthographiques (Ramineni et Williamson, 2013).

Dans ces systèmes d'évaluation automatique, les expressions polylexicales (Baldwin et Kim, 2010), aussi appelées unités préformées, sont très largement négligées. Tout au plus peut-on mentionner qu'une variable indicatrice de leur emploi adéquat dans un texte a été récemment incluse dans e-Rater (version 10.1), mais son utilité pour estimer la qualité d'un texte est très limitée. Elle n'a reçu qu'un poids très faible dans le modèle prédictif et Higgins *et al.* (2015, p. 597) précisent que « *the new feature showed only very slight improvements in the previously established empirical performance of e-rater [...], but there was no evidence of degradation compared to the earlier version of the engine* ».

L'intérêt limité et le peu d'efficacité des expressions polylexicales dans les systèmes d'évaluation automatique de textes rédigés par des apprenants d'une langue étrangère sont particulièrement étonnants pour deux raisons. Tout d'abord, leur importance dans l'emploi du langage est bien établie comme l'ont montré les recherches en phraséologie, branche de la linguistique qui étudie la structure, le sens et l'emploi de ces expressions (Cowie, 1994 ; Pawley et Syder, 1983 ; Sinclair, 1991). Si certaines de ces séquences se signalent par leur figement et leur opacité, la très grande majorité d'entre elles sont des manières habituelles de s'exprimer (Smiskova *et al.*, 2012). On peut citer à titre d'exemple : *cell phone* et *mobile phone, dramatic increase, committed suicide, depend on, out of, such as, by the way, first of all* ou encore *as far as I know*. La maîtrise de ces expressions est une composante majeure de l'apprentissage d'une deuxième langue. Elle permet aux apprenants de communiquer même lorsque leurs moyens linguistiques sont limités et d'apparaître comme plus avancés qu'ils ne le sont en termes de complexité (« *the ability to use a wide and varied range of sophis-*

ticated structures and vocabulary in the L2 ») et d'exactitude (« *the ability to produce target-like and error-free language* »), deux des trois dimensions majeures de l'évaluation de l'apprentissage et de la maîtrise d'une langue étrangère (Housen *et al.*, 2012, p. 2 ; Myles, 2012, p. 71). De plus, le nombre d'expressions phraséologiques dans un texte d'apprenants est un excellent prédicteur du niveau de compétence globale dans la langue à apprendre (Forsberg, 2010 ; Stengers *et al.*, 2011 ; Verspoor *et al.*, 2012).

Ensuite, ces mêmes expressions ont fait l'objet d'une série d'études en linguistique computationnelle visant à détecter automatiquement des erreurs dans leur emploi en raison des difficultés que celles-ci posent aux apprenants d'une langue étrangère (Higgins *et al.*, 2015). L'objectif principal de ces travaux est de pouvoir détecter des erreurs potentielles afin de les signaler à l'auteur du texte et de lui suggérer une autre formulation (Liu *et al.*, 2009 ; Wible *et al.*, 2003 ; Wu *et al.*, 2010). À première vue, le pas à effectuer pour intégrer ces informations dans l'évaluation semble petit.

La raison principale de leur négligence dans les systèmes d'évaluation réside probablement dans la difficulté à mettre en œuvre des mesures automatiques capables de prendre en compte la très large diversité des expressions conventionnelles. D'une part, les procédures automatiques de détection d'erreurs, souvent focalisées sur un type particulier d'expressions, ne s'intéressent pas aux séquences non erronées. D'autre part, dans presque toutes les études linguistiques, l'identification des séquences conventionnelles est effectuée manuellement, au moyen de dictionnaires ou en demandant l'avis de locuteurs natifs, une décision complexe et coûteuse en temps. Un exemple particulièrement illustratif est donné par l'étude de Santos *et al.* (2012). Ces auteurs ont montré qu'une procédure d'apprentissage supervisé sélectionnait parmi un grand nombre de mesures linguistiques un indice phraséologique comme un des meilleurs prédicteurs de la qualité d'un texte, mais cet indice avait été obtenu au moyen d'une annotation des textes par des experts (Verspoor *et al.*, 2012). Une deuxième raison est que la prise de conscience de l'importance de la phraséologie dans l'enseignement d'une langue étrangère et dans l'évaluation des compétences d'apprenants ne s'est réellement développée que récemment (Cavalla, 2009 ; Myles, 2012), même si des plaidoyers en ce sens ont été formulés bien plus tôt (Pawley et Syder, 1983). Tout particulièrement, il est symptomatique de constater que dans l'échantillon représentatif de quarante études publiées entre 1995 et 2008, analysé par Bulté et Housen (2012) afin de déterminer comment la complexité est mesurée en apprentissage d'une langue étrangère, aucune étude ne mentionne la dimension phraséologique du langage.

Récemment, toutefois, une procédure automatique pour évaluer l'emploi d'expressions conventionnelles dans des textes rédigés par des apprenants a été proposée (Durrant et Schmitt, 2009 ; Granger et Bestgen, 2014 ; Somasundaran et Chodorow, 2014 ; Somasundaran *et al.*, 2015). Elle repose sur l'analyse des n-grammes de mots présents dans un texte d'apprenants auxquels un score d'association collocationnelle, comme l'information mutuelle ou le score-*t*, est attribué sur la base d'un grand corpus de référence. L'emploi d'un corpus natif permet de distinguer les séquences de mots typiques de la langue de celles qui sont peu ou pas appropriées. Se limiter aux seuls bigrammes de mots, comme cela est fait dans la présente étude, facilite grandement l'automatisa-

tion de l'analyse, mais le prix à payer n'est pas négligeable : seule une petite partie de la phraséologie d'une langue peut être prise en compte et le plus souvent d'une manière indirecte. Les mesures proposées ne sont donc pas immunes aux critiques formulées par Bulté et Housen (2012, p. 40) de la manière suivante : « *none of the complexity measures employed or recommended in the L2 research is unproblematic, neither in its computation nor in its interpretation* ».

Il est à noter qu'une procédure très similaire a été utilisée par Bernardini (2007) en traductologie afin de comparer l'emploi d'expressions phraséologiques dans des textes traduits et non traduits. Elle présente aussi une grande parenté avec une approche, déjà ancienne en TAL, conçue pour identifier des erreurs dans des textes sur la base des séquences de mots (ou d'étiquettes syntaxiques) qui sont extrêmement peu probables lorsqu'on les compare à celles qui se trouvent dans un corpus natif (Chodorow et Leacock 2000). Toutefois, dans le cas présent, les séquences sont vues comme se situant sur un continuum allant des séquences les plus typiques de la langue, celles qui ont les scores d'association les plus élevés, jusqu'aux séquences les moins typiques, les anticollocations dans la terminologie d'Evert (2008).

L'objectif général de la présente recherche est d'évaluer d'une manière approfondie l'utilité de prendre en compte ces mesures totalement automatiques de la compétence phraséologique pour estimer la qualité de textes d'apprenants d'une langue étrangère. La section suivante décrit les travaux à la source de cette étude. La troisième section énonce les questions de recherche auxquelles la quatrième section essaie d'apporter une réponse au moyen de l'analyse de trois ensembles de textes d'apprenants de l'anglais langue étrangère. Afin d'illustrer plus concrètement l'approche employée pour obtenir les indices phraséologiques, la cinquième section présente une application en ligne librement disponible (http ://collgram.pja.edu.pl/Default, consulté le 7 mai 2017), développée par Lenko-Szymanska et Wolk (2016), qui permet d'obtenir de tels indices pour un texte d'apprenants ou tout autre texte en anglais.

2. Travaux antérieurs

Durrant et Schmitt (2009) ont, à notre connaissance, été les premiers à proposer d'attribuer aux bigrammes de mots présents dans un texte rédigé par un apprenant d'une langue étrangère des scores d'association collocationnelle calculés sur la base d'un grand corpus de référence contenant des textes rédigés par des natifs. Comme la fréquence d'occurrence d'un bigramme est connue pour ne pas être un indice très fiable du degré d'association entre deux mots (parce qu'elle est fortement influencée par la fréquence des mots qui le compose, voir Evert, 2008), Durrant et Schmitt emploient deux mesures d'association : l'information mutuelle (*IM*) et le score-*t*. Ces deux mesures comparent la fréquence avec laquelle deux mots apparaissent l'un à la suite de l'autre dans le corpus à la fréquence attendue si les mots du corpus sont ordonnés d'une manière aléatoire. Chacune de ces mesures met en évidence un type de collocation différent : *IM* privilégie les bigrammes composés de mots relativement rares tels que *inversement proportionnel* et *violation flagrante* alors que le score-*t* fait

ressortir ceux composés de mots relativement fréquents tels que *par exemple* et *avoir besoin*. Durrant et Schmitt ont observé que, comparés à des locuteurs natifs, des apprenants de l'anglais ont tendance à employer moins souvent les bigrammes ayant un score *IM* élevé et à employer plus souvent ceux qui ont un score-*t* élevé.

