Revue Française de Linguistique Appliquée, XXIV-1, 81-94, 2019.

Evaluation de textes en anglais langue étrangère et séries phraséologiques : comparaison de deux procédures automatiques librement accessibles

Yves Bestgen, Centre for English Corpus Linguistics, Université catholique de Louvain

Résumé : Lors de l'évaluation automatique de la qualité d'un texte rédigé en langue étrangère, les séries phraséologiques sont fréquemment négligées alors que leur maîtrise est une composante majeure de l'apprentissage. Récemment, deux systèmes automatiques capables de les prendre en compte en anglais ont été mis librement à disposition : le COCA Parser et TAALES. L'étude évalue l'efficacité et l'utilité de ces systèmes en les appliquant à deux ensembles de textes d'apprenants. Si les deux systèmes se sont révélés relativement efficaces, seul le COCA Parser s'est montré capable d'apprendre un modèle prédictif sur un ensemble de données et de l'appliquer avec succès à un autre. C'est également le seul des deux systèmes qui produit un fichier permettant une analyse qualitative des textes évalués.

Title: Evaluation of texts in English as a foreign language and phraseological series: comparison of two freely accessible automatic procedures

Abstract: When automatically evaluating the quality of a text written in a foreign language, phraseological series are frequently neglected while their mastery is a major component of learning. Recently, two automatic systems capable of taking them into account in English have been made freely available: the COCA Parser and TAALES. The study evaluates the effectiveness and utility of these systems by applying them to two sets of learner texts. While both systems proved to be relatively effective, only the COCA Parser was able to learn a predictive model on one dataset and successfully apply it to another. It is also the only one of the two systems that produces a file allowing a qualitative analysis of the evaluated texts.

Mots-clés : apprentissage des langues étrangères, évaluation, phraséologie, traitement automatique du langage (TAL)
Keywords: foreign language learning, evaluation, phraseology, natural language processing (NLP)

1. Introduction

L'évaluation automatique de textes en apprentissage d'une langue étrangère (L2) est un challenge, mais aussi une source d'opportunité tant en éducation qu'en recherche. Dans notre société, rédiger des textes de qualité est une compétence importante (Weigle 2013). L'acquérir demande du temps et, le plus souvent, un apprentissage formel dans un cadre scolaire où l'évaluation, tout particulièrement en L2, est une nécessité comme le souligne Myles (2002, 11) : « L2 writers require and expect specific overt feedback from teachers not only on content, but also on the form and structure of writing. If this feedback is not

part of the instructional process, then students will be disadvantaged in improving both writing and language skills. » La difficulté majeure est qu'évaluer la qualité rédactionnelle d'un texte est compliqué et coûteux en temps (Ramineni & Williamson 2013). En recherche, de telles évaluations sont un prérequis pour pouvoir étudier l'évolution longitudinale de l'apprentissage de la compétence rédactionnelle, les différences entre des populations d'apprenants se distinguant par exemple par la langue maternelle ou encore l'impact d'un dispositif didactique innovant. Dans ce cadre aussi, la lourdeur de l'évaluation pose problème. Les désaccords au moins partiels entre évaluateurs imposent le recours à plusieurs experts. Lorsque des dimensions spécifiques de la qualité rédactionnelle sont au centre de l'étude comme la richesse lexicale, la complexité syntaxique ou la cohérence et l'organisation, un entraînement spécifique des annotateurs est indispensable.

L'évaluation automatique est une réponse au moins partielle à ces difficultés (Ramineni & Williamson 2013). Si cette approche est loin d'être neuve puisque Page a déjà proposé le Project Essay Grade dans les années 1960, ces 25 dernières années ont vu apparaître toute une série de systèmes commerciaux basés sur les progrès de la recherche en traitement automatique du langage (Shermis & Hammer 2013). Ces systèmes extraient des textes à évaluer des indices linguistiques potentiellement corrélés avec la qualité rédactionnelle comme la présence d'erreurs typographiques, orthographiques et syntaxiques, mais aussi la longueur du texte, la présence de phrases très courtes, l'absence d'une introduction ou d'une conclusion. Un matériel d'apprentissage, composé au minimum de plusieurs centaines de textes évalués par des experts humains, est employé pour établir l'équation la plus efficace pour prédire ces évaluations humaines au moyen des indices linguistiques. Cette équation est utilisée pour évaluer de nouveaux textes. Si ces systèmes sont principalement mis en œuvre dans des examens standardisés employés à grande échelle, en compléments d'évaluateurs humains, certains d'entre eux sont aussi utilisés comme outil pédagogique, proposant aux élèves une évaluation globale, mais aussi pointant les erreurs lexicales, grammaticales ou stylistiques présentes dans leurs textes.

Ces systèmes ont toutefois fait l'objet de virulentes critiques (Condon 2013 ; Herrington & Moran 2014 ; Perelman 2014), principalement parce que leur utilité est déterminée sur la base de leur efficacité (les scores attribués sont-ils corrélés avec ceux d'évaluateurs humains?) alors que la validité théorique ou conceptuelle des indices employés n'est pas prise en compte (les indices employés mesurent-ils vraiment la qualité d'un texte?). En d'autres mots, la procédure fonctionne, mais non pour les bonnes raisons parce que les indices obtenus automatiquement ne sont qu'indirectement et que très partiellement liés à la qualité du texte. L'exemple le plus connu est la longueur des textes (Perelman 2014). C'est un indice de qualité particulièrement efficace et donc fréquemment employé par les systèmes automatiques. Il va de soi qu'un apprenant plus avancé peut rédiger des textes plus longs qu'un apprenant moins avancé, surtout si le temps disponible est limité. Toutefois, ceci ne signifie pas qu'un texte plus court est nécessairement moins bon, ni, surtout, que la longueur d'un texte reflète intrinsèquement sa qualité rédactionnelle. D'autres indices employés par ces systèmes présentent le même défaut comme le nombre moyen de syllabes par mot ou la présence de phrases très courtes ou encore l'emploi du

passif qui peut être parfaitement justifié ou non. Comme l'a souligné Somasundaran *& al.* (2015), il est nécessaire que les systèmes d'évaluation automatiques, en plus d'assigner aux textes des scores aussi similaires que possibles à ceux d'évaluateurs humains, s'appuient sur des recherches linguistiques pour se focaliser sur des indices qui « allow for clear interpretation and explanation of scores, which is especially important if the automated scoring is to be employed for educational purposes » (Somasundaran *ibid.*, 42).

