Indices d'association collocationnelle et évaluation de textes en langue étrangère : comparaison des bigrammes et trigrammes

Yves Bestgen
CECL, Place du Cardinal Mercier 10, B-1348 Louvain-la-Neuve, Belgique
yves.bestgen@uclouvain.be

RÉSUMÉ

Cette recherche a pour principal objectif 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. Les analyses, menées sur plus de 1000 copies d'examen du First Certificate in English, librement mises à disposition par Yannakoudakis et coll., confirment que l'approche qui consiste à assigner aux bigrammes et aux trigrammes de mots présents dans un texte des scores d'association collocationnelle calculés sur la base d'un grand corpus de référence natif est particulièrement efficace. Si les indices extraits des trigrammes sont moins efficaces que ceux extraits des bigrammes, ils apportent une contribution utile à ces derniers. Les analyses soulignent aussi les bénéfices apportés par un emploi simultané de plusieurs mesures d'association collocationnelle.

ABSTRACT

Collocation measures and automated scoring of foreign language texts : Comparing bigrams and trigrams.

The main objective of this study is to assess the utility to take into account automatic measures of the phraseological competence to estimate the quality of texts written by learners of English as a foreign language. The analyzes, carried out on more than 1000 scripts of the First Certificate in English, made freely available by Yannakoudakis and collaborators, confirm that the approach of assigning to bigrams and trigrams of words present in a text association scores computed on the basis of a large native reference corpus is particularly effective. If the features extracted from the trigrams are less effective than those extracted from the bigrams, they make a useful contribution to the latter. The analyzes also highlight the benefits of the simultaneous use of several measures of collocational association.

MOTS-CLÉS : Evaluation automatique de textes ; Expressions conventionnelles ; Mesures d'association collocationnelle ; N-grammes ; Double validation croisée ; Information mutuelle.

KEYWORDS: Automatic text scoring ; Formulaic sequences ; Measures of collocation strength ; N-grams ; Double cross-validation ; Mutual information.

1 Introduction et état de l'art

La mise au point et le déploiement de systèmes automatiques capables d'évaluer efficacement les compétences rédactionnelles en langue apprise sont l'objet d'intenses recherches depuis de nombreuses années en TAL (Weigle, 2013). Ces systèmes se basent classiquement sur des indices linguistiques corrélés avec la qualité du texte. Par exemple, e-Rater, développé par l'Educational

Testing Service (ETS), 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 & Williamson, 2013).

Dans ces systèmes, la dimension phraséologique du langage est très largement négligée. Or, l'importance des unités préformées dans l'emploi du langage est bien établie (Sinclair, 1991). Si certaines de ces séquences relèvent de l'approche traditionnelle en phraséologie, se signalant par leur figement et leur opacité, la très grande majorité de celles-ci sont des manières habituelles de s'exprimer comme *dramatic increase, committed suicide, depend on, such as, by the way* ou encore *as far as I know* (Smiskova *et al.*, 2012). Cette dimension phraséologique du langage a d'importantes conséquences en linguistique (pour une synthèse voir les volumes 32 de l'Annual Review of Applied Linguistics (2012) et 189 de Langages (2013)) et en traitement automatique du langage (pour une synthèse, voir Ramisch, 2015). Tout particulièrement, elle est une composante majeure de l'apprentissage d'une langue étrangère comme l'ont montré de nombreuses études en linguistique appliquée. Elle permet de distinguer les locuteurs natifs des non-natifs et l'emploi d'expressions phraséologiques par des apprenants évolue en parallèle avec la compétence en langue apprise (p. ex., Chen & Baker, 2014 ; Verspoor *et al.*, 2012).

En traitement automatique de textes d'apprenants d'une langue étrangère, les expressions phraséologiques ont presque exclusivement été employées dans le cadre de la détection d'erreurs où elles se sont d'ailleurs montrées extrêmement efficaces (Futagi *et al.*, 2008 ; Wu *et al.*, 2010). La méthode employée est inspirée d'une approche développée pour identifier les erreurs grammaticales dans les textes. Elle signale comme potentiellement problématique les séquences de mots qui sont improbables lorsqu'on les compare à celles qui se trouvent dans un grand corpus natif (Chodorow & Leacock, 2000). Ces séquences peuvent être des erreurs grammaticales, mais aussi des juxtapositions de mots atypiques. Dans le cadre plus spécifique de l'évaluation automatique de la qualité de texte, une variable indicatrice de l'emploi adéquat d'expressions phraséologiques, obtenue de cette manière, a été incluse dans e-Rater, mais son utilité pour estimer la qualité d'un texte est très limitée (Higgins *et al.*, 2015). Récemment toutefois, Somasundaran et Chodorow (2014) et Somasundaran *et al.* (2015) ont mis en évidence les bénéfices apportés par des indices phraséologiques pour évaluer automatiquement des phrases produites à partir de deux mots et d'une image et de brèves narrations produites à partir d'une série d'images. Granger et Bestgen (2014 ; Bestgen & Granger, 2014) et Bestgen (2016) ont abouti aux mêmes conclusions en appliquant une approche très similaire à des textes de différents genres (essais argumentatifs, lettres, histoires rédigées en réponse à un énoncé).