Plus récemment, Bestgen et Granger (2014) ont montré que la moyenne des scores *IM* d'un texte et la proportion de bigrammes présents dans le texte, mais absents du corpus de référence sont significativement corrélées avec la qualité d'un texte telle que déterminée par des évaluateurs humains. Toutefois, les corrélations (en valeurs absolues) obtenues ne sont pas très élevées (entre 0,28 et 0,38) et surtout l'efficacité des indices phraséologiques n'a pas été comparée à celle d'autres indices connus pour prédire la qualité d'un texte. Il est donc impossible de savoir si ces indices phraséologiques sont vraiment utiles ou si, à tout le moins, ils permettent d'améliorer la prédiction lorsqu'ils sont combinés avec d'autres indices.

Ces limitations ont été dépassées par Somasundaran et Chodorow (2014) et Somasundaran *et al.* (2015) qui ont démontré le bénéfice apporté par des mesures phraséologiques pour évaluer automatiquement des phrases produites à partir de deux mots et d'une image et des narrations produites à partir d'une série d'images. Comme Durrant et Schmitt, Somasundaran *et al.* (2015) ont employé un grand corpus de référence pour obtenir les scores *IM* des bigrammes, ainsi que des trigrammes, présents dans les réponses orales retranscrites par un système automatique. Ils en ont dérivé une série d'indices : le maximum, le minimum et la médiane des scores *IM* ainsi que la proportion de scores *IM* se situant dans un intervalle parmi huit déterminés manuellement comme [-inf,-20],]-20,-10],]-10,-1],]-1,0] ou]20, +inf]. Utilisant une procédure d'apprentissage supervisé, ils ont observé que ces indices étaient très efficaces pour évaluer les réponses des apprenants, même lorsqu'ils sont comparés à une approche état de l'art fondée principalement sur des traits acoustiques extraits automatiquement.

S'appuyant sur toutes ces études antérieures, Bestgen (2016) a combiné leurs spécificités. Comme Durrant et Schmitt (2009), il a employé les bigrammes auxquels plusieurs scores d'association ont été affectés, dont *IM* et le score-*t*, mais aussi le score *z* ou la probabilité issue du test exact de Fisher. Comme Somasundaran *et al.* (2015), il a extrait de ces scores des traits plus riches que la moyenne en appliquant une procédure automatique et non supervisée de discrétisation des distributions des scores en intervalles de même fréquence. Enfin, il a comparé l'efficacité des indices phraséologiques à celle obtenue en utilisant comme traits la simple fréquence des unigrammes et des bigrammes (Yannakoudakis *et al.*, 2011). L'évaluation a été effectuée sur des textes de 200 à 400 mots du *Cambridge Learner Corpus*, plus particulièrement ceux du *First Certificate in English Examination* décrit dans Yannakoudakis *et al.* (2011). Le modèle prédictif, construit au moyen d'une procédure d'apprentissage supervisé, a été appris sur 1 141 textes recueillis en 2000 et ensuite testé sur 97 textes de 2001. Les résultats indiquent qu'employer simultanément une série d'indices d'association produit de meilleures performances que leur emploi isolé et que discrétiser les distributions de scores d'association est bénéfique. Si les fréquences des mots et des bigrammes dans les textes sont plus efficaces que la meilleure combinaison de traits phraséolo-

giques, leur ajouter ces traits phraséologiques donne lieu à une performance nettement meilleure.

3. Questions de recherche

Tous ces résultats sont encourageants et suggèrent que ces indices phraséologiques sont utiles pour estimer la qualité de textes écrits par des apprenants. Ils ont toutefois été obtenus au moyen de procédures de calcul des indices partiellement différentes : bigrammes ou bigrammes et trigrammes, un seul indice d'association ou une combinaison d'indices, moyenne des scores ou discrétisation des distributions. Déterminer si l'on peut confirmer l'efficacité de cette approche en appliquant exactement la même procédure à trois ensembles de textes d'apprenants se différenciant par le genre de textes, le thème, les conditions de production et les caractéristiques des apprenants est le premier objectif de cette recherche.

L'intérêt majeur d'analyser simultanément trois ensembles de données réside cependant dans la possibilité d'évaluer l'efficacité de la technique, non par une procédure de validation croisée sur le même ensemble de textes, mais au moyen d'une validation externe en apprenant et en testant le modèle prédictif sur des ensembles de textes différents. Confirmer l'efficacité de la technique dans ce contexte, et donc sa capacité de généralisation, est une condition nécessaire pour développer un système d'évaluation de textes capable de traiter des textes très diversifiés comme ceux qui sont produits tout au long de l'apprentissage d'une langue étrangère.

La présente recherche vise aussi à mieux évaluer l'utilité des mesures phraséologiques en les comparant, non seulement à la fréquence des mots et des bigrammes (Yannakoudakis *et al.*, 2011 ; Bestgen, 2016), mais aussi à un ensemble de statistiques lexicales connues pour être de bons prédicteurs de la qualité d'un texte d'apprenants comme la longueur du texte et des mesures de diversité lexicale fondées sur les mots isolés (Engber, 1995 ; Lu, 2012). Ces comparaisons permettront de déterminer si les mesures phraséologiques apportent une information plus riche ou différente de celle apportée par des mesures lexicales nettement plus simples à extraire d'un texte.

Enfin, comme expliqué ci-dessus, les indices phraséologiques nécessitent pour être calculés le recours à un corpus de référence. Dans les travaux antérieurs, plusieurs corpus différents ont été employés, mais l'impact de ceux-ci sur les résultats n'a jamais été évalué. Dans la présente étude, trois corpus fréquemment employés en linguistique de corpus et en TAL sont comparés.

La section suivante tente d'apporter une réponse, au moins partielle, à ces questions au moyen de l'analyse de trois ensembles de textes d'apprenants analysés séparément et différemment dans les travaux antérieurs.

4. Validation interne et externe

4.1. *Données*

4.1.1. *Ensembles de textes d'apprenants*

Trois ensembles de données, composés de textes rédigés par des apprenants de l'anglais, ont été analysés afin de pouvoir évaluer le degré de généralité des conclusions.

– FCE : le premier ensemble de textes est extrait du *First Certificate in English examination* décrit par Yannakoudakis *et al.* (2011). Ce test a pour fonction d'évaluer des apprenants de l'anglais au niveau intermédiaire supérieur. Extrait du *Cambridge Learner Corpus*, il est composé de documents de 200 à 400 mots correspondant à la section de rédaction de l'examen. Les participants doivent rédiger deux textes de 120 à 180 mots comme une lettre, une composition ou une histoire en réponse à une instruction précisant le thème. Les 1 141 documents de l'année 2000 pour lesquels on dispose des deux textes ont été employés dans les analyses[1]. Dans le cadre de l'examen, une note globale a été attribuée à chaque document sur une échelle allant de 0 à 40. On ne dispose pas d'information à propos de la fiabilité de cette évaluation, mais s'agissant d'un test de niveau réputé, on peut penser qu'elle est élevée. De plus, Yannakoudakis *et al.* (2011) ont estimé la fiabilité des évaluations de 97 autres textes extraits de cet ensemble en demandant à quatre experts de les réévaluer et ils ont obtenu une corrélation du produit des moments de Pearson (Howell, 2008, p. 244–245) entre ceux-ci et le score original de 0,80.

– ICLE : cet ensemble de textes, décrit en détail par Thewissen (2013), est composé de 223 essais argumentatifs d'une longueur de 500 à 900 mots à propos de thèmes comme le rôle de l'argent ou l'importance de l'imagination dans le monde actuel. Ils ont été sélectionnés aléatoirement dans trois sous-corpus de l'International Corpus of Learner English (ICLE v2, Granger *et al.*, 2009) : 74 essais du composant français, 71 du composant allemand et de 78 du composant espagnol. Deux évaluateurs professionnels ont attribué à chacun des textes un score holistique sur la base des descripteurs du Cadre européen commun de référence pour les langues (CECRL) (Council of Europe, 2001) en employant les niveaux B1, B2, C1 et C2, soit les niveaux intermédiaire et avancé. Les évaluateurs pouvaient employer un signe + ou un signe – pour spécifier plus finement la qualité des textes dans chaque niveau de compétence. Les évaluations obtenues ont été transformées en nombre sur une échelle numérique de 11 points allant de B1– = 1 jusqu'à C2 = 11. La corrélation entre les deux évaluateurs sur cette échelle de notation est de 0,69. Les 34 textes (15 %) sur lesquels les deux évaluateurs étaient en désaccord de plus d'un niveau (par exemple B1–C1) ont été soumis à un troisième évaluateur. Le score final d'un essai est la moyenne des deux (ou trois) scores disponibles.