Une deuxième critique fréquente est que ces systèmes mettent l'accent sur les erreurs des apprenants et prennent nettement moins en compte leurs succès dans la tâche rédactionnelle (Cheville 2004). Cette prépondérance s'explique par la plus grande efficacité des systèmes automatiques à détecter les erreurs lexicales et grammaticales, mais aussi par l'importance de celles-ci dans l'évaluation humaine. De plus, les erreurs potentielles peuvent être aisément et utilement signalées à l'apprenant par les outils pédagogiques d'aide à la rédaction. Néanmoins, il en résulte une insistance sur l'exactitude formelle des textes au détriment de l'originalité et de la créativité (Vojak *& al.* 2011).

Enfin, ces systèmes commerciaux, qui ont été développés pour répondre à des objectifs souvent très éloignés de ceux des acteurs pédagogiques de terrain, sont coûteux à l'emploi et ne mettent à la disposition des utilisateurs qu'une partie des informations nécessaires pour comprendre leur fonctionnement véritable (Cheville 2004). Leur emploi dans le champ éducationnel ou en recherche s'en trouve donc fortement limité. D'un autre côté, les systèmes développés en recherche académique s'apparentent le plus souvent à une preuve du concept. Ils sont tout aussi rarement utilisables par les praticiens ou d'autres chercheurs.

L'objectif principal de la présente étude est de déterminer si la prise en compte d'un type spécifique d'indices, les séries phraséologiques ou manières conventionnelles de s’exprimer, indices systématiquement négligés jusqu'il y a peu en évaluation automatique de textes, permet de répondre à ces trois critiques. La section suivante résume brièvement les travaux qui ont montré que ces séquences de mots sont des indices linguistiquement justifiés de la qualité rédactionnelle, qu'elles peuvent être mesurées au moyen de procédures automatiques et qu'elles prennent en compte tant les aspects positifs que négatif de la rédaction. La troisième critique, la libre disponibilité de systèmes les implémentant, est abordée dans les deux sections suivantes.

2. Séries phraséologiques et évaluation automatique de textes

L'importance des unités préformées dans l'emploi du langage fait l'objet d'un large consensus tant en linguistique qu'en psycholinguistique (Pawley & Syder 1983 ; Sinclair 1991 ; Wray 2012). La production d'énoncés ne repose pas seulement sur le principe du choix ouvert, mais également, et sans doute même plus, sur le principe idiomatique défini par Sinclair (1991, 110) comme « a language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analysable into segments ». Les unités préformées sont très majoritairement des manières conventionnelles de s’exprimer, conventionnel s'interprétant dans le sens saussurien du terme : une habitude collective et non un contrat (Engler 1980,

11). Il s'agit donc de séquences de mots considérées comme statistiquement typiques de la langue parce qu'elles occurrent « with markedly high frequency, relative to the component words or alternative phrasings of the same expression » (Baldwin & Kim 2010, 270). On peut citer à titre d'exemple : *greater than*, *depend on*, *as much as*, *a lot of* ou encore *on the other hand*. Dans le présent article, ces séquences seront appelées *séries phraséologiques*, le terme proposé par Bally dès 1909 (voir Legallois et Tutin (2013) pour une mise en contexte historique) pour désigner des groupements usuels dans lesquels les « éléments du groupe conservent leur autonomie, tout en laissant voir une affinité évidente qui les rapproche, de sorte que l'ensemble présente des contours arrêtés et donne l'impression du *déjà vu* ». (Bally 1909, 70). La maitrise de ces séries phraséologiques est une composante majeure de l'apprentissage d'une langue étrangère, comme l'avait déjà souligné Bally (1909, 73). Grâce à elle, les apprenants peuvent faire preuve de complexité, d'exactitude et de fluence dans leurs productions, les trois dimensions majeures de l'évaluation du niveau de compétence dans une langue étrangère (Myles 2012). De nombreuses études ont montré que l'emploi de ces séries phraséologiques distinguait les locuteurs natifs des non-natifs et les apprenants avancés des apprenants moins avancés (p.ex. Durrant & Schmitt 2009 ; Forsberg 2010 ; Verspoor *& al.* 2012).

L'analyse de ces séries phraséologiques dans les textes d'apprenants devrait donc être particulièrement utile pour l'évaluation de leur qualité. Cependant, ces séquences sont presque totalement ignorées par les procédures automatiques (Shermis & Hammer 2013). La principale raison de cette négligence provient de la difficulté de développer des mesures appropriées de leur emploi. Dans presque toutes les études en apprentissage des langues étrangères, la nature phraséologique d'une séquence lexicale est déterminée manuellement, à l'aide de dictionnaires ou en demandant à des locuteurs natifs, ce qui est une approche notoirement complexe et coûteuse en temps (Verspoor *ibid.*). Récemment toutefois, des mesures phraséologiques pouvant être automatisées ont été proposées. Ces mesures reposent toutes sur l'approche fréquentielle en phraséologie qui propose de distinguer les séries phraséologiques des combinaisons libres sur la base de leur fréquence dans un grand corpus de référence (Gyllstad 2007). Une autre caractéristique commune à ces mesures est de se baser sur les n-grammes de mots, les séquences contigües d'au moins deux mots, pour identifier ces séries phraséologiques.