L'approche employée dans ces travaux présente trois caractéristiques importantes. Tout d'abord, elle utilise un grand corpus de référence natif pour essayer de déterminer parmi les séquences de plusieurs mots (n-grammes) présentes dans un texte d'apprenant celles qui sont typiques de la langue de celles qui sont peu ou pas appropriées. Elle supplée ainsi à l'évaluation manuelle des séquences, basée sur des dictionnaires ou sur l'avis de locuteurs natifs, habituellement employée en linguistique appliquée (Verspoor *et al.*, 2012). Comme mentionnée ci-dessus, l'utilisation d'un corpus de référence natif était déjà une des approches employées en TAL pour identifier des erreurs dans des textes (Chodorow & Leacock, 2000). En traductologie, Bernardini (2007) a également employé un corpus de référence natif pour identifier des paires de mots collocationnelles dans des textes traduits et non traduits. Le recours à un corpus natif distingue aussi ces travaux de la simple prise en compte de la fréquence des bigrammes de mots dans des textes d'apprenants comme l'ont proposé Yannakoudakis *et al.* (2011 ; Yannakoudakis & Briscoe, 2012). L'approche développée par ces auteurs pour estimer la qualité

d'un texte est basée sur l'extraction automatique d'un grand nombre d'indices linguistiques des plus simples comme les fréquences des unigrammes et bigrammes de mots aux plus complexes comme des indices extraits d'une analyse syntaxique des phrases ou reflétant la cohérence et la cohésion des textes. Le modèle prédictif proprement dit est construit par une procédure d'apprentissage supervisé (voir section 3.3). Dans leur étude, Yannakoudakis *et al.* (2011) ne font pas référence à la compétence phraséologique, mais les bigrammes de mots qu'ils utilisent sont évidemment liés à celle-ci.

Ensuite, pour évaluer le caractère plus ou moins typique d'une séquence de mots dans la langue, l'approche se base sur des indices d'association collocationnelle comme l'information mutuelle (IM) ou le score-t (Evert, 2008), qui sont obtenus à partir des fréquences d'occurrences dans le corpus de référence. Ils permettent de ne pas se fier exclusivement à la fréquence de la séquence dont on sait qu'elle peut être trompeuse parce qu'une séquence peut-être observée fréquemment dans un corpus, non en raison de sa nature phraséologique, mais parce qu'elle est composée de mots très fréquents (Evert, 2008). Inversement, une séquence relativement rare, composée de mots eux aussi rares, peut être typique de la langue.

Enfin, un ensemble d'indices statistiques sont extraits des scores d'association collocationnelle assignés aux n-grammes présents dans un texte à analyser. Il s'agit de statistiques globales, comme la moyenne ou le maximum, ainsi que de la proportion de scores d'association se situant dans des intervalles donnés. L'intérêt majeur d'une discrétisation des distributions de scores est qu'elle permet d'extraire des informations utiles même lorsque la relation entre ces scores et la qualité des textes n'est pas linéaire ou même monotone. Granger et Bestgen (2014) ont ainsi observé que les meilleurs textes d'apprenant contiennent plus de bigrammes ayant un score t moyen et moins de bigrammes ayant un score t faible ou élevé. Cette observation peut être mise en relation avec le fait que les bigrammes ayant des scores t faibles sont souvent des formulations erronées et que les scores élevés correspondent à des bigrammes extrêmement courants dans la langue comme *of the* dont l'emploi pose peu de problèmes à l'apprenant.

Somasundaran et Chodorow (2014) et Somasundaran *et al.* (2015) ont employé cette approche pour analyser de brèves productions orales. Ils ont affecté aux bigrammes et aux trigrammes présents dans les réponses orales retranscrites par un système automatique les scores IM calculés sur la base de *Google 1T web corpus*. Ils ont discrétisé les distributions des scores en huit intervalles déterminés manuellement comme [-inf,-20],]-20,-10], ou]-1,0]. 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 basée principalement sur des traits acoustiques extraits automatiquement. Bestgen (2016) a extrait les bigrammes présents dans des textes de 200 à 400 mots du Cambridge Learner Corpus, un corpus mis librement à disposition par Yannakoudakis *et al.* (2011). Il a analysé huit indices d'association collocationnelle différents ainsi que la combinaison de ceux-ci et les a discrétisés au moyen d'une procédure automatique. Utilisant également une procédure d'apprentissage supervisé, il a montré 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. Il a aussi observé qu'une discrétisation en 8 à 50 intervalles est optimale.

2 Questions de recherche

Si ces études ont employé la même approche, elles se distinguent sur deux points importants :

— Les analyses portent soit sur les seuls bigrammes, soit sur les bigrammes et sur les trigrammes.
— Une seule mesure d'association collocationnelle est employée ou bien plusieurs mesures sont combinées afin d'obtenir les meilleures performances.

L'objectif de la présente recherche est d'évaluer l'impact de ces deux facteurs sur les performances de la procédure. Il s'agit de questions importantes, non seulement afin d'essayer d'améliorer les performances de la procédure, mais aussi parce qu'elles sont susceptibles d'orienter les recherches futures dans le domaine. En premier lieu, l'intérêt de prendre en compte les trigrammes, seuls ou en combinaison avec les bigrammes n'a pas été évalué. Or, il est bien établi que nombre d'expressions phraséologiques comptent plus de deux mots et que, souvent, les séquences de deux mots fortement associées sont des tronçons d'expressions plus longues. Certains linguistes considèrent même que les séquences de deux mots sont trop courtes pour donner lieu à des analyses intéressantes (Conrad & Biber, 2004, p. 58). Il est donc possible que les trigrammes soient plus performants que les bigrammes.

Ensuite, les mesures d'association pour les n-grammes de plus de 2 mots ont retenu nettement moins l'attention que celles pour les bigrammes (Evert, 2008 ; Nerima *et al.*, 2010). Mettre en évidence l'utilité des trigrammes dans ce genre de tâches pourrait favoriser une intensification de la réflexion à propos des mesures d'association pour de telles séquences.

Enfin, l'information mutuelle, la mesure largement privilégiée dans ces études, a été fréquemment critiquée en raison de sa tendance à privilégier des expressions très rares et les difficultés posées par sa généralisation à des séquences de plus de deux mots ont été soulignées (Biber, 2009, mais voir Van de Cruys (2011) pour une analyse approfondie). L'efficacité de ces mesures, mais aussi des autres mesures employées par Bestgen (2016), lorsqu'elles sont appliquées aux trigrammes, mérite donc d'être analysée. Tout particulièrement, on peut se demander si la simple fréquence d'occurrence de la séquence, qui est un indice d'association peu fiable dans le cas des bigrammes comme rappelé ci-dessus, ne devient pas un indice compétitif dans le cas des trigrammes parce que plus une séquence est longue, moins elle est susceptible de se produire par hasard, même quand elle est composée de mots fréquents. Une réponse positive à cette question s'accorderait avec l'approche classiquement employée en linguistique appliquée pour étudier les expressions conventionnelles qui est basée sur les paquets lexicaux (*lexical bundles*), définis comme des n-grammes de plus de deux mots identifiés sur la base de ce critère de fréquence (Conrad & Biber, 2004).