1. Estimer la qualité de ces textes est nettement plus difficile qu'estimer celle des 97 textes de 2001 employés par Bestgen (2016) comme l'ont montré les analyses de Yannakoudakis et Briscoe (2012), la différence entre les performances à système identique étant de 14 %.

– MSU : le troisième ensemble de textes est le Michigan State University, composé de 171 textes de 140 à 650 mots écrits par 57 apprenants durant un semestre de cours universitaire d'anglais intensif avec un intervalle de six à huit semaines entre les trois rédactions. Il s'agit de textes descriptifs rédigés en 30 minutes au maximum en réponse à des énoncés comme *Décrivez le campus de l'université* ou *Décrivez votre famille* (voir Connor-Linton et Polio (2014) pour des détails). Chaque texte a été évalué sur une échelle allant de 0 à 100 par deux experts au moyen d'une grille analytique prenant en compte principalement le contenu, l'organisation, le vocabulaire, la syntaxe et l'orthographe. La corrélation entre leurs jugements est de 0,88 et la mesure de qualité employée dans la suite est la moyenne des notes des deux juges.

4.1.2. *Corpus de référence*

La procédure employée pour estimer la force collocationnelle d'un bigramme impose le recours à un corpus de référence suffisamment grand pour permettre une estimation précise des scores d'association et aussi représentatif que possible de la langue telle qu'elle est employée par les locuteurs natifs. Afin d'évaluer l'impact du corpus de référence employé, trois corpus de taille et d'origine différentes ont été comparés :

– le *British National Corpus* (BNC), un corpus de 100 millions de mots constitué de manière à représenter dans une large mesure la diversité de l'anglais britannique de la fin du vingtième siècle ;

– le *Corpus of Contemporary American English* (COCA), un corpus de plus de 400 millions de mots constitué de 20 millions de mots récoltés chaque année entre 1990 et 2011 de manière à représenter dans une large mesure la diversité de l'anglais américain ;

– un corpus issu du projet *WaCKy* (Baroni *et al.*, 2009), le *UkSubset* qui est composé de 100 millions de mots. Ce corpus a été construit par une procédure d'extraction de textes d'Internet fondée sur des listes de mots germes choisis de manière à maximiser la diversité des textes.

4.2. *Construction des modèles prédictifs*

Les modèles prédictifs ont été construits par apprentissage supervisé sur la base de trois ensembles de traits : phraséologiques, statistiques lexicales et n-grammes. Nous avons employé la procédure proposée par Yannakoudakis *et al.* (2011) et Yannakoudakis et Briscoe (2012) qui ont évalué l'efficacité d'une série d'indices linguistiques pour prédire la qualité des textes dans l'ensemble de données FCE, également analysé dans la présente recherche. Ces auteurs ont montré que traiter la tâche d'évaluation automatique comme un problème d'apprentissage de préférence de rang, au moyen du package SVMRank (Joachims, 2006), était plus efficace que de la traiter au moyen d'une régression à vecteurs de support. La suite de cette section décrit les indices employés.

4.2.1. *Indices phraséologiques*

Les indices phraséologiques ont été obtenus au moyen des trois étapes suivantes.

1) Extraction des bigrammes. Tous les bigrammes composés de mots sont extraits de chaque texte. Les signes de ponctuation et toute suite de caractères qui ne correspond pas à un mot interrompent les bigrammes. Pour la suite des analyses, seuls les bigrammes différents (types) sont conservés afin de donner plus de poids à leur diversité (Durrant et Schmitt, 2009).

2) Obtention des scores d'association. Chaque bigramme est recherché dans le corpus de référence. S'il y est trouvé, six mesures d'association, bien établies en extraction d'expressions polylexicales à partir de corpus (Evert, 2008 ; Pecina, 2008) et qui ont montré leur efficacité (Bestgen, 2016), sont calculées :

– *IM* qui correspond à la transformation logarithmique du rapport entre la fréquence observée du bigramme et sa fréquence attendue sous l'hypothèse d'indépendance complète ;

– Score-*t* qui est calculé en divisant la différence entre la fréquence observée du bigramme et sa fréquence attendue par la racine carrée de la fréquence observée ;

– *Z* qui est calculé en divisant la différence entre la fréquence observée du bigramme et sa fréquence attendue par la racine carrée de la fréquence attendue ;

– *Simplell* comme l'a appelé Evert (2008), la contribution apportée par la fréquence du bigramme à la statistique du test du maximum de vraisemblance recommandée par Dunning (1993) pour analyser les collocations dans des corpus ;

– *Fisher* qui correspond à la probabilité, obtenue par le test exact de Fisher, d'observer sous l'hypothèse nulle d'indépendance au moins autant d'occurrences du bigramme que le nombre réellement observé ;

– *Mutual rank ratio*, une mesure non paramétrique d'association proposée par Dean (2005) qui a été employée avec succès pour détecter des erreurs phraséologiques dans des textes d'apprenants de l'anglais (Futagi *et al.*, 2008).

Toutes ces mesures d'association ont été calculées au moyen des formules rassemblées par Evert (2008) et en appliquant la procédure décrite par Dean (2005) pour la dernière.

Enfin, si le bigramme n'est pas trouvé dans le corpus de référence, il est pris en compte dans le calcul de la proportion des bigrammes présents dans un texte, mais absents du corpus de référence.

3) Extraction des traits phraséologiques. Ils sont composés des trois types suivants :

– la proportion de bigrammes présents dans un texte, mais absents du corpus de référence ;

– quatre statistiques descriptives globales : la moyenne, la médiane, le maximum et le minimum des scores d'association ;

– les traits extraits par la procédure de discrétisation. Dans un premier temps, la distribution des scores de chaque mesure d'association est discrétisée au moyen de la procédure automatique de discrétisation en intervalles de même fréquence. Ensuite, on compte dans chaque texte la proportion de bigrammes dont le score d'association se trouve dans chaque intervalle en employant comme dénominateur le nombre total de bigrammes présents dans le texte. Cette discrétisation a été effectuée en employant dix, vingt et cinquante intervalles afin de contrôler l'impact de ce paramètre sur l'efficacité.

En résumé, l'ensemble des traits phraséologiques fournis à la procédure d'apprentissage supervisé est composé a) de la proportion de bigrammes absents (1 trait), b) des quatre statistiques globales pour chaque mesure d'association (24 traits) et c) d'un ensemble de traits produits par la discrétisation pour chaque mesure d'association (de 60 à 300 traits au total selon le nombre d'intervalles employés).

4.2.2. *Statistiques lexicales : diversité lexicale et nombre de mots*

Parmi les très nombreuses mesures de diversité lexicale disponibles (Lu, 2012), nous avons sélectionné les trois recommandées par McCarthy et Jarvis (2010) dans leur étude comparative parce qu'elles apportent des points de vue complémentaires sur la diversité lexicale tout en étant peu affectées par d'autres facteurs comme la longueur des textes, contrairement au classique Type-Token Ratio (TTR) :

– l'indice de Maas, une transformation logarithmique du TTR ;

– HD-D qui estime la probabilité que de nouveaux mots soient introduits dans des échantillons de plus en plus longs extraits du texte ;

– MTLD, la mesure de la diversité lexicale textuelle qui est égale à la longueur moyenne des séquences de mots dans un texte qui ont une valeur de TTR de 0,72.

Nous avons ajouté à ceux-ci l'indice de Guiraud, qui est une transformation racine carrée du TTR, parce qu'il est un des plus efficaces pour prédire la qualité des textes (Lu, 2012). Pour la même raison, le nombre total de mots dans un texte (*tokens*) et le nombre de mots différents (types) ont aussi été employés (Santos *et al.*, 2012 ; Lu, 2012).

4.2.3. *Unigrammes et bigrammes de mots*

Nous avons également employé des traits lexicaux, dont Yannakoudakis *et al.* (2011) ont montré qu'ils permettaient à eux seuls une excellente prédiction de la qualité d'un texte. Ils sont composés de la fréquence des unigrammes et bigrammes de mots dans chaque texte. Ce point de comparaison est particulièrement pertinent parce qu'il inclut les bigrammes qui sont à la base des scores phraséologiques. Ces traits ont été extraits en suivant la procédure de Yannakoudakis *et al.* (2011) et donc en employant un seuil de fréquence minimal de 4 et une pondération de la fréquence par TF$\times$IDF.