Les approches techniquement les plus simples sont basées sur la seule fréquence de ces n-grammes dans un corpus de référence comme le BNC ou le COCA (Kyle & Crossley 2015 ; Garner *& al.* 2019). On procède en extrayant, dans un premier temps, tous les n-grammes différents de la longueur voulue dans le corpus de référence et on en détermine la fréquence. Ensuite, une approche discrète et une approche scalaire sont possibles. Selon la première, seuls les n-grammes dont la fréquence atteint au moins un seuil fixé arbitrairement comme 10 occurrences par million de mots, sont considérés comme des séries phraséologiques. Cette liste est ensuite employée pour identifier dans chaque texte d'apprenant à analyser les n-grammes considérés comme phraséologiques. La proportion de n-grammes ainsi identifiés dans un texte donne son niveau de qualité phraséologique (selon cette mesure bien évidemment). L'approche scalaire n'emploie pas de

seuil de fréquence, mais utilise directement la fréquence moyenne dans le corpus de référence de tous les n-grammes du texte qui sont présents dans celui-ci. Kyle et Crossley (2015) ont observé que de tels indices, tout particulièrement la fréquence moyenne et la proportion de n-grammes présents dans le corpus de référence, étaient de bons prédicteurs de la qualité de textes produits par des apprenants de l'anglais. Ces résultats ont été confirmés par Garner (*ibid.*).

La limitation principale des indices basés sur la seule fréquence des n-grammes dans le corpus de référence est qu'ils ne prennent pas en compte la fréquence des mots qui composent le n-gramme (Benigno *& al.* 2015). Un n-gramme peut être fréquent, non pas en raison de son statut phraséologique, mais parce qu'il est composé de mots très fréquents, qui ont donc une probabilité non négligeable de se suivre dans un texte (voir Bestgen (2018) pour une analyse détaillée). Cette limitation peut être dépassée en employant des indices d'association lexicale qui mesurent la force d'attraction entre les mots qui composent le n-gramme d'une manière indépendante de la fréquence de ceux-ci (Evert 2008). En évaluation de textes, les deux indices les plus souvent employés sont l'information mutuelle (IM) et le score-t (Durrant & Schmitt 2009 ; Granger & Bestgen 2014). Ces deux indices présentent l'intérêt d'être complémentaires, IM privilégiant les n-grammes composés de mots relativement rares comme *volcanic eruption, double-edged sword, wrongful dismissal* alors que le score-t privilégie les séquences composées de mots fréquents comme *you know, as well* ou *able to*. Pour employer ce type d'indices, la première étape consiste à les calculer pour l'ensemble des n-grammes différents présents dans le corpus de référence. Ensuite, une approche discrète et une approche scalaire sont possibles. Elles sont appliquées exactement comme expliqué ci-dessus à la seule différence que l'indice sous-jacent n'est pas la fréquence dans le corpus de référence, mais le score d'association. On peut donc obtenir la proportion de n-grammes dans un texte qui dépasse un score IM donné (Durrant & Schmitt 2009) ou le score IM moyen d'un texte, calculé sur la base de tous les n-grammes de celui-ci qui sont présents dans le corpus de référence (Bestgen & Granger 2014). Ces indices se sont révélés être des prédicteurs particulièrement efficaces de la qualité d'un texte (Bestgen 2016, 2017 ; Bestgen & Granger 2014 ; Garner *ibid.* ; Lenko-Szymanska &Wolk 2016 ; Somasundaran *ibid.*). Bestgen et Granger (2014) ont proposé d'appeler cette approche *CollGram* parce qu'elle combine deux techniques classiques en linguistique computationnelle : l'extraction de n-grammes dans lesquels les mots sont nécessairement contigus, mais aucun indice d'association lexicale n'est employé, et l'identification de collocations, qui s'appuie sur des indices d'association, mais les mots qui les composent ne sont pas nécessairement contigus.

Ces mesures de l'emploi des séries phraséologiques par des apprenants d'une langue étrangère sont immunes aux deux principales critiques formulées à l'encontre des systèmes d'évaluation automatique de textes. Non seulement ces mesures sont justifiées linguistiquement et pédagogiquement, mais elles permettent de mettre en évidence tant des emplois avancés (de bonnes séries phraséologiques) que des erreurs (Granger & Bestgen 2017). En ce qui concerne la troisième critique, la nécessité de systèmes librement disponibles pour l'éducation et la recherche, la situation a évolué tout récemment d'une

manière très favorable puisque deux systèmes les implémentant sont depuis peu librement disponibles pour l'anglais. Tant les chercheurs que les praticiens peuvent donc calculer ces indices sur les textes qui les intéressent malgré leur grande complexité (besoin d'un corpus de référence de grande taille, prétraitement des textes à analyser, etc.). La question du choix entre ces deux systèmes se pose toutefois. Or, le peu de documentation disponible sur leur fonctionnement ainsi que d'études ayant démontré l'efficacité de ces systèmes spécifiques rend un tel choix difficile d'autant plus que les indices calculés et les résultats fournis à l'utilisateur sont différents. La suite de cet article a pour objectif de permettre un choix informé en présentant tant les avantages que les limitations de chaque système, mais aussi et surtout les bénéfices qu'ils peuvent apporter tant en recherche appliquée que dans le champ éducationnel. La section suivante présente brièvement ces deux systèmes. Ensuite, la section empirique rapporte une série d'analyses menées afin de comparer leur efficacité et leur utilité lorsqu'ils sont appliqués à deux collections de textes d'apprenants de l'anglais.