Les sections suivantes tentent de répondre à ces questions en comparant l'efficacité, pour prédire la qualité de textes, de mesures d'association obtenues à partir des bigrammes et des trigrammes de mots. Cette efficacité sera aussi comparée à celle atteinte par des niveaux de base composés d'ensembles de n-grammes auxquels aucun score d'association n'est attribué.

3 Méthode

3.1 Ensemble de données analysé

L'ensemble de données analysé est constitué des copies d'examen du *First Certificate in English* (FCE), décrit dans Yannakoudakis *et al.* (2011). Ce test, qui fait partie du *Cambridge English Language Assessment for Speakers of Other Languages*, est conçu pour évaluer les apprenants de niveau intermédiaire de l'anglais, qui relèvent du niveau B2 dans le Cadre européen commun de référence pour les langues. L'ensemble de données se compose de 1235 documents de 200 à 400 mots extraits de la section écrite de l'examen, pour un total de 460 964 mots. Chaque document comprend

deux textes écrits par le même apprenant en réponse à des énoncés, le premier étant obligatoirement une lettre et le second un choix libre entre une autre lettre ou une composition, une histoire, un rapport ou un article de magazine. Un énoncé typique pour la première tâche était : *Votre classe d'anglais va passer trois jours à Londres. Le directeur de votre collège, M. Robertson, a déjà organisé le programme. Cependant, les étudiants de votre classe ont vu une publicité pour le London Fashion and Leisure Show et vous aimeriez tous aller au spectacle. Votre classe vous a demandé d'écrire à M. Robertson à ce sujet. Lisez l'extrait du programme de M. Robertson, la publicité et vos notes. Puis, en utilisant ces informations, écrivez une lettre à M. Robertson*. Un énoncé fréquent pour une histoire (15% de tous les textes pour la tâche 2) est : *Votre professeur vous a demandé d'écrire une histoire pour Revue de langue anglaise de l'école. L'histoire doit commencer par les mots suivants : Malheureusement, Pat n'était pas très douée pour garder des secrets. Rédigez votre histoire.* Les participants ont eu 80 minutes pour écrire les deux textes. Ils étaient de 16 langues maternelles différentes, les plus fréquentes étant l'espagnol (16%) et le français (12%), mais il y avait aussi au moins 5% d'allemand, chinois, japonais, coréen, thaï, turc et polonais. Les deux tiers des candidats étaient âgés de 16 à 25 ans.

Dans le cadre de l'évaluation certificative menée au moyen du *First Certificate in English*, une note globale a été attribuée à chaque document sur une échelle allant de 0 à 40. Les critères de notation reposent principalement sur le succès des apprenants dans la tâche de communication qu'ils ont dû accomplir en mettant l'accent sur l'organisation et la cohésion du texte, la mise en page, l'emploi du registre de langue approprié, la correction et la richesse de la langue. Les scores se distribuent sur toute l'échelle avec une moyenne de 27,92 et un écart-type de 5,34.

Deux textes, rédigés en réponse à la première consigne donnée plus haut, sont reproduits ci-dessous de manière à donner un aperçu du matériel qui sera analysé. Le premier texte a obtenu un score de qualité de 11, soit une des valeurs les plus faibles de l'ensemble de données. Le deuxième texte (abrégé ici à la longueur du premier) a obtenu un score de qualité de 39, soit pratiquement le maximum possible (malgré qu'il contienne l'une ou l'autre erreur). Ce sont ces scores de qualité, attribués par des évaluateurs experts, que la procédure automatique proposée essayera de prédire le mieux possible.

— TR964*0100*2000*02 Score=11 : *Mr. Robertson : First of all our class would like to give you our special thanks about the programme you gave us yesterday afternoon. After reading the hole organized programme, we all agreed about the sightseeing by bus around London wich will give us the knowledge about this fantastic city we all want to meet since time ago. The main purpose of our letter it's to explain about the London Fashion and Leisure Show. We found the advertisment in a local newspaper article and have been thinking it could be such a good opportunity, because the show consists on these topics : Latest Fashions — Leisure and sports wear — Make up, Hairstyles. On the other hand, and the most important, it is free for students. After being discussing , we want to suggest how we think the programme should be re-arranged. Instead of going to the science museum, we could go to the London Fashion and Leisure Show wich is opened on Tuesday, march 14 from 10 am to 19 pm . After that we could continue our trip as it had been arranged previously . Yours sincirely*

— TR875*0100*2000*02 Score=39 : *Dear Mr Robertson, I would like first, in the name of my whole class, to thank you for all you've done for us and for our trip to London. We've just been told about your programme and I must say that we all like it ! What I am most looking forward to , is the River trip to Greenwich ; but I would also really like to visit the Science Museum as well as the National Gallery ! This will be very enriching for us ! You must wonder why I'm writing to you, them ! ... well actually we've must received an advertisement for the London Fashion and Leisure Show, and the vast majority of the class would like to go and*

see it. This will take place in the Central Exhibition Hall on Tuesday 14th March. We thought it would be a great experience for us since we quite don't have such opportunities in our everyday life ! Moreover, the entrance is free for the students so that won't raise the cost for the trip. I thought we should visit the Science Museum on Monday afternoon so that [...]

3.2 Corpus de référence

Pour estimer la force collocationnelle d'un n-gramme, un corpus de référence suffisamment grand et aussi représentatif que possible de la langue telle qu'elle est employée par les locuteurs natifs est nécessaire. Nous avons choisi d'utiliser 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, parce que les auteurs des textes analysés se sont inscrits à un examen britannique.