4.2.4. *Synthèse des ensembles de traits employés*

Ces ensembles de traits ont été employés séparément ou en combinaison produisant sept conditions expérimentales : lexicaux (L), n-grammes (G), phraséologiques (P), lexicaux + n-grammes (LG), lexicaux + phraséologiques (LP), n-grammes + phraséologiques (GP) et lexicaux + n-grammes + phraséologiques (LGP). Les traits phraséologiques et les statistiques lexicales ont été centrés (moyenne de 0). Tous les traits, y compris les n-grammes, ont été pondérés par la norme L2 suivant la procédure employée par Yannakoudakis *et al.* (2011).

4.3. *Évaluation de l'efficacité des modèles prédictifs*

4.3.1. *Mesure d'efficacité*

Étant donné que les textes des trois ensembles de données ont été évalués sur des échelles composées de 11 à 101 valeurs, la corrélation entre les scores réels et les scores prédits, calculée au moyen du coefficient de Pearson, est utilisée comme mesure de performance du modèle.

4.3.2. *Estimation de l'efficacité par validation interne*

Pour estimer l'efficacité des modèles prédictifs construits sur la base de chaque ensemble de traits et de leur combinaison, nous avons employé une procédure de validation croisée répétée. Chaque ensemble de données a été divisé aléatoirement en dix blocs. Chacun d'eux a servi à tour de rôle pour évaluer l'efficacité du modèle construit sur la base des neuf autres blocs. Afin de limiter l'impact potentiel de la partition effectuée, cette procédure est répétée quinze fois en employant des germes différents pour le générateur aléatoire. La mesure d'efficacité est la moyenne des 150 valeurs ainsi obtenues (10 blocs $\times$ 15 germes).

Cette procédure ne résout toutefois pas un problème : la fixation du métaparamètre C de régularisation qui ajuste le rapport entre les capacités de généralisation du modèle et l'efficacité sur le matériel d'apprentissage. Sélectionner la valeur de celui-ci sur la base de la performance du modèle lors de l'analyse de l'échantillon de test induit un biais positif. Afin de l'éviter, une deuxième boucle de validation croisée, interne à la première, a été employée. Chaque échantillon d'apprentissage (composé des neuf blocs initiaux) a été divisé aléatoirement en cinq sous-blocs. Chacun d'eux a servi à tour de rôle pour évaluer l'efficacité du modèle construit sur la base des quatre autres sous-blocs en faisant varier le paramètre C auquel les valeurs suivantes étaient successivement affectées : 0,00001, 0,0001, 0,001, 0,01, 0,1 et 1. Cette procédure est répétée dix fois en employant des germes différents pour le générateur aléatoire. L'efficacité est déterminée en calculant la moyenne des cinquante valeurs obtenues et le paramètre C, employé dans la boucle externe, est fixé à la valeur qui a obtenu la moyenne la plus élevée.

L'intérêt d'employer plusieurs répétitions de chaque boucle de validation croisée est évidemment de permettre l'obtention d'estimations plus précises tant du paramètre C que de l'efficacité du modèle, mais aussi, dans ce dernier cas, de pouvoir tester la significativité statistique des différences d'efficacité entre les modèles puisque quinze valeurs sont à chaque fois obtenues pour chaque modèle (une par valeur germe). De plus, l'emploi des mêmes valeurs germes garantit que les valeurs comparées sont issues des mêmes échantillons d'apprentissage et de test.

4.3.3. *Estimation de l'efficacité par validation externe*

L'ensemble de données contenant le plus de documents a été employé comme échantillon d'apprentissage et les deux autres sont employés séparément comme échantillon de test[2]. Pour fixer le paramètre C, il semble risqué d'utiliser une procédure de validation croisée à l'intérieur de l'échantillon d'apprentissage étant donné qu'on peut penser que généraliser à une partie de celui-ci ou généraliser à des textes issus d'un autre ensemble de données sont deux situations très différentes. Nous avons donc choisi de construire un modèle prédictif sur la base de l'échantillon d'apprentissage pour chacune des valeurs C testées et de l'appliquer aux deux échantillons de tests afin de pouvoir comparer les performances et de déterminer si les valeurs du paramètre C qui produisent les meilleures performances dans les deux échantillons de tests sont ou non identiques. Les valeurs de C testées sont identiques à celles employées lors de la validation interne.

4.4. *Résultats*

4.4.1. *Comparaison de l'efficacité des différents modèles prédictifs par validation interne*

La figure 1 présente toutes les mesures d'efficacité (corrélations de Pearson) obtenues par la procédure de double validation croisée et donc en apprenant et en testant le modèle sur des sections différentes d'un même ensemble de données. Dans ces graphiques, un pour chaque ensemble de données (FCE, ICLE et MSU), les points représentent une corrélation pour un des quinze ensembles de traits employés. Ils sont également différenciés selon le corpus de référence (CR) employé dans les calculs. De gauche à droite, on trouve, tout d'abord, les statistiques **L**exicales seules, les n-**G**rammes seuls, et les statistiques lexicales et n-grammes. Ces trois ensembles de traits ne nécessitant pas de corpus de référence, une seule valeur est donnée (et l'identifiant du corpus de référence est « ___ ». Suivent les douze ensembles de traits incluant les indices phraséologiques et donc calculés avec l'aide d'un des trois corpus de référence (BNC, COCA et WACKY) et en employant dix, vingt ou cinquante intervalles lors de la discrétisation. Les corpus de référence sont distingués au moyen des traits qui relient les corrélations et les nombres d'intervalles divisent les points en trois blocs sur

2. Comme les ensembles de textes ont été évalués au moyen de grilles différentes, il n'est pas possible de combiner les textes de deux ensembles différents.

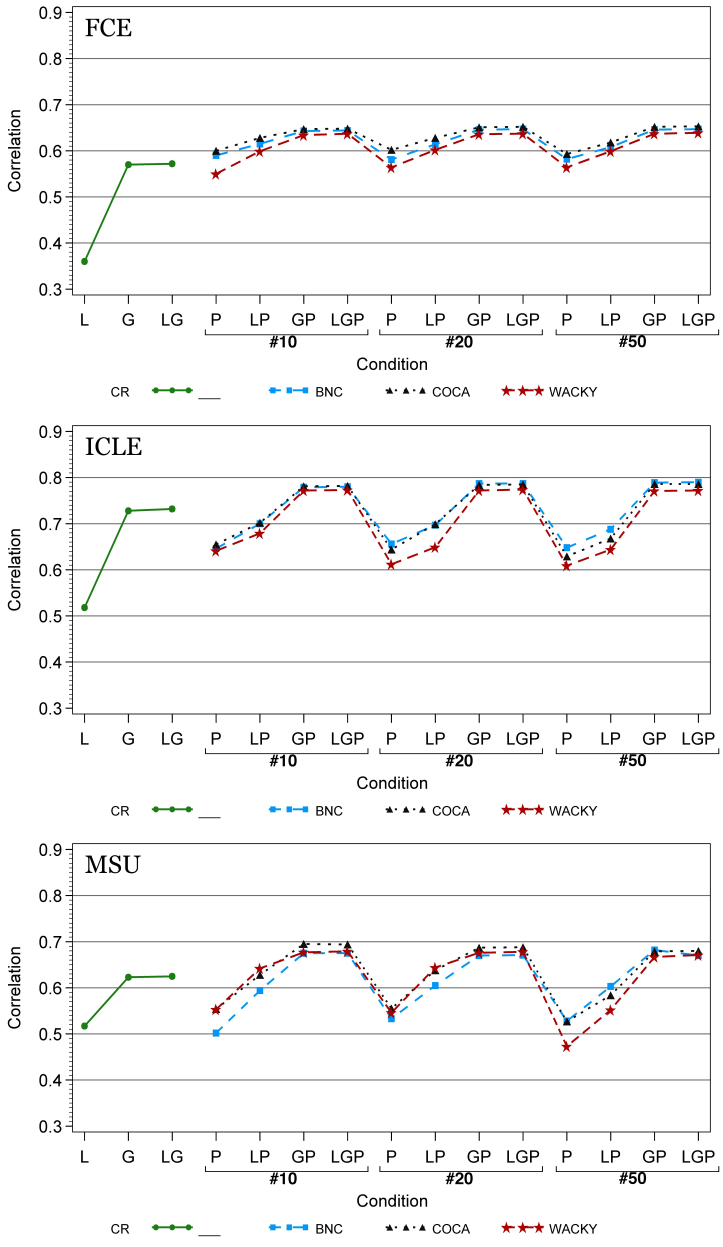

Figure 1. *Performances pour les trois ensembles de données en validation interne. Note : CR = corpus de référence.*

l'axe horizontal. Dans chaque bloc, on donne d'abord la corrélation pour les traits **P**hraséologiques seuls et puis leur combinaison avec les statistiques lexicales et avec les n-grammes et, en dernier lieu, la combinaison des trois ensembles de traits.