3. Les deux systèmes librement accessibles

3.1. COCA Parser

Le COCA Parser est un site internet (http://collgram.pja.edu.pl, Lenko-Szymanska &Wolk 2016 ; Wolk *& al.* 2017), qui calcule pour les n-grammes de 2 à 4 mots présents dans un texte les scores IM et t moyens sur la base de COCA comme corpus de référence. L'analyse d'un texte ou d'un ensemble de textes est très facile à effectuer puisqu'il suffit de sélectionner les fichiers à analyser dans un dossier de son propre ordinateur. Il est important de choisir l'option *Use CLAWS tagger* qui permet un meilleur prétraitement des textes. Le traitement d'un ensemble de textes prend un temps variable, pouvant aller jusqu'à trois quarts d'heure pour plusieurs centaines de textes.

Le site renvoie des documents Excel compressés, dont une synthèse contenant les scores globaux pour chaque texte soumis. Il retourne aussi un fichier par texte analysé qui donne un ensemble d'informations pour chaque n-gramme du texte dont les plus importantes sont la fréquence du n-gramme dans le texte, sa fréquence dans le COCA, les scores IM et t de ce n-gramme ainsi que ses scores Dice et Gravity, deux autres mesures d'association (bien que celles-ci ne soient pas repris dans le fichier général de synthèse).

La limitation majeure du site réside dans l'absence de documentation fournie à l'utilisateur. Si l'approche CollGram qu'il implémente est décrite dans la littérature, aucune information n'est donnée sur les mesures supplémentaires calculées comme le score Gravity. De plus, le fichier de synthèse ne contient que les deux mesures d'association préconisées par CollGram et seules celles fournies dans les fichiers par texte peuvent être analysées en plus, moyennant l'obtention par l'utilisateur des scores moyens. Un autre problème potentiel du COCA Parser est que les textes soumis sont envoyés par internet à un serveur, ce qui peut poser des problèmes de confidentialité.

3.2. TAALES

TAALES 2.0 est un logiciel téléchargeable (www.kristopherkyle.com/taales.html), développé par Kyle et Crossley (Kyle & Crossley 2015 ; Kyle *& al.* 2018) et disponibles

pour Linux, Windows et Mac OS. TAALES est l'acronyme de *Tool for the Automatic Analysis of Lexical Sophistication*, mais l'outil est bien plus polyvalent puisqu'il permet le calcul de plus de 400 indices lexicaux obtenus en comparant les mots présents dans un texte à ceux présents dans des listes très diversifiées comme la fréquence dans des corpus de référence, les mots académiques de Coxhead, des normes psycholinguistiques couvrant le degré de concrétude ou la familiarité, le nombre de voisins orthographiques... TAALES facilite aussi le calcul d'indices de la compétence phraséologique en comparant les bigrammes et trigrammes de mots présents dans un texte à ceux qui sont présents dans les différentes sections du COCA telles que les sections académiques et orales.

Le programme est accessible via une interface graphique, ce qui le rend très facile d'emploi. Il traite les fichiers au format *txt* et produit en sortie deux fichiers de synthèse au format *csv* (valeurs séparées par des virgules). Le premier donne les valeurs des indices pour chaque texte et le second, le nombre de mots ou de n-grammes de chaque texte qui ont été employés pour calculer ces indices.

TAALES est accompagné d'un mode d'emploi succinct et d'une description rudimentaire de l'ensemble des indices calculés. Fonctionnant sur l'ordinateur de l'utilisateur et ne nécessitant pas d'accès réseau, il ne présente pas de problème de confidentialité. Par contre, il nécessite une version spécifique de Python et ne fonctionne donc pas nécessairement sur tous les ordinateurs. Il ne produit pas en sortie de fichiers reprenant les mots et n-grammes spécifiques qui ont été identifiés dans chaque texte analysé avec leurs valeurs dans les listes. Il est donc impossible de déterminer par exemple quels sont les n-grammes les plus sophistiqués d'un texte donné, ni de s'assurer de ce que les textes ont été correctement traités. Ceci est d'autant plus regrettable que très peu d'information à propos des prétraitements des textes tels que la segmentation en mots est disponible.

4. Etude empirique

4.1. Méthode

4.1.1. Matériel

Les analyses ont été effectuées sur deux collections de textes d'apprenants de l'anglais afin d'évaluer l'efficacité des deux systèmes pour des textes se différenciant par le genre, le thème et les conditions de production. ICLE, le premier ensemble de données, est composé de 223 essais (approximativement 150 000 mots au total) extraits de l'*International Corpus of Learner English* (Granger *& al.* 2002). Les essais analysés, sélectionnés dans le cadre d'une thèse de doctorat (Thewissen 2013), étaient tous de nature argumentative et composés de 500 et 900 mots. Les apprenants étaient des étudiants de premier cycle, principalement âgés de 20 à 25 ans. Ces essais ont été évalués par deux évaluateurs professionnels, qui leur ont attribué une note basée sur les descripteurs du Cadre européen commun de référence pour les langues (CECR). Les évaluateurs ont été invités à juger indépendamment chaque essai comme étant à l'un des niveaux de compétence B1, B2, C1 ou C2 (c'est-à-dire les niveaux intermédiaire et avancé du CECR). Ils pouvaient utiliser les signes + et - pour distinguer les sous-niveaux. Ces catégories ont été recodées sous forme d'échelle allant de 1

(pour B1-) à 11 (pour C2). Le coefficient de corrélation de Pearson entre les notes attribuées par les deux évaluateurs sur cette échelle est de 0,69. Les 34 textes sur lesquels les évaluateurs étaient en désaccord de plus d'un niveau CECR ont été soumis à un troisième évaluateur. Le score final d'un texte est la moyenne des scores disponibles.