3.3 Construction des modèles prédictifs

Les modèles prédictifs ont été construits par apprentissage supervisé sur la base de deux ensembles de traits : phraséologiques 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 de textes dans l'ensemble de données FCE. 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.

3.3.1 Indices phraséologiques

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

— Extraction des n-grammes. Tous les bigrammes et trigrammes 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 l'extraction. Seuls les n-grammes différents (types) dans un texte sont analysés afin de donner plus de poids à leur diversité.

— Obtention des scores d'association. Chaque n-gramme 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 du corpus (Evert, 2008 ; Pecina, 2010), sont calculées : IM et score-t (Church *et al.*, 1991), score-z (Berry-Rogghe, 1973), simple-ll (Evert, 2008), IM3 (Daille, 1994) et la fréquence brute [1]. Pour les bigrammes, toutes ces mesures d'association ont été calculées au moyen des formules données dans Evert (2008). Les mêmes formules ont été employées pour les trigrammes, la fréquence attendue étant calculée au moyen de l'extension classique de la formule pour les bigrammes (Ramish, 2015, p. 64). Enfin, si le n-gramme n'est pas trouvé dans le corpus de référence, il est pris en compte dans le calcul de la proportion de bigrammes ou de trigrammes présents dans un texte, mais absents du corpus de référence.

1. Les analyses rapportées ci-dessous ont également été effectuées après avoir appliqué une transformation logarithmique aux fréquences brutes, mais les performances étaient presque identiques (en fait très légèrement inférieures) à celles obtenues sur la base des simples fréquences, ce qui n'est pas étonnant parce que cette transformation monotone n'affecte que les statistiques descriptives globales

— Extraction des traits phraséologiques. Ils sont composés des trois types suivants, extraits séparément sur la base des bigrammes et des trigrammes :
  — La proportion de n-grammes 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 (Dougherty *et al.*, 1995). Cette procédure est non supervisée et ne dépend que d'un seul paramètre, le nombre d'intervalles qui a été fixé à 20, une valeur particulièrement efficace selon l'analyse comparative de Bestgen (2016). Ensuite, on compte dans chaque texte la proportion de n-grammes dont le score d'association se trouve dans chaque intervalle en employant comme dénominateur le nombre total de n-grammes présents dans le texte.

Tous ces traits sont fournis à la procédure d'apprentissage supervisée pour chaque indice d'association analysé isolément. Pour la condition *Tous*, les traits extraits des six indices d'association sont employés dans une seule et même analyse. Un des bénéfices apportés par l'extraction de traits phraséologiques au moyen de scores d'association calculés sur la base d'un corpus de référence est que le nombre de traits ne dépend en aucune manière de la longueur du n-gramme analysé. Dans le cas présent, tant pour les bigrammes que pour les trigrammes, le nombre de traits est de 25 pour les six scores d'association considérés isolément et de 145 lorsqu'ils sont utilisés simultanément, la proportion de n-grammes absents du corpus de référence étant la même pour tous les scores d'association.

3.3.2 Niveaux de base

Comme niveaux de base, on a employé des traits lexicaux et morphosyntaxiques 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, bigrammes et trigrammes de mots et d'étiquettes morphosyntaxiques (PoS-Tag) dans chaque texte. Ces derniers sont issus du prétraitement du corpus par Claws 7 [2]. Yannakoudakis *et al.* (2011) n'ont pas employé les trigrammes de mots, mais ceux-ci sont particulièrement pertinents dans la présente étude pour évaluer l'utilité des scores d'association calculés sur des trigrammes. Nous les avons donc aussi extraits. Ces traits ont été obtenus au moyen de la procédure décrite dans Yannakoudakis *et al.* (2011) et donc en employant un seuil de fréquence minimal de 4 et une pondération de la fréquence par TF-IDF.

3.4 Évaluation de l'efficacité des modèles prédictifs

3.4.1 Mesure d'efficacité

Étant donné que les textes ont été évalués sur une échelle allant de 0 à 40, 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.

2. http ://ucrel.lancs.ac.uk/claws/

3.4.2 Estimation de l'efficacité par validation croisée

Pour estimer l'efficacité des modèles prédictifs et fixer le métaparamètre C de régularisation qui ajuste le rapport entre les capacités de généralisation du modèle et l'efficacité sur le matériel d'apprentissage, une procédure de double validation croisée emboîtée a été employée. La boucle externe estime l'efficacité du modèle en divisant aléatoirement l'ensemble de données en dix blocs d'une taille aussi égale que possible. Chacun de ceux-ci est employé 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 division effectuée, cette procédure est répétée dix fois en employant des germes différents pour le générateur aléatoire. La mesure d'efficacité est la moyenne des valeurs ainsi obtenues.

La boucle interne fixe le paramètre C pour chaque itération de la boucle externe. Chaque échantillon d'apprentissage employé dans une itération externe (et donc composé des neuf blocs initiaux) est divisé aléatoirement en cinq sous-blocs. Chacun de ceux-ci sert 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 sont affectées : 0,00001, 0,0001, 0,001, 0,01, 0,1 et 1. Cette procédure est répétée cinq fois en employant des germes différents pour le générateur aléatoire. L'efficacité est déterminée en calculant la moyenne des 25 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, comme cela est illustré à la section 4.3, mais aussi, dans ce dernier cas, de pouvoir tester la significativité statistique des différences d'efficacité entre les modèles puisque dix valeurs sont 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. Le test de comparaison de moyennes de Student pour échantillons appariés a été employé. Dans la suite, toutes les différences mentionnées sont statistiquement significatives au seuil de 0,0001 sauf indication contraire. Un seuil de signification strict est employé en raison du grand nombre de tests effectués.