Comme expliqué dans la section 4.3.2, les corrélations rapportées correspondent à la moyenne de toutes les valeurs issues de la double validation croisée. Ces valeurs ont été employées pour tester la signification statistique des différences entre les corrélations au moyen d'un test *t* de Student pour mesures répétées appliqué après transformation des corrélations en score *z* de Fisher afin de normaliser les distributions. Dans la suite, seuls les effets confirmés par des différences statistiquement significatives au seuil de 0,0001 sont soulignés.

À un niveau très général, on remarque que tous les graphiques se ressemblent fortement, ce qui autorise de commencer par analyser ceux-ci en considérant successivement et indépendamment chaque facteur manipulé dans l'étude.

En ce qui concerne les différences entre les sept conditions (ou ensembles de traits), on observe que :

– les statistiques lexicales (L) sont toujours nettement moins efficaces que les n-grammes (G). Elles sont également moins efficaces que les indices phraséologiques (P), mais seulement dans deux corpus sur trois, MSU faisant exception. Ajouté à P, L améliore systématiquement les performances et le gain est même substantiel dans MSU. En revanche, ajouter L à G ou à GP n'améliore pas ou très faiblement la performance ;

– les indices phraséologiques (P) sont moins performants que les n-grammes (G) dans deux corpus sur trois, FCE faisant exception. Les ajouter à G améliore nettement la performance ;

– dans toutes les analyses, GP et LGP obtiennent les meilleurs résultats.

On observe peu de différences selon le corpus de référence utilisé pour calculer les indices phraséologiques tout particulièrement lorsque G est ajouté à ceux-ci, ce qui est logique. Lorsque les traits phraséologiques sont utilisés seuls, WaCKy est le plus souvent moins efficace que les deux autres corpus, mais il y a une exception notable : la discrétisation en 10 intervalles pour MSU où c'est l'emploi du BNC qui produit la moins bonne performance.

Le nombre d'intervalles employés pour discrétiser les distributions de scores n'affecte que très faiblement les résultats, à tout le moins pour la plage de valeurs testées.

Les similarités entre les graphiques, soulignées ci-dessus, ne doivent toutefois pas occulter l'existence de différences relativement importantes entre les ensembles de données. Tout particulièrement, l'efficacité des modèles prédictifs est nettement plus grande dans ICLE que dans FCE, MSU occupant une position intermédiaire. On note aussi que la différence entre les conditions L et G est nettement plus petite pour MSU que pour les deux autres ensembles de données. Enfin, les différences entre les quatre conditions incluant les indices phraséologiques sont plus faibles pour FCE que pour les deux autres.

Si ces différences soulignent l'impact potentiel de l'ensemble de données employé sur les performances, elles ne modifient pas les réponses aux principales questions de recherche qui peuvent être formulées de la manière suivante :

– les indices phraséologiques sont utiles pour l'évaluation automatique de textes produits par des apprenants de l'anglais langue étrangère ;

– ces indices apportent une information différente de celles apportées par des mesures lexicales classiques fondées sur les mots isolés et par la fréquence des mots et des bigrammes de mots dont ils sont issus ;

– l'impact du corpus de référence employé est perceptible, mais faible.

4.4.2. *Comparaison de l'efficacité des différents modèles prédictifs par validation externe*

Dans cette analyse, l'échantillon d'apprentissage est FCE et les deux autres ensembles de données servent d'échantillon de test. Comme corpus de référence, nous avons utilisé COCA qui est celui qui produit d'une manière générale les performances les plus élevées en validation interne (voir la figure 1).

Le tableau 1 présente la meilleure corrélation obtenue pour chaque échantillon de test et la valeur du paramètre C de régularisation correspondante ainsi que la corrélation pour la valeur du paramètre optimale dans l'autre corpus de test. Cette valeur est indiquée en caractères gras. La colonne Rang donne la place occupée par la valeur en question parmi les six valeurs testées de C et la colonne Diff rapporte la différence entre la corrélation obtenue dans cette validation externe et celle obtenue dans la validation interne (voir la figure 1).

La comparaison des performances dans les deux échantillons de test pour les différents ensembles de traits indique que, dans dix cas sur quinze, la même valeur C produit la meilleure performance dans les deux échantillons. Dans trois autres cas, la meilleure valeur pour un échantillon est la deuxième plus performante pour l'autre et la différence entre les deux meilleures performances pour un échantillon donné est au maximum de 0,014. Dans le quatorzième cas, la condition P#50, la différence de performance entre la valeur C optimale pour ce corpus et pour l'autre corpus est de 0,03. La seule condition problématique est L pour laquelle la solution optimale pour un corpus arrive en avant-dernière ou en dernière position pour l'autre et les différences de performances sont très élevées (de 0,07 à 0,09).

Comme on pouvait s'y attendre en raison des différences importantes entre les ensembles de données au niveau du genre de textes et des thèmes de ceux-ci, les n-grammes se généralisent très mal d'un ensemble à un autre. Plus étonnamment, un problème similaire est rencontré par les statistiques lexicales lorsque l'on emploie la valeur C optimale pour l'autre échantillon de test. Les indices phraséologiques sont clairement l'ensemble de traits qui se généralise le mieux, la perte par rapport aux performances lors de la validation interne étant la plus faible.

		ICLE				MSU			
Cond.	**Int.**	**C**	**Rg**	**r**	**Diff**	**C**	**Rg**	**r**	**Diff**
L		0,00001	1	0,504	−0,014	0,1	1	0,483	−0,034
		0,1	5	**0,438**	−0,080	0,00001	6	**0,388**	−0,129
G		0,00001	1	**0,565**	−0,158	0,00001	1	**0,452**	−0,171
LG		0,00001	1	**0,582**	−0,150	0,00001	1	**0,471**	−0,154
P	10	1	1	**0,621**	−0,035	1	1	**0,516**	−0,040
	20	0,1	1	0,624	−0,020	0,01	1	0,535	−0,020
		0,01	2	**0,617**	−0,027	0,1	2	**0,524**	−0,031
	50	0,001	1	0,602	−0,031	0,01	1	0,535	+0,007
		0,01	2	**0,591**	−0,042	0,001	4	**0,506**	−0,022
LP	10	0,01	1	**0,691**	−0,012	0,01	1	**0,578**	−0,050
	20	0,1	1	**0,691**	−0,008	0,1	1	**0,583**	−0,055
	50	0,01	1	**0,672**	+0,002	0,01	1	**0,585**	+0,002
GP	10	0,00001	1	**0,673**	−0,109	0,00001	1	**0,509**	−0,182
	20	0,00001	1	0,681	−0,103	0,0001	1	0,513	−0,174
		0,0001	2	**0,668**	−0,116	0,00001	2	**0,513**	−0,174
	50	0,00001	1	0,690	−0,097	0,0001	1	0,517	−0,158
		0,0001	2	**0,677**	−0,110	0,00001	2	**0,515**	−0,160
LGP	10	0,00001	1	**0,686**	−0,097	0,00001	1	**0,523**	−0,167
	20	0,00001	1	**0,693**	−0,092	0,00001	1	**0,526**	−0,162
	50	0,00001	1	**0,701**	−0,086	0,00001	1	**0,529**	−0,148

Tableau 1. *Performances pour les deux ensembles de test en validation externe. Note : Cond. = conditions, Int. = nombre d'intervalles, Rg = gang, Diff = différence.*

Parmi les sept conditions testées, c'est la combinaison LP qui produit les meilleures performances lorsque l'on prend en compte les deux échantillons de test, GP et LGP fonctionnant nettement moins bien dans MSU. Par rapport à la validation interne, la perte de performance de LP est quasi nulle pour ICLE et de seulement 0,038 pour MSU. Il n'en reste pas moins que la meilleure condition en validation externe est nettement moins performante que la meilleure condition en validation interne, mais ce résultat doit être mis à charge des n-grammes. En conclusion, la validation externe confirme l'efficacité des indices phraséologiques et de leur emploi combiné avec les statistiques lexicales.

5. Une application en ligne pour calculer des indices phraséologiques

La technique d'analyse de textes d'apprenants décrite ci-dessus est relativement complexe tout particulièrement en ce qui concerne l'extraction des indices phraséologiques. Ceci découle de l'objectif principal de la recherche qui est de développer le système le plus efficace possible pour évaluer des textes. Son application dans l'enseignement des langues étrangères est donc difficilement envisa-

geable à court terme. Toutefois, une composante importante de cette approche a été mise à disposition librement sur Internet par Lenko-Szymanska et Wolk (2016 ; http ://collgram.pja.edu.pl/Default, consulté le 7 mai 2017). Ce CollGram[3] Calculator permet d'évaluer très facilement des textes d'apprenants. Si, contrairement à la procédure employée ici, il ne prend en compte que deux indices d'association et n'utilise pas la procédure de discrétisation, il présente le grand intérêt de permettre l'identification automatique dans des textes d'apprenants des bigrammes qui s'apparentent le plus ou le moins à un emploi natif selon les scores *IM* et *t*. Il peut également être utilisé pour mettre en évidence dans des textes à visée pédagogique les séquences de mots rarement employées par les natifs, mais fortement associées selon *IM*, séquences qui sont souvent particulièrement difficiles pour les apprenants (Durrant et Schmitt, 2009).