FCE, le second ensemble de données, est composé de textes rédigés dans le cadre du *First Certificate in English* (Yannakoudakis *& al.* 2011). Ce test a été conçu pour évaluer les apprenants de l'anglais au niveau B2 du CECR. Pour les 1 237 apprenants, on dispose de deux brefs textes, chacun de 100 à 200 mots, pour un total de plus ou moins 460 000 mots. Les participants disposaient de 80 minutes pour rédiger leurs deux textes. La majorité des apprenants avaient entre 16 et 25 ans. Dans le cadre du premier certificat en anglais, une note globale a été attribuée à la paire de textes rédigés par chaque apprenant sur une échelle allant de zéro à 40. Les critères de notation étaient principalement basés sur le succès dans la tâche de communication à accomplir, avec un accent particulier sur l'organisation et la cohésion du texte ainsi que sur l'exactitude et l'étendue du langage. Yannakoudakis (*ibid.*) a estimé la fiabilité des évaluations des textes d'une centaine d'apprenants en les comparant aux évaluations de quatre experts et a obtenu une corrélation moyenne de Pearson de 0,80.

4.1.2. Indices employés et techniques statistiques

Les 50 indices phraséologiques recommandés par Garner (*ibid.*) pour TAALES ont été utilisés. Il s'agit d'indices basés tant sur la fréquence (moyenne, dispersion et proportion, après ou non une transformation logarithmique), que sur des mesures d'association dont IM et le score-t, mais aussi DeltaP. Ces indices sont calculés sur la base des bigrammes et des trigrammes présents dans deux corpus de référence : les sections orale et académique du COCA.

Pour le COCA Parser, les 16 indices disponibles pour les bigrammes et les trigrammes de mots ont été utilisés. Il s'agit des indices d'association employés dans l'approche CollGram et de deux indices de fréquence. Ces indices sont calculés en prenant en compte toutes les occurrences des n-grammes dans un texte (*tokens*), et donc potentiellement plusieurs fois le même, ou seulement les n-grammes différents (*types*).

Les analyses ont été effectuées au moyen de régressions linéaires multiples dont la fonction est de prédire le plus efficacement possible les scores réels des textes en sélectionnant les meilleurs prédicteurs parmi l'ensemble des indices phraséologiques calculés par chacun des deux systèmes. Les meilleurs prédicteurs ont été sélectionnés par une procédure pas-à-pas qui, lors de chaque étape, ajoute au modèle en cours de construction l'indice ayant le pouvoir prédictif le plus élevé pour autant que cet indice soit statistiquement significatif au seuil de 0,05. Comme l'ajout d'un prédicteur au modèle peut modifier l'utilité des prédicteurs déjà présents, la procédure vérifie après chaque ajout que le prédicteur du modèle qui présente l'utilité la plus faible contribue toujours de manière significative à la prédiction. Sinon, il est retiré du modèle.

Ces analyses ont d'abord été réalisées sur toutes les données disponibles afin d'obtenir le point de vue le plus complet possible sur l'efficacité des deux systèmes. Pour

réduire le risque d'une estimation favorablement biaisée de cette efficacité, les systèmes ont également été évalués à l'aide d'une procédure de validation interne, dite croisée. Pour ce faire, chaque ensemble de données a été divisé de manière aléatoire en trois parties de taille égale et, de manière répétée, deux de ces parties ont été employées pour construire le modèle permettant de prédire les scores des observations de la troisième partie. La mesure d'efficacité finale est la corrélation entre l'ensemble des scores prédits et les scores réels. Il est néanmoins important de noter que, si une validation interne permet d'évaluer la possibilité de généraliser les résultats à un autre échantillon extrait de la même population, elle ne garantit en rien que les résultats puissent être généralisés à une autre population. Tout particulièrement, elle ne nous informe pas à propos de l'efficacité du système dans les cas où on ne dispose pas, pour le nouveau matériel à traiter, d'un échantillon qui a été évalué par des juges humains de sorte qu'on puisse calibrer le système. Il s'agit pourtant d'un scénario important lorsque l'objectif est de développer un système d'évaluation automatique de textes qui puisse être employé dans un contexte de recherche appliquée et dans le domaine de l'enseignement. Apprendre le modèle prédictif sur un ensemble de données et le tester sur un tout autre ensemble de données est le seul moyen d'évaluer véritablement les capacités de généralisation du système. Afin d'obtenir cette information, on a procédé à une validation externe en appliquant le modèle prédictif construit par chaque système pour un ensemble de données à l'autre ensemble de données.

4.2. Résultats

4.2.1 Analyses quantitatives

Le tableau 1 présente les corrélations entre les scores prédits par les régressions multiples et les scores des évaluateurs humains pour les deux systèmes et les deux corpus. Toutes ces corrélations sont statistiquement significatives pour un seuil de 0,0001 et la majorité d'entre elles sont élevées puisqu'elles sont au moins égales à 0,50, valeur correspondant à une taille d'effet importante selon les critères de Cohen (1988). Comme attendu, on observe des performances plus faibles en validation interne et encore plus faibles en validation externe. Toutefois, les performances, dans ce dernier cas, du COCA Parser restent parfaitement acceptables alors qu'elles sont nettement plus problématiques pour TAALES.

Tableau 1 : Corrélations entre les valeurs prédites et les évaluations humaines

		COCA Parser r	TAALES r	Z
ICLE	Tout	0,67	0,63	1,00
	Validation interne	0,64	0,50	3,03*
	Validation externe : FCE	0,49	0,15	12,57**
FCE	Tout	0,62	0,48	8,05**
	Validation interne	0,60	0,44	8,80**
	Validation externe : ICLE	0,50	0,29	4,66**

Notes : * 0,01, ** 0,0001.