4 Analyses et résultats

4.1 Comparaison des mesures d'association appliquées aux bigrammes et aux trigrammes

Les premières analyses visent à comparer l'efficacité des traits phraséologiques extraits des bigrammes et des trigrammes ainsi que de leur combinaison. Le tableau 1 donne l'ensemble des valeurs de corrélation entre les scores de qualité des textes attribués par des juges experts et les scores de qualité prédits sur la base des traits phraséologiques. Si on s'intéresse d'abord aux colonnes du tableau, les traits phraséologiques extraits des bigrammes sont plus efficaces pour prédire la qualité des textes que ceux extraits des trigrammes, sauf dans le cas de la fréquence, mais il s'agit aussi de l'indice de loin le moins efficace. Toutefois, la combinaison des trigrammes et des bigrammes produit de meilleures corrélations que l'emploi des seuls bigrammes.

Les différents indices d'association donnent lieu à des corrélations très similaires dans le cas des bigrammes alors que des différences plus importantes sont observées pour les trigrammes et la

combinaison des deux longueurs de n-grammes, mais seules quelques-unes de ces différences sont statistiquement significatives. Pour les trigrammes, IM est significativement plus corrélé avec la qualité des textes que IM3 et que le score-z. Pour la combinaison des deux longueurs, simple-ll l'est par rapport à IM et le score-z.

Lorsqu'on considère l'ensemble des corrélations obtenues, c'est la combinaison de tous les indices d'association qui produit les meilleurs résultats. On notera toutefois que le gain est plus faible pour les bigrammes que pour les trigrammes et leur combinaison.

La dernière ligne du tableau présente les corrélations obtenues au moyen des n-grammes de mots de chaque longueur (bigrammes ou trigrammes) et de leur combinaison, sans employer donc d'indices d'association collocationnelle. Si ceux-ci sont généralement plus efficaces que chaque mesure d'association prise individuellement (mais les seules différences significatives s'observent dans le cas des trigrammes), ils sont moins efficaces que les modèles qui sont basés sur l'ensemble des indices d'association. Ce résultat souligne l'intérêt d'employer des scores d'association calculés sur la base d'un corpus de référence natif pour prédire la qualité de textes rédigés en anglais langue étrangère. Il souligne aussi l'utilité de combiner plusieurs indices d'association.

Condition	Bigrammes	Trigrammes	Bi- et Trigrammes
Fréquence	0,415	0,447	0,524
IM	0,541	0,486	0,543
IM3	0,540	0,462	0,567
Simple-ll	0,548	0,477	0,570
Score-t	0,543	0,479	0,570
Score-z	0,541	0,460	0,552
Tous	0,562	0,522	0,591
N-grammes lexicaux	0,549	0,487	0,555

TABLE 1 – Performances (corrélation) des traits phraséologiques

.

4.2 Comparaison des mesures d'association et des traits lexicaux et morphosyntaxiques

Afin de sélectionner le niveau de base pour la comparaison des mesures d'association et des n-grammes lexicaux et morphosyntaxiques, le tableau 2 présente les corrélations avec le jugement de qualité des textes pour une série de modèles prédictifs basés sur des combinaisons de traits lexicaux et morphosyntaxiques pertinentes pour la présente étude. On observe que les traits morphosyntaxiques sont légèrement plus efficaces que les traits lexicaux, mais la différence n'est pas statistiquement significative. La meilleure combinaison de n-grammes lexicaux inclut les trois longueurs, mais elle est juste meilleure que celle qui ne se base pas sur les trigrammes. Lorsque les n-grammes d'étiquettes morphosyntaxiques sont ajoutés, c'est la combinaison, employée par Yannakoudakis *et al.* (2011), qui n'inclut pas les trigrammes de mots qui est très légèrement la plus efficace et qui sera donc employée dans la suite comme niveau de base. Cette observation souligne qu'ajouter des prédicteurs ne produit pas toujours une amélioration de la performance, la condition basée sur l'ensemble des n-grammes lexicaux et morphosyntaxiques n'étant pas plus efficace que la combinaison sélectionnée.

Le tableau 2 donne aussi le nombre de traits correspondant à chaque ensemble de prédicteurs. On

remarque que les trigrammes de mots fournissent moins de traits que les bigrammes, en raison du seuil de fréquence minimale fixé à 4 comme dans Yannakoudakis *et al.* (2011). On note aussi que la condition contrôle sélectionnée est composée de moins de traits que deux autres ensembles de prédicteurs testés.

Lexical			Morpho.			#traits	r
1	2	3	1	2	3		
X						3513	0,486
	X					12690	0,549
		X				10947	0,487
X	X					16203	0,572
	X	X				23637	0,555
X	X	X				27150	0,578
			X	X	X	8189	0,585
X	X		X	X	X	24392	0,622
X	X	X	X	X	X	35339	0,620

TABLE 2 – Nombre de traits et performances (corrélation) des traits lexicaux et morphosyntaxiques

Les traits lexicaux et morphosyntaxiques, uni- et bigrammes de mots et uni-, bi- et trigrammes d'étiquettes morphosyntaxiques permettent d'obtenir une corrélation de 0.622 avec la qualité des textes. Cette valeur est significativement supérieure à la meilleure corrélation obtenue au moyen d'indices phraséologiques (0,591 pour *Tous* sur la base des bi- et trigrammes).