À titre d'illustration, le CollGram Calculator a été appliqué à deux courts extraits de textes de l'ensemble de données MSU, partiellement analysés par Bestgen et Granger (2014). Le premier est l'extrait de vingt mots qui a obtenu le plus haut score *IM* moyen ; il provient d'un texte qui, avec une évaluation de qualité de 70 sur 100, fait partie des 4 % des textes ayant obtenu la meilleure note. Le second extrait est celui qui a obtenu le score *IM* moyen le plus proche de 0 ; il provient d'un texte qui, avec une évaluation de qualité de 50, fait partie des 45 % des textes ayant obtenu la moins bonne note.

> Extrait 1 : *He used to investigate illegally parked cars. Now, he came here with me in order to improve his ability to speak English fluently.*
>
> Extrait 2 : *In the morning everything are peaceful Nothing is show to you the difference in the night. In the nights everything are changed.*

Chaque extrait a été soumis au CollGram Calculator qui emploie le COCA comme corpus de référence. Bien que l'implémentation soit totalement différente de celle de Bestgen et Granger (2014), les scores obtenus sont très similaires, la corrélation de Pearson entre ceux-ci est de 0,99, soit pratiquement égale à la valeur maximale de 1 de ce coefficient.

L'outil renvoie des documents Excel compressés, dont une synthèse qui contient les scores globaux pour chaque texte soumis, et une série de fichiers, un pour chaque texte, donnant les scores *IM* et *t* pour chaque bigramme. Ce sont ces derniers fichiers qui sont particulièrement intéressants. Le tableau 2 présente les principales informations qu'ils contiennent. En plus des scores *IM* et *t*, on y trouve la fréquence du bigramme dans le texte ainsi que dans le corpus de référence.

3. Bestgen et Granger (2014) ont proposé ce terme parce que l'unité phraséologique analysée combine des propriétés des collocations, en s'appuyant sur des scores d'association collocationnelle, et des n-grammes puisqu'elle est composée de paires de mots contigus.

Extrait 1				
Bigramme	#EX	#CR	IM	Score-t
he used	1	7 083	2,2	65,5
used to	1	81 621	3,9	266,3
to investigate	1	5 930	4,7	74,1
investigate illegally	1	0	.	.
illegally parked	1	32	9,2	5,7
parked cars	1	320	9,1	17,9
he came	1	12 994	3,2	101,6
came here	1	3 189	4,0	52,9
here with	1	6 100	1,0	38,7
with me	1	25 990	2,3	127,6
me in	1	16 123	0,1	10,9
in order	1	43 218	4,6	199,5
order to	1	39 301	4,0	185,9
to improve	1	14 713	4,4	115,6
improve his	1	479	2,0	16,3
his ability	1	2 360	3,4	43,9
ability to	1	33 591	4,7	176,3
to speak	1	19 454	4,0	130,9
speak english	1	1 246	8,0	35,2
english fluently	1	20	9,4	4,5
Extrait 2				
Bigramme	#EX	#CR	IM	Score-t
in the	3	2 043 382	2,2	1 119,0
the morning	1	24 363	1,8	112,6
morning everything	1	10	– 1,7	– 7,0
everything are	2	30	– 4,5	– 122,1
are peaceful	1	65	0,3	1,7
peaceful nothing	1	0	.	.
nothing is	1	3 356	0,3	11,2
is show	1	534	– 2,6	– 117,4
show to	1	1 205	– 1,8	– 88,8
to you	1	49 839	– 0,8	– 168,1
you the	1	11 815	– 4,0	– 1 618,1
the difference	1	15 914	2,5	104,1
difference in	1	7 437	3,0	75,5
the night	1	31 499	1,6	119,4
the nights	1	585	– 0,5	– 9,6
nights everything	1	1	– 2,0	– 2,0
are changed	1	236	– 0,4	– 5,3

Tableau 2. *Bigrammes et scores CollGram pour les deux extraits. Note : #EX = fréquence dans l'extrait, #CR = fréquence dans le corpus de référence.*

La comparaison des extraits et du tableau met en évidence une série de caractéristiques actuelles de l'approche CollGram telle que définie par Granger et Bestgen (2014) et appliquée par le CollGram Calculator :

– la présence de signes de ponctuation interrompt l'extraction des bigrammes comme pour *Now, he*. L'omission d'un signe nécessaire (*peaceful Nothing*) n'est pas corrigée automatiquement et entraîne la prise en compte du bigramme en question ;

– certains bigrammes, comme *illegally parked*, ne reçoivent pas de scores d'association parce qu'ils ne se trouvent pas dans le corpus de référence ;

– certains bigrammes, comme *you the* ou *show to*, reçoivent des scores négatifs. Il s'agit de paires de mots qui cooccurrent dans le corpus de référence moins fréquemment que ce que le hasard ne prédirait (voir Evert, 2008, p. 1224-1230, pour une explication détaillée). Un score négatif ne signifie pas nécessairement que la séquence est erronée ; il peut aussi s'agir d'une combinaison particulièrement créative, et donc très rare, mais tout à fait acceptable dans la langue.

On observe aussi que les valeurs issues des deux scores d'association ne sont pas directement comparables, les scores-t étant nettement plus dispersés. Ceci résulte des formules de calcul et tout particulièrement du fait que la formule de IM inclut une transformation logarithmique (voir à ce sujet Stubbs, 1995).

Une analyse qualitative des bigrammes qui obtiennent en général les scores *IM* et les scores *t* les plus élevés a été effectuée par Bestgen et Granger (2014) et par Granger et Bestgen (2014). Elle montre que les bigrammes qui obtiennent des scores *IM* élevés sont généralement composés de mots relativement rares, mais fortement associés comme *vacuum cleaner, alcoholic beverages* ou *traffic jam* alors que ceux qui obtiennent des scores *t* élevés sont fréquemment composés de mots très fréquents comme *of the, do not, going to, you know, out of* ou *more than*.

Il est important de souligner à nouveau que la recherche présentée ici emploie les indices CollGram d'une manière globale afin d'estimer automatiquement la qualité des textes. Comme le montre le tableau 2, une partie des bigrammes analysés a un intérêt pédagogique limité. Des recherches complémentaires sont indispensables pour évaluer les scores d'association attribués à des combinaisons spécifiques de mots.

6. Conclusion

L'objectif de la présente recherche était d'évaluer l'utilité de prendre en compte des mesures totalement automatiques de la compétence phraséologique pour estimer la qualité de textes d'apprenants de l'anglais langue étrangère. Au travers de l'analyse de trois ensembles de données se distinguant par le nombre de textes, le genre et le thème de ceux-ci, la situation de production et les caractéristiques sociolinguistiques des apprenants, les analyses ont montré que les indices phraséologiques apportaient une information utile. De plus, cette information est différente et, donc, complémentaire à celles apportées par des mesures lexicales plus simples comme le nombre de

mots différents ou la diversité lexicale. La validation externe indique que ces indices présentent un degré de généralisation important et qu'ils peuvent donc être employés pour évaluer de nouveaux textes sans devoir réapprendre le modèle sur un ensemble de textes similaires à ceux-ci. Ce résultat suggère que les traits phraséologiques et les statistiques lexicales pourraient être intégrés dans un modèle prédictif, accessible par Internet, qui serait capable d'évaluer des textes très diversifiés et donc pourrait faciliter l'évaluation formative en apprentissage d'une langue étrangère. Il est toutefois évident que d'autres dimensions de la qualité d'un texte doivent être prises en compte. La composante syntaxique (complexité, erreurs) est certainement la première qui mériterait d'être ajoutée (Kuiken et Vedder, 2012 ; Grant et Ginther, 2000).