Les corrélations obtenues par les deux systèmes pour les mêmes conditions (donc, sur une même ligne du tableau 1), ont été comparées au moyen du test Z proposé par Steiger (1980) pour des paires de corrélations ayant une variable commune. Les corrélations obtenues par le COCA Parser sont toujours supérieures à celles obtenues par

TAALES et la différence est statistiquement significative dans cinq cas sur six[1]. La seule exception est l'analyse basée sur l'échantillon complet pour ICLE. Il s'agit toutefois de l'ensemble de données dans lequel les différences entre l'estimation de performance optimiste, obtenue sur toutes les données, et celles plus crédibles, obtenues en validation, sont les plus fortes. On peut donc penser que c'est la taille modeste de l'échantillon (N = 223) en comparaison du grand nombre de prédicteurs disponibles dans TAALES (50) qui permet à ce système d'être plus performant, mais uniquement sur les données employées pour l'apprentissage. Le tableau 1 montre aussi que les performances du COCA Parser sont très similaires pour les deux ensembles de données alors que des différences beaucoup plus nettes sont observées pour TAALES.

Tableau 2 : prédicteurs sélectionnés par les régressions multiples

COCA Parser		
Spécifiques à ICLE	Communs	Spécifiques à FCE
	IM_types_2g IM_tokens_2g freq_types_2g freq_tokens_2g	IM_types_3g t_types_3g
TAALES		
Spécifiques à ICLE	Communs	Spécifiques à FCE
oral_3g_IM acad_3g_2_T acad_3g_2_IM acad_3g_prop oral_3g_DP acad_2g_disp_log	acad_2g_prop oral_3g_DP	oral_2g_IM oral_3g_freq oral_3g_2_DP oral_2g_freq_log oral_3g_prop oral_2g_freq_log oral_2g_DP oral_3g_2_IM oral_2g_disp oral_2g_T

Notes : acad = académique, log = logarithme, disp = dispersion, freq = fréquence, prop = proportion. Le *2* qui suit le *3g* dans certains noms de prédicteurs issus de TAALES indique qu'il s'agit d'une version spécifique d'un prédicteur.

Les prédicteurs les plus utiles pour la prédiction, c'est-à-dire ceux qui sont sélectionnés par la régression multiple lors de l'analyse de l'échantillon complet, sont donnés dans le tableau 2. Pour le COCA Parser, deux tiers des prédicteurs sont basés sur des bigrammes. Parmi les indices sélectionnés, on trouve des indices qui ne sont pas présents dans l'approche CollGram habituelle, à savoir les fréquences moyennes. On observe aussi que les indices calculés sur les *types* et sur les *tokens* apportent des informations au moins partiellement complémentaires puisqu'ils sont sélectionnés par la procédure de régression pas à pas. Les indices sélectionnés dans TAALES sont nettement plus diversifiés. On y trouve trois mesures d'association différentes, des indices de fréquence, de proportion et de dispersion. Ces indices sont majoritairement basés sur les trigrammes. Lorsqu'on compare les modèles prédictifs pour les deux ensembles de données, on observe qu'il y a deux fois plus de prédicteurs communs pour le COCA Parser que pour TAALES, ce qui explique pourquoi le COCA Parser est nettement plus efficace que

[1] Les tests sont nettement plus significatifs pour FCE par rapport à ICLE parce que l'ensemble de données FCE est beaucoup plus grand et qu'il existe une relation bien établie en statistique entre la taille de l'échantillon et le degré de significativité statistique.

TAALES lorsqu'il s'agit de prédire les scores d'un ensemble de données à partir du modèle construit sur l'autre. En ce qui concerne TAALES, on observe que tous les prédicteurs pour FCE sont issus de la section orale de COCA alors qu'ils sont majoritairement issus de la section académique pour ICLE.

4.2.2 Analyses qualitatives

Les fichiers produits par le COCA Parser pour chaque texte permettent de nombreuses analyses qualitatives qui sont impossibles avec la version actuelle de TAALES. Ils peuvent évidemment être employés pour identifier les n-grammes les plus extrêmes présents dans les textes analysés. On note ainsi que parmi les bigrammes les plus fortement associés pour IM dans FCE on trouve des noms propres comme *hercule poirot* et *sagrada familia*, dont on peut se demander s'ils ne devraient pas être éliminés de l'analyse (Granger & Bestgen 2014). On y trouve aussi des bigrammes comme c*arbon dioxide*, *soap operas* et f*irst-prize winner* qui ont clairement leur place dans ce genre de listes. Parmi les trigrammes ayant le score IM le plus élevé, on trouve *near piccadilly circus* et *alexander graham bell*, mais aussi *stained glass windows*, *favorite soap operas*, *pixel by pixel* et *famous tourist attractions*. Parmi les trigrammes ayant le score IM le plus faible, des trigrammes qui sont donc particulièrement inattendus dans des textes rédigés par des natifs, on trouve six *because is the*, deux *because is a* et un *because was the,* produits par dix apprenants différents.

Un autre usage évident de ce fichier est de permettre de vérifier concrètement comment chaque texte a été analysé par le système et d'effectuer, si nécessaire, des ajustements. A titre d'exemple, dans un des textes de ICLE, un bigramme s'est vu attribuer un score-t de -85 475, ce qui indique qu'il est observé dans le corpus de référence extrêmement moins souvent que ce qu'il devrait l'être étant donné la fréquence des deux mots qui le composent. Il s'agit de la séquence *s he* qui est, de fait, hautement improbable en anglais. Elle résulte, dans le texte de l'apprenant, d'une mauvaise segmentation par le COCA Parser de la forme graphique *s/he*. Il pourrait donc être judicieux de transformer cette forme avant de resoumettre le texte.