Condition	Bigrammes	Trigrammes	Bi- et Trigrammes
Fréquence	0,628	0,632	0,637
IM	0,643	0,640	0,655
IM3	0,635	0,635	0,646
Simple-ll	0,636	0,636	0,647
Score-t	0,636	0,635	0,647
Score-z	0,637	0,636	0,647
Tous	0,665	0,664	0,681

TABLE 3 – Performances (corrélation) des traits phraséologiques, lexicaux et morphosyntaxiques

Comme le montre le tableau 3, la meilleure corrélation (0,681), significativement supérieure à toutes les autres, est obtenue en combinant, aux traits lexicaux et morphosyntaxiques, l'ensemble des traits phraséologiques. Ajouter ces indices phraséologiques permet d'améliorer fortement la qualité de la prédiction, le gain en termes de variabilité des scores de qualité réels expliquée par le modèle prédictif étant de 8% (différence entre les R^2, Howell, 2008 : 254-257). Ce résultat confirme donc l'utilité des scores d'association lexicale pour la tâche en jeu. Il est intéressant de comparer cette corrélation avec la valeur la plus élevée rapportée par Yannakoudakis et Briscoe (2012) même si on ignore comment ces auteurs ont fixé le paramètre de régularisation. Leur meilleur modèle, qui incluait les mêmes n-grammes lexicaux et morphosyntaxiques (mais non les traits phraséologiques), ainsi qu'une série d'indices extraits d'une analyse syntaxique des phrases, d'erreurs potentielles détectées automatiquement et issus d'une analyse automatique de la cohérence et de la cohésion des textes, a obtenu une corrélation de 0,677 avec la qualité des textes selon les experts. La performance du modèle proposé dans la présente recherche est d'autant plus remarquable qu'il n'inclut pas nombre

de ces prédicteurs.

Le tableau 3 indique aussi qu'il y a peu de différences entre les mesures d'association lorsqu'elles sont employées isolément, même si la fréquence brute reste la moins efficace et que IM est systématiquement la plus efficace. En raison de la très grande stabilité des prédictions, induite par la présence dans tous les modèles des nombreux traits lexicaux et morphosyntaxiques, ces différences sont, malgré leur faiblesse, statistiquement significatives.

4.3 Utilité de la procédure de double validation croisée

Dans cette étude, une procédure de double validation croisée emboîtée a été employée pour estimer l'efficacité des modèles prédictifs et fixer le métaparamètre C de régularisation. Il s'agit de l'approche la plus objective, mais elle est coûteuse en temps-calcul en raison du très grand nombre de modèles qui doivent être construits, les résultats présentés ici ayant nécessité 76 500 appels de l'algorithme d'apprentissage supervisé. On peut légitimement se demander si une procédure de validation croisée simple, n'employant donc pas de boucle interne pour fixer le paramètre C, mais le fixant d'une manière arbitraire, n'aurait pas permis d'atteindre des résultats similaires pour un temps calcul nettement moindre puisque le nombre de modèles à construire aurait été divisé par 150.

Des analyses complémentaires ont été menées pour répondre à cette question en comparant les corrélations obtenues à la suite de la procédure de double validation croisée à celles obtenues en employant, dans la seule boucle externe, les six valeurs de C qui sont systématiquement testées. Le tableau 4 présente d'abord les résultats pour la comparaison de la condition combinant le niveau de base et toutes les mesures d'association pour les bigrammes et de la condition combinant le niveau de base et toutes les mesures d'association pour les bigrammes et trigrammes (les première et troisième valeurs de la septième ligne du tableau 3). Cette comparaison porte sur un cas dans lequel les traits employés sont très nombreux et très largement les mêmes dans les deux conditions (24 537 sur 24682 sont communs, soit 99,4%). On observe que la double validation croisée est peu utile : un choix adéquat pour la valeur de C est aisé. La deuxième comparaison porte sur la condition combinant toutes les mesures d'association pour les bi- et trigrammes et sur la condition basée sur les bi- et trigrammes de mots (les troisièmes valeurs des lignes 7 et 8 du tableau 1). Ces conditions incluent des ensembles de traits très différents et en nombre variable (290 versus 23 637). Comme l'indique la ligne *Différence*, suivant le choix de la valeur de C, les conclusions sont diamétralement opposées. La double validation croisée, lorsqu'elle est computationnellement faisable, présente donc un avantage certain en raison de son objectivité, mais elle n'est pas toujours indispensable.

Condition	Double	Paramètre C					
	CV	0,00001	0,0001	0,001	0,01	0,1	1
Niveau de base + Tous_23	0,681	0,634	0,681	0,675	0,670	0,670	0,670
Niveau de base + Tous_2	0,665	0,612	0,667	0,660	0,654	0,655	0,655
Différence	0,016	0,022	0,015	0,015	0,015	0,015	0,015
Tous_23	0,591	0,532	0,559	0,591	0,569	0,539	0,517
Bi- + trigrammes de mots	0,555	0,460	0,555	0,554	0,548	0,549	0,549
Différence	0,036	0,072	0,003	0,036	0,021	-0,010	-0,032

TABLE 4 – Performances (corrélation) de différentes conditions en fonction de la valeur de C. Note : Tous_23 correspond à tous les indices phraséologiques pour les bi- et trigrammes.

5 Discussion et conclusion

L'objectif de la présente recherche est 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. Les analyses ont montré que l'approche qui consiste à assigner aux bigrammes et aux trigrammes de mots présents dans un texte des scores d'association collocationnelle calculés sur la base d'un grand corpus de référence natif est particulièrement efficace. Si les indices extraits des trigrammes sont moins efficaces que ceux extraits des bigrammes, ils apportent une contribution utile à ces derniers, validant la proposition de Somasundaran et Chodorow (2014) de combiner des n-grammes de plusieurs longueurs. Ces indices semblent [3] même apporter une contribution plus importante à l'efficacité de la prédiction qu'une combinaison d'indices syntaxiques et discursifs (Yannakoudakis & Briscoe, 2012). Il s'agit là du résultat le plus intéressant de cette étude et qui va au-delà des espoirs initiaux.

Les analyses soulignent aussi l'utilité d'employer simultanément plusieurs mesures d'associations, rejoignant les recommandations de Evert (2008) et de Pecina (2010) dans d'autres domaines d'application. Il faut toutefois reconnaître que l'étude présente à ce sujet plusieurs faiblesses. On n'a employé qu'une poignée d'indices classiques et surtout on les a combinés tous, laissant à la procédure d'apprentissage supervisé le travail de construire le meilleur modèle. Il est possible et même probable que certains indices n'apportent pas de contribution utile à la prédiction. Si on n'a pas tenté de les identifier, c'est en raison du très grand nombre de modèles qu'il serait nécessaire d'évaluer. De plus, une autre question semble bien plus importante : est-ce que d'autres indices, plus complexes ou développés spécifiquement pour traiter les n-grammes de plus de deux mots (Bestgen, sous presse ; Koch, 2015) n'apporteraient pas un surcroit d'efficacité ?