Cette synthèse conduit à ce qui nous semble être la limite principale de cette étude : le point de vue très spécifique qui est choisi pour évaluer automatiquement la qualité des textes. Ce point de vue est très partiel et non exempt de problèmes. Tout d'abord, les mesures proposées peuvent être aisément trompées en leur soumettant des phrases, ou même de simples tournures, issues de textes natifs. Ensuite et surtout, l'approche employée, qui compare les textes à un corpus de référence, tombe sous les critiques d'une série d'auteurs (Condon, 2013 ; Herrigton et Morran, 2012) qui soulignent que rédiger un texte est d'abord un acte de communication entre humains qui met en jeu de nombreuses compétences (raisonnement, rhétorique) et connaissances (à propos du thème du texte, mais aussi à propos du genre de texte à rédiger et du contexte socioculturel dans lequel il est produit et ensuite lu). Dans la présente recherche, on ne s'intéresse qu'à l'évaluation des compétences en langue étrangère au travers de la rédaction de textes et non aux compétences en rédaction de textes, que ceux-ci soient rédigés dans la langue apprise ou native (Weigle, 2013). Ces deux compétences interagissent nécessairement lors de la rédaction d'un texte en langue apprise comme l'a souligné Galbraith (2009, p. 12-13), « *L2 language proficiency would be expected to affect not just how well-formed the written product is from a linguistic point of view, but also the writer's capacity to engage in the higher level problem-solving activities characteristic of expert writing* ». L'auteur d'un texte peut privilégier la correction du texte produit au détriment de sa pertinence par rapport au thème et de son intérêt pour un lecteur. Des mesures automatiques qui ne prennent en compte qu'une seule de ces deux compétences sont donc nécessairement partielles.

Plusieurs résultats obtenus dans cette étude méritent une discussion plus approfondie. Tout d'abord, si les réponses aux principales questions de recherche sont globalement identiques quelles que soient les données analysées, des différences importantes en termes d'efficacité sont observées entre les trois ensembles de données puisqu'elles atteignent 0,13 de corrélation entre ICLE et FCE pour les conditions GP et LGP. Des différences de 0,10 s'observent également pour la condition P. Ces différences rappellent combien il est risqué de se fier à un seul ensemble de données pour évaluer l'efficacité d'une procédure. La recherche ne permet hélas pas de déterminer quel facteur en est responsable. Les ensembles de données ont été recueillis dans des situations très différentes. La longueur des textes, le genre, le thème et même les critères pour évaluer les textes sont différents dans les trois ensembles de données. Tout au plus peut-on noter que les textes de ICLE sont les plus longs et qu'ils ont, dans 80 % des

cas, été produits sans contrainte temporelle, contrairement aux textes des deux autres ensembles. Il est donc possible qu'ils contiennent une plus grande quantité d'informations de meilleure qualité pour effectuer une estimation automatique.

L'évaluation par validation externe indique que les n-grammes se généralisent très mal d'un ensemble de données à un autre. Ce résultat s'explique très probablement par les différences de genres et de thèmes entre les textes qui les composent. La possibilité pour les n-grammes de mots d'encoder, partiellement, des erreurs potentielles plus ou moins fréquentes comme *their are* ou *resemble with* ne suffit donc pas à leur donner une capacité de généralisation. On notera toutefois que, dans le cadre de tests standardisés largement diffusés portant sur un nombre limité de sujets, employer les unigrammes et les bigrammes se révèle particulièrement efficace, une conclusion qui rejoint les observations de Yannakoudakis *et al.* (2011).

Le faible impact des corpus de référence est aussi quelque peu inattendu. Les corpus employés varient fortement par leur taille, la variété de la langue anglaise et le degré supposé de représentativité. Malgré cela, les différences sont faibles et dans des directions variables selon l'ensemble de données et le nombre d'intervalles employés pour la discrétisation. Le niveau de performance atteint par WaCKy mérite d'être souligné en raison de son accès libre et il serait intéressant, dans des travaux futurs, d'employer la version complète de ce corpus qui contient 2 milliards de mots.

D'autres pistes mériteraient d'être suivies. Somasundaran *et al.* (2015) ont analysé les bigrammes et les trigrammes, alors que seuls les bigrammes ont été employés dans cette étude. En revanche, Somasundaran *et al.* (2015) n'ont employé qu'une seule mesure d'association alors que nous en avons employé six. Prendre en compte les trigrammes imposera une analyse approfondie des mesures d'association pour les séquences de plus de deux mots, celles-ci ayant bien moins retenu l'attention des chercheurs que celles pour les bigrammes (Bestgen, sous presse ; Gries, 2010).

Les analyses effectuées traitent les scores de qualité des textes comme des variables quantitatives présentant un nombre important de valeurs différentes. Ceci contraste avec la manière habituelle de présenter le degré de compétence d'un apprenant au moyen de quelques niveaux différents, par exemple B1 ou C2 dans le CECRL. Il serait intéressant de déterminer si l'approche proposée permet aussi de catégoriser efficacement des apprenants dans ces niveaux. Le pourcentage de classification correcte serait une information plus *parlante* pour les praticiens que les corrélations rapportées ici, qui, comme indiqué ci-dessus, ont été employées en raison de la nature des scores de qualité des textes dans certains des ensembles de données analysés.

Il est également indispensable de déterminer si les indices phraséologiques peuvent améliorer non seulement le niveau de base utilisé ici, mais aussi un modèle prédictif qui inclut de nombreuses autres caractéristiques connues pour leur efficacité. Les indices phraséologiques, mais aussi les statistiques lexicales (diversité et nombre de mots) analysés ne constituent qu'une petite partie des traits linguistiques pertinents pour évaluer la qualité de textes d'apprenants en général ou dans le cadre du modèle *Complexité, Exactitude et Fluence* (Bestgen *et al.*, 2010 ; Bulté et Housen, 2012 ;

Grant et Ginther, 2000 ; Higgins *et al.*, 2015 ; Neumann, 2014). Déterminer les bénéfices éventuels résultant de l'ajout des indices phraséologiques à un système comme e-Rater serait particulièrement informatif étant donné que, comme indiqué dans l'introduction, les premières tentatives dans ce sens des concepteurs du e-Rater ne se sont pas révélées fructueuses (Higgins *et al.*, 2015). Il faut toutefois rappeler que, si le fonctionnement de e-Rater est relativement bien documenté, il s'agit néanmoins d'un logiciel propriétaire.

Il serait aussi intéressant de déterminer si ces indices sont aussi performants dans d'autres langues comme le français. *A priori*, on peut penser que les statistiques lexicales fonctionneront aussi bien. Il en est probablement de même pour les indices phraséologiques. Toutefois, en français, il pourrait être pertinent de lemmatiser les bigrammes avant de procéder à leur recherche dans le corpus de référence. La sélection du corpus de référence elle-même risque de poser quelques problèmes. Le fait que le corpus issu du projet WaCKy est presque aussi performant pour cet usage que le BNC ou le COCA laisse toutefois augurer d'une option au moins satisfaisante en attendant la disponibilité d'un véritable corpus de référence pour le français. Toutefois, le problème le plus important est que nous ne disposons pas de textes d'apprenants du français évalués en terme de qualité. Recueillir ce matériel est le premier point dans notre agenda.

Remerciements

Cette recherche a bénéficié du soutien du Fonds de la recherche scientifique (crédit F.R.S.-FNRS J.0025.16). L'auteur est chercheur qualifié de cette institution. Une partie des ressources informatiques utilisées ont été fournies par les installations de calcul intensif de l'Université catholique de Louvain (CISM/UCL) et du Consortium des Équipements de Calcul intensif en Fédération Wallonie-Bruxelles (CÉCI) financé par le F.R.S.-FNRS.

7. Bibliographie

Baldwin T., Kim S. N., « Multiword Expressions », *in* N. Indurkhya, F. J. Damerau (eds), *Handbook of Natural Language Processing*, CRC Press, p. 267-292, 2010.

Baroni M., Bernardini S., Ferraresi A., Zanchetta E., « The WaCky wide web : A collection of very large linguistically processed web-crawled corpora », *Language Resources and Evaluation*, vol. 43, p. 209-226, 2009.

Bernardini S., « Collocations in translated language. Combining parallel, comparable and reference corpora », *Proceedings of the Corpus Linguistics Conference*, p. 1-16, 2007.

Bestgen Y., « Using collocational features to improve automated scoring of EFL texts », *Proceedings of the 12th Workshop on Multiword Expressions*, p. 84-90, 2016.

Bestgen Y., « Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test », *Corpora*, in press.

Bestgen Y., Granger S., « Quantifying the development of phraseological competence in L2 English writing : An automated approach », *Journal of Second Language Writing*, vol. 26, p. 28-41, 2014.

Bestgen Y., Lories G., Thewissen J., « Using latent semantic analysis to measure coherence in essays by foreign language learners ? », *Proceedings of 10th International Conference on Statistical Analysis of Textual Data*, p. 385-395, 2014.

Bulté B., Housen A., « Defining and operationalising L2 complexity », *in* A. Housen, F. Kuiken, I. Vedder (eds), *Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA*, John Benjamins Publishing, p. 21-46, 2012.