Disposer de ces listes permet aussi de modifier les indices calculés par le COCA Parser. Il est ainsi possible de ne prendre en compte lors du calcul des IM et score-t moyens que les bigrammes qui ont été observés au moins cinq fois dans le corpus de référence, comme l'ont recommandé Durrant et Schmitt (2009). C'est aussi grâce à ces listes qu'il a été possible de se rendre compte que la version actuelle du COCA Parser calcule les scores IM et t moyens en utilisant comme dénominateur non le nombre de scores connus, comme cela est effectué habituellement dans CollGram, mais le nombre total de mots (t*okens*).

Ces listes permettent également d'aller au-delà des scores moyens fournis par le COCA Parser pour obtenir, pour chaque texte, le pourcentage de paires de mots qui sont très fortement, fortement ou faiblement associés. Il est alors possible d'identifier des textes qui présentent des profils extrêmes, par exemple beaucoup trop de *mauvais* n-grammes et beaucoup trop peu de *bons* ou le profil inverse, et surtout d'identifier les n-grammes en question. Un exemple de ce deuxième type de profil est donné dans le tableau 3. Il s'agit

d'un extrait du texte de ICLE, évalué par les experts comme étant du niveau C2, qui a la proportion la plus élevée de paires de mots ayant un score IM supérieur à 6, un des seuils employés par Durrant et Schmitt (2009). L'extrait présenté est le suivant : *Forget the image of children created by TV ads: cherubic faces, blond curly hair, ocean blue eyes and smiles that makes you melt.* Nombre de bigrammes y témoignent d'un haut degré de maîtrise de la langue, même si on y trouve une erreur grammaticale (*makes*).

Tableau 3 : Bigrammes et scores IM

Bigramme	IM	Bigramme	IM	Bigramme	IM
forget the	1,19	tv ads	8,42	eyes and	1,55
the image	2,16	cherubic faces	8,93	and smiles	1,82
image of	3,33	blond curly	8,41	smiles that	-1,20
of children	0,98	curly hair	10,50	that makes	3,15
children created	-1,51	ocean blue	3,88	makes you	2,75
created by	5,17	blue eyes	7,51	you melt	-1,05
by tv	-0,52				

La même analyse peut être effectuée sur les n-grammes qui sont absents du corpus de référence afin de déterminer leurs propriétés. Les deux extraits donnés dans la figure 1 proviennent des deux textes de ICLE qui contiennent le plus grand pourcentage de ces n-grammes absents dans COCA. Comme on peut le voir, les bigrammes absents, qui sont soulignés ou surlignés dans la figure, ont une origine très différente et ceci mène à des conclusions opposées quant au niveau de compétence des deux apprenants tel qu'il peut être estimé sur la base des textes. Le premier extrait inclut toute une série de bigrammes contenant une erreur et qui, pour cette raison, ne se trouvent pas dans le corpus de référence ; ce texte a d'ailleurs été évalué comme étant du niveau B1-. Les bigrammes du deuxième extrait (texte du niveau C2), qui sont absents du corpus de référence, sont complètement différents : il ne s'agit pas d'erreurs, mais d'énoncés créatifs s'inscrivant dans des séries phraséologiques plus longues.

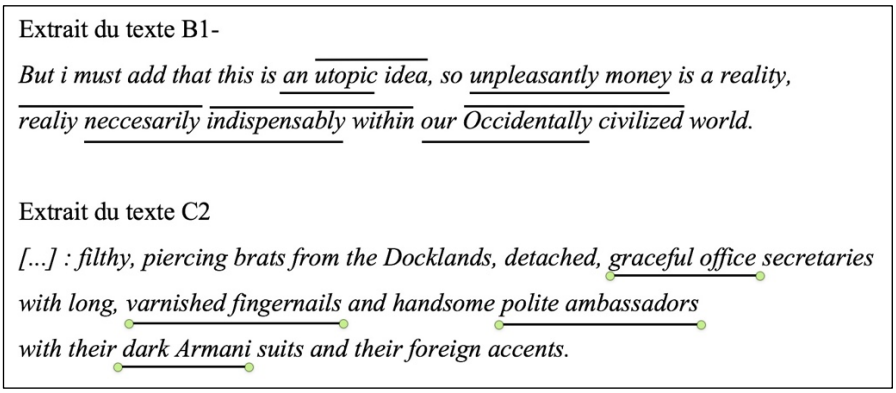

Figure 1 : Extraits des deux textes contenant le plus grand nombre de bigrammes absents

5. Discussion et conclusion

Les analyses présentées ci-dessus plaident en faveur de la prise en compte des séries phraséologiques, les manières conventionnelles de s'exprimer dans une langue, lors de l'évaluation automatique de textes d'apprenants d'une langue étrangère. Non seulement les indices dérivés de ces séquences sont d'excellents prédicteurs de la qualité rédactionnelle

d'un texte, mais l'analyse plus qualitative des séquences de mots ainsi identifiées apporte plusieurs points de vue complémentaires sur les succès comme les échecs rédactionnels d'un apprenant. Puisque nombre d'auteurs ont souligné l'importance de la composante phraséologique dans la maîtrise d'une langue étrangère (voir la section 2), ces indices semblent répondre aux deux critiques majeures formulées à l'encontre des systèmes automatiques d'évaluation de textes. Grâce au développement récent de deux systèmes librement disponibles implémentant ces techniques, les chercheurs en linguistique et en didactique des langues étrangères ont maintenant accès à ceux-ci, pouvant les évaluer, suggérer des améliorations ou attirer l'attention sur des problèmes. A terme, cette approche de la qualité rédactionnelle d'un texte pourra entrer en classe et contribuer à l'évaluation de textes par les enseignants. Il faut toutefois souligner que, si les possibilités d'emploi semblent nombreuses et potentiellement fructueuses, les deux systèmes actuellement disponibles présentent des limitations qui risquent de fortement freiner une telle adoption. Tout particulièrement, TAALES ne permet d'obtenir les informations qualitatives nécessaires et les sorties, suffisamment riches du COCA Parser, nécessitent des traitements complémentaires pour en tirer bénéfice. Il y a donc encore du travail à effectuer pour rendre cette approche pleinement disponible, d'autant plus que les deux systèmes ne peuvent analyser qu'un très petit nombre de langues différentes, l'anglais seul pour TAALES et l'anglais et le polonais pour le COCA Parser. Il faut aussi souligner que cette approche pour estimer le degré de maîtrise phraséologique d'un apprenant est encore en cours de développement, même si elle semble déjà efficace. Tout particulièrement, malgré que les deux systèmes disponibles aillent au-delà de CollGram en prenant en compte d'autres indices d'association lexicale et en considérant non seulement les bigrammes, mais aussi les séquences plus longues, ils présentent le défaut de ne prendre en compte que les séquences de mots contigus alors que les expressions conventionnelles ne sont qu'exceptionnellement entièrement figées (Legallois & Tutin 2013).