Une autre limitation de l'approche employée pour extraire les traits phraséologiques est qu'elle repose sur la combinaison en une vingtaine de traits quantitatifs de l'ensemble des informations apportées par les scores d'association des nombreux bigrammes et trigrammes présents dans chaque texte. Il s'ensuit qu'il n'est pas possible de donner des exemples de trigrammes particulièrement efficaces pour prédire la qualité d'un texte selon les indices phraséologiques. Il est par contre possible, comme le montre le tableau 5, de donner des exemples de trigrammes ayant reçu des scores très élevés ou très faibles pour chacun des six indices employés. La tendance de IM à privilégier des séquences composées de mots rares y est manifeste de même que celle du score-t, de simple-ll et de la fréquence brute à privilégier des séquences composées de mots fréquents. On note aussi que les séquences qui obtiennent les scores d'association les plus faibles sont clairement problématiques et similaires pour tous les indices, ce qui pourrait être à l'origine, en partie au moins, du peu de différences entre les mesures d'association quant à leur efficacité. Ces séquences ne sont pas données pour la fréquence brute parce qu'il s'agit de séquences qui ne sont observées qu'une seule fois dans le corpus de référence. Elles sont donc très nombreuses et elles sont toutes aussi extrêmes les unes que les autres.

En ce qui concerne les niveaux de bases, les traits lexicaux et morphosyntaxiques sont plus efficaces que les traits phraséologiques, mais les traits lexicaux sont beaucoup plus difficilement généralisables à des textes d'autres genres que les traits phraséologiques (Bestgen, 2017), imposant de réentraîner le modèle prédictif pour tout nouvel ensemble de données. On peut penser que les traits morphosyntaxiques se généralisent mieux, mais le tableau 2 montre qu'employés seuls, ils sont nettement moins efficaces (0,578) que la combinaison des traits lexicaux et morphosyntaxiques (0,622).

3. La prudence est de mise puisqu'on ne peut pas garantir que les expérimentations effectuées sont parfaitement comparables, tout particulièrement la fixation du paramètre C

Indice	Type	Exemples
mi	+	*pros_and_cons spread_like_wildfire door_bell_rang spanish_civil_war burst_into_tears*
	-	*the_to_have were_and_of and_the_had an_in_the in_were_the*
st	+	*i_do_n't do_n't_know one_of_the the_end_of as_well_as*
	-	*the_to_have and_the_had with_the_in that_the_the in_on_of*
mi3	+	*i_do_n't do_n't_know per_cent_of as_well_as the_united_states*
	-	*the_to_have were_and_of and_the_had an_in_the in_were_the*
sz	+	*m_gon_na pros_and_cons second_world_war and_vice_versa burst_into_tears*
	-	*that_the_the in_to_the and_of_the to_is_the in_in_the*
ll	+	*i_do_n't as_well_as do_n't_know one_of_the per_cent_of*
	-	*that_the_the in_to_the to_is_the and_of_the on_of_the*
fg	+	*i_do_n't one_of_the the_end_of as_well_as part_of_the*

TABLE 5 – Exemples de trigrammes pour les 6 indices d'association

Une des questions de recherche portait sur la possibilité que la fréquence brute soit un indice nettement meilleur pour les trigrammes que pour les bigrammes. Les analyses confirment cette hypothèse, la fréquence étant le seul indice évalué dont la performance s'améliore avec la longueur du n-gramme. Il n'en reste pas moins l'indice le moins efficace. Évaluer également des quadrigrammes aurait peut-être permis de répondre encore plus positivement à cette question. Plus généralement, on peut se demander si les quadrigrammes n'apporteraient pas eux aussi une contribution utile à la prédiction. Par principe, plus un n-gramme est long et moins souvent il est employé par les apprenants, mais le recours à un grand corpus de référence natif permet de remédier à ce problème en attribuant à ces séquences des scores d'association. Prendre en compte les quadrigrammes s'accorderait avec les nombreux travaux en linguistique appliquée qui ont montré que les locuteurs natifs et non natifs, mais aussi les apprenants à différents niveaux de compétence, pouvaient être distingués en analysant l'emploi de paquets lexicaux composés de 4 mots (p. ex., Chen & Baker, 2014).

Une propriété importante de l'ensemble de données n'a pas été exploitée dans les analyses effectuées : la langue maternelle des apprenants qui est très diversifiée comme indiqué à la section 3.1. Or, plusieurs études ont mis en évidence l'impact du transfert depuis la langue maternelle sur l'emploi de collocations et de n-grammes de mots et leur efficacité pour distinguer la langue maternelle d'apprenants de l'anglais (p.ex., Brooke & Hirst, 2012 ; Jarvis *et al.*, 2013). Il serait donc intéressant d'introduire cette variable dans les analyses afin de déterminer si l'efficacité des traits phraséologiques est identique dans tous les groupes qui peuvent être formés sur cette base.

Enfin, la technique employée pour obtenir les indices phraséologiques est basée sur les séquences contiguës de mots. Les collocations, telles que définies classiquement en linguistique (Sinclair, 1991, p. 170), n'imposent pas une telle contrainte. Au moins deux options sont envisageables pour s'en affranchir : calculer des scores d'association entre des mots non contigus, en prenant en compte les catégories morphosyntaxiques, ou autoriser la présence de variations dans la séquence comme dans les cadres collocationnels (Renouf & Sinclair, 1991).

Remerciements

L'auteur souhaite remercier les relecteurs pour les commentaires qui ont permis d'étoffer le texte en abordant des aspects absents de la première version. 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 (CECI) financé par le F.R.S.-FNRS.