Cavalla C., « La phraséologie en classe de FLE », *Les Langues Modernes*, 2009.

Chodorow M., Leacock C., « An unsupervised method for detecting grammatical errors », *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, p. 140-147, 2000.

Condon W., « Large-scale assessment, locally-developed measures, and automated scoring of essays : Fishing for red herrings ? », *Assessing Writing*, vol. 18, n° 1, p. 100-108, January, 2013.

Connor-Linton J., Polio C., « Comparing perspectives on L2 writing : Multiple analyses of a common corpus », *Journal of Second Language Writing*, vol. 26, p. 1-9, 2014.

Council of Europe, *Common European Framework of Reference for Languages : Learning, Teaching, Assessment.*, Cambridge University Press, 2001.

Cowie A. P., « Phraseology », *in* R. E. Asher (ed.), *The encyclopedia of language and linguistics*, Oxford University Press, p. 3168-3171, 1994.

Deane P., « A nonparametric method for extraction of candidate phrasal terms », *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, p. 605-613, 2005.

Dunning T. E., « Accurate Methods for the Statistics of Surprise and Coincidence », *Computational Linguistics*, vol. 19, p. 61-74, 1993.

Durrant P., Schmitt N., « To what extent do native and non-native writers make use of collocations ? », *International Review of Applied Linguistics in Language Teaching*, vol. 47, p. 157-177, 2009.

Engber C. A., « The relationship of lexical proficiency to the quality of ESL compositions », *Journal of Second Language Writing*, vol. 4, n° 2, p. 139-155, 1995.

Evert S., « Corpora and collocations », *in* A. Lüdeling, M. Kytö (eds), *Corpus Linguistics. An International Handbook*, Mouton de Gruyter, p. 1211-1248, 2008.

Forsberg F., « Using conventional sequences in L2 French », *International Review of Applied Linguistics in Language Teaching*, vol. 48, p. 25-50, 2010.

Futagi Y., Deane P., Chodorow M., Tetreault J., « A computational approach to detecting collocation errors in the writing of non-native speakers of English », *Computer Assisted Language Learning*, vol. 21, p. 353-367, 2008.

Galbraith D., « Cognitive models of writing », *GFL - German as a Foreign Language*, vol. 2, p. 7-22, 2009.

Granger S., Bestgen Y., « The use of collocations by intermediate vs. advanced non-native writers : A bigram-based study », *International Review of Applied Linguistics in Language Teaching*, vol. 52, p. 229-252, 2014.

Granger S., Dagneaux E., Meunier F., Paquot M., *The International Corpus of Learner English. Handbook and CD-ROM. Version 2*, Presses universitaires de Louvain, 2009.

Grant L., Ginther A., « Using computer-tagged linguistic features to describe L2 writing differences », *Journal of Second Language Writing*, vol. 9, p. 123-145, 2000.

Gries S. T., « Useful statistics for corpus linguistics », *in* A. Sánchez, M. Almela (eds), *A Mosaic of Corpus Linguistics : Selected Approaches*, Peter Lang, Frankfurt au Main, Germany, p. 269-291, 2010.

Herrington A., Moran C., « Writing to a Machine is Not Writing At All », *in* N. Elliot, L. Perelman (eds), *Writing Assessment in the 21st Century : Essays in Honor of Edward M. White*, Hampton, p. 425-438, 2014.

Higgins D., Ramineni C., Zechner K., « Learner corpora and automated scoring », *in* S. Granger, G. Gilquin, F. Meunier (eds), *Cambridge Handbook of Learner Corpus Research*, Cambridge University Press, 2015.

Housen A., Kuiken F., Vedder I., « Complexity, accuracy and fluency : Definitions, measurement and research », *in* A. Housen, F. Kuiken, I. Vedder (eds), *Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA*, John Benjamins Publishing, p. 1-20, 2012.

Howell D., *Méthodes statistiques en sciences humaines*, De Boeck Université, Bruxelles, 2008.

Joachims T., « Training Linear SVMs in Linear Time », *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*, 2006.

Kuiken F., Vedder I., « Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking », *in* A. Housen, F. Kuiken, I. Vedder (eds), *Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA*, John Benjamins Publishing, p. 143-170, 2012.

Lenko-Szymanska A., Wolk A., « A corpus-based analysis of the development of phraseological competence in EFL learners using the CollGram profile », *Paper presented at the 7th Conference of the Formulaic Language Research Network (FLaRN)*, 2016.

Liu A. L.-E., Wible D., Tsao N.-L., « Automated suggestions for miscollocations », *Proceedings of the 4th Workshop on Innovative Use of NLP for Building Educational Applications*, p. 47-50, 2009.

Lu X., « The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives », *The Modern Language Journal*, vol. 96, p. 190-208, 2012.

McCarthy P. M., Jarvis S., « MTLD, vocd-D, and HD-D : a validation study of sophisticated approaches to lexical diversity assessment », *Behavior Research Methods*, vol. 42, p. 381-392, 2010.

Myles F., « Complexity, accuracy and fluency : The role played by formulaic sequences in early interlanguage development », *in* A. Housen, F. Kuiken, I. Vedder (eds), *Dimensions of L2 Performance and Proficiency : Complexity, Accuracy and Fluency in SLA*, John Benjamins Publishing, p. 71-94, 2012.

Neumann H., « Teacher assessment of grammatical ability in second language academic writing : A case study », *Journal of Second Language Writing*, vol. 24, p. 83-107, 2014.

Pawley A., Syder F. H., « Two puzzles for linguistic theory : nativelike selection and nativelike fluency », *in* J. C. Richards, R. W. Schmidt (eds), *Language and Communication*, Longman, 1983.

Pecina P., « Lexical association measures and collocation extraction », *Language Resources & Evaluation*, vol. 44, p. 137-158, 2010.

Ramineni C., Williamson D. M., « Automated Essay Scoring : Psychometric Guidelines and Practices », *Assessing Writing*, vol. 18, n° 1, p. 25-39, January, 2013.

Santos V., Verspoor M., Nerbonne J., « Identifying important factors in essay grading using machine learning », *in* D. Sagari, S. Papadima-Sophocleous, S. Ioannou-Georgiou (eds), *International Experiences in Language Testing and Assessment—Selected Papers in Memory of Pavlos Pavlou*, Peter Lang, Frankfurt au Main, Germany, p. 295-309, 2012.

Sinclair J., *Corpus, Concordance, Collocation*, Oxford University Press, 1991.

Smiskova H., Verspoor M., Lowie W., « Conventionalized ways of saying things (CWOSTs) and L2 development », *Dutch Journal of Applied Linguistics*, vol. 1, p. 125-142, 2012.

Somasundaran S., Chodorow M., « Automated measures of specific vocabulary knowledge from constructed responses (use these words to write a sentence based on this picture) », *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 1-11, 2014.

Somasundaran S., Lee C. M., Chodorow M., Wang X., « Automated Scoring of Picture-based Story Narration », *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 42-48, 2015.

Stengers H., Boers F., Housen A., Eyckmans J., « Formulaic sequences and L2 oral proficiency : Does the type of target language influence the association ? », *International Review of Applied Linguistics*, vol. 49, p. 239-263, 2011.

Stubbs M., « Collocations and semantic profiles : On the cause of the trouble with quantitative studies », *Functions of Language*, vol. 2, p. 23-55, 1995.

Thewissen J., « Capturing L2 accuracy developmental patterns : Insights from an error-tagged EFL learner corpus », *Modern Language Journal*, vol. 97, p. 77-101, 2013.

Verspoor M., Schmid M. S., Xu X., « A dynamic usage based perspective on L2 writing », *Journal of Second Language Writing*, vol. 21, p. 239-263, 2012.

Weigle S. C., « English language learners and automated scoring of essays : Critical considerations », *Assessing Writing*, vol. 18, n° 1, p. 85-99, January, 2013.

Wible D., Kwo C.-H., Tsao N.-L., Liu A., Lin H.-L., « Bootstrapping in a language learning environment », *Journal of Computer Assisted Learning*, vol. 19, n° 4, p. 90-102, 2003.

Williamson D. M., Xi X., Breyer F. J., « A framework for evaluation and use of automated scoring », *Educational Measurement : Issues and Practices*, vol. 31, n° 1, p. 2-13, 2012.

Wu J.-C., Chang Y. C., Mitamura T., Chang J. S., « Automatic collocation suggestion in academic writing », *Proceedings of the Association for Computational Linguistics Conference*, p. 115-119, 2010.

Yannakoudakis H., Briscoe T., « Modeling coherence in ESOL learner texts », *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, p. 33-43, 2012.

Yannakoudakis H., Briscoe T., Medlock B., « A New Dataset and Method for Automatically Grading ESOL Texts », *The 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 180-189, 2011.