Références

BERNARDINI S. (2007). Collocations in translated language. combining parallel, comparable and reference corpora. In *Proceedings of the Corpus Linguistics Conference*, p. 1–16.

BERRY-ROGGHE G. L. M. (1973). The computation of collocations and their relevance in lexical studies. In A. J. AITKEN, R. W. BAILEY & N. HAMILTON-SMITH, Eds., *The Computer and Literary Studies*. Edinburgh University Press.

BESTGEN Y. (2016). Using collocational features to improve automated scoring of EFL texts. In *Proceedings of the 12th Workshop on Multiword Expressions*, p. 84–90.

BESTGEN Y. (2017). Validation interne et externe d'indices phraséologiques pour l'évaluation automatique de textes rédigés en anglais langue étrangère. *Traitement automatique des langues*, **57**(3).

BESTGEN Y. (in press). Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test. *Corpora*, **13**(3).

BESTGEN Y. & GRANGER S. (2014). Quantifying the development of phraseological competence in L2 English writing : An automated approach. *Journal of Second Language Writing*, **26**, 28–41.

BIBER D. (2009). A corpus-driven approach to formulaic language in English : Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, **14**, 275–311.

BROOKE J. & HIRST G. (2012). Robust, lexicalized native language identification. In *Proceedings of COLING 2012*, p. 391–408 : John Benjamins Publishing.

CHEN Y.-H. & BAKER P. (2014). Investigating criterial discourse features across second language development : Lexical bundles in rated learner essays. *Applied Linguistics*, **37**(6), 849–880.

CHODOROW M. & LEACOCK C. (2000). An unsupervised method for detecting grammatical errors. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)*, p. 140–147.

CHURCH K., GALE W. A., HANKS P. & HINDLE D. (1991). Using statistics in lexical analysis. In U. ZERNIK, Ed., *Lexical Acquisition : Using On-line Resources to Build a Lexicon*, p. 115–164. Lawrence Erlbaum.

CONRAD S. & BIBER D. (2004). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica*, **20**(1), 56–71.

DAILLE B. (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques*. PhD thesis, Université Paris 7.

DOUGHERTY J., KOHAVI R. & SAHAMI M. (1995). Supervised and unsupervised discretization of continuous features. In *Proceedings of 12th ICML*, p. 194–202.

EVERT S. (2008). Corpora and collocations. In A. LÜDELING & M. KYTÖ, Eds., *Corpus Linguistics. An International Handbook*, p. 1211–1248. Mouton de Gruyter.

FUTAGI Y., DEANE P., CHODOROW M. & TETREAULT J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning*, **21**, 353–367.

GRANGER S. & BESTGEN Y. (2014). The use of collocations by intermediate vs. advanced non-native writers : A bigram-based study. *IRAL*, **52**, 229–252.

HIGGINS D., RAMINENI C. & ZECHNER K. (2015). Learner corpora and automated scoring. In S. GRANGER, G. GILQUIN & F. MEUNIER, Eds., *Cambridge Handbook of Learner Corpus Research*. Cambridge University Press.

HOWELL D. (2008). *Méthodes statistiques en sciences humaines*. Bruxelles : De Boeck Université.

JARVIS S., BESTGEN Y. & PEPPER S. (2013). Maximizing classification accuracy in native language identification. In *Proceedings of The 8th Workshop on Innovative Use of NLP for Building Educational Applications (NAACL-HLT)*, p. 111–118.

JOACHIMS T. (2006). Training linear SVMs in linear time. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD)*.

KOCH C. (2015). Routines in lexis and grammar : A 'gravity' approach within the International Corpus of English. In *Paper presented at ICAME 36, Trier, 27-31 May 2015*.

NERIMA L., WEHRLI E. & SERETAN V. (2010). A recursive treatment of collocations. In *Proceedings of LREC 2010*, p. 634–638.

PECINA P. (2010). Lexical association measures and collocation extraction. *Language Resources & Evaluation*, **44**, 137–158.

RAMINENI C. & WILLIAMSON D. M. (2013). Automated essay scoring : Psychometric guidelines and practices. *Assessing Writing*, **18**(1), 25–39.

RAMISH C. (2015). *Multiword Expressions Acquisition : A Generic and Open Framework*. Springer.

RENOUF A. & SINCLAIR J. (1991). Collocational frameworks in english. In K. AIJMER & B. ALTENBERG, Eds., *English Corpus Linguistics*, p. 128–143. Longman.

SINCLAIR J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

SMISKOVA H., VERSPOOR M. & LOWIE W. (2012). Conventionalized ways of saying things (CWOSTs) and L2 development. *Dutch Journal of Applied Linguistics*, **1**, 125–142.

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

SOMASUNDARAN S., LEE C. M., CHODOROW M. & WANG X. (2015). Automated scoring of picture-based story narration. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, p. 42–48.

VAN DE CRUYS T. (2011). Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality (DiSCo'2011)*, p. 16–20.

VERSPOOR M., SCHMID M. S. & XU X. (2012). A dynamic usage based perspective on l2 writing. *Journal of Second Language Writing*, **21**, 239–263.

WEIGLE S. C. (2013). English language learners and automated scoring of essays : Critical considerations. *Assessing Writing*, **18**(1), 85–99.

WU J.-C., CHANG Y. C., MITAMURA T. & CHANG J. S. (2010). Automatic collocation suggestion in academic writing. In *Proceedings of the Association for Computational Linguistics Conference*, p. 115–119.

YANNAKOUDAKIS H. & BRISCOE T. (2012). Modeling coherence in ESOL learner texts. In *Proceedings of the 7th Workshop on the Innovative Use of NLP for Building Educational Applications*, p. 33–43.

YANNAKOUDAKIS H., BRISCOE T. & MEDLOCK B. (2011). A new dataset and method for automatically grading ESOL texts. In *The 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies*, p. 180–189.