Pour conclure, il est utile de revenir sur les résultats de la comparaison des deux systèmes afin d'en approfondir l'interprétation, mais aussi de la nuancer. Tant les analyses quantitatives que qualitatives suggèrent que le COCA Parser est plus efficace et plus utile que TAALES. Il faut toutefois garder en mémoire que TAALES est d'abord un outil pour l'analyse de la sophistication du vocabulaire d'un texte et qu'il inclut un très grand nombre de mesures lexicales basées sur les mots isolés. TAALES permet donc beaucoup plus d'analyses que le COCA Parser. S'il ne semble pas évident que nombre des mesures disponibles soient particulièrement utiles pour la didactique des langues, leur utilité en linguistique appliquée semble indéniable. La présente étude suggère seulement que le COCA Parser est un meilleur choix pour l'analyse des séries phraséologiques. Un des résultats obtenus mérite dans ce cadre une analyse approfondie. On peut s'étonner que les indices qui sont calculés par les deux instruments, et tout particulièrement, ceux issus de CollGram, soient nettement moins efficaces lorsqu'ils sont obtenus par TAALES que par le COCA Parser. Il faut toutefois se souvenir que les deux systèmes n'emploient pas exactement le même corpus de référence. On peut aussi penser que les prétraitements des textes sont différents, mais cette hypothèse, vu le peu d'information disponible à ce sujet,

est difficile à vérifier. En ce qui concerne TAALES, une analyse approfondie du logiciel montre qu'il ne prend en compte que les bigrammes qui occurrent au moins 51 fois dans le corpus de référence, une information non disponible dans la documentation et qui peut certainement expliquer ses performances moindres puisqu'il néglige un grand nombre de n-grammes présents dans les textes des apprenants. Comme on l'a vu plus haut, les indices calculés par le COCA Parser sont basés sur un dénominateur discutable, le nombre total de bigrammes et non le nombre de bigrammes qui sont réellement intervenus dans le calcul de l'indice en question. Ce point n'est pas signalé sur le site. Mettre à disposition une documentation détaillée semble donc la première amélioration indispensable et ce pour les deux systèmes. Plus généralement, les analyses présentées suggèrent une série d'améliorations possibles. La suite logique, et déjà en cours, du présent travail est le développement d'un système, tout aussi librement disponible, qui les implémente.

Remerciements. L'auteur est chercheur qualifié du F.R.S-FNRS de la fédération Wallonie-Bruxelles.

Yves Bestgen
CECL, Université catholique de Louvain
Place Cardinal Mercier, 10 1348 Louvain-la-Neuve Belgique
yves.bestgen@uclouvain.be, Tél. : +3210473005

Références

Baldwin, T. & Kim, S. N. (2010). Multiword Expressions. In Indurkhya, N. & Damerau, F. J. (eds), *Handbook of Natural Language Processing*, Boca Raton, CRC Press, 267-292.

Bally, C. (1909). *Traité de stylistique française*. Paris, Klincksieck.

Benigno, V., Grossmann, F. & Kraif, O. (2015). Les collocations fondamentales : une piste pour l'apprentissage lexical. *Revue française de linguistique appliquée*, XX, 81-96.

Bestgen, Y. (2016). Evaluation automatique de textes. Validation interne et externe d'indices phraséologiques pour l'évaluation automatique de textes rédigés en anglais langue étrangère. *Traitement automatique des langues*, 57(3), 91-115.

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System*, 69, 65-78.

Bestgen, Y. (2018). Evaluating the frequency threshold for selecting lexical bundles by means of an extension of the Fisher's exact test. *Corpora*, 13(2), 205-228.

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.

Cheville, J. (2004). Automated scoring technologies and the rising influence of error. *The English Journal*, 93(4), 47-52.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Mahwah, Lawrence Erlbaum.

Condon, W. (2013). Large-scale assessment, locally-developed measures, and automated scoring of essays: Fishing for red herrings? *Assessing Writing*, 18(1), 100-108.

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

Engler, R. (1980). Sémiologies saussuriennes II. *Cahiers Ferdinand de Saussure*, 34, 1-16.

Evert, S. (2008). Corpora and collocations. In Lüdeling, A. & Kytö, M. (eds), *Corpus Linguistics. An International Handbook*, Berlin, De Gruyter Mouton, 1211-1248.

Forsberg, F. (2010). Using conventional sequences in L2 French. *International Review of Applied Linguistics in Language Teaching*, 48, 25-50.

Garner, J., Crossley, S. & Kyle, K. (2019). N-gram measures and L2 writing proficiency. *System*, 80, 176-187.

Granger, S. & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers : a bigram-based study. *International Review of Applied Linguistics in Language Teaching*, 52, 229-252.

Granger, S. & Bestgen, Y. (2017). Using collgrams to assess L2 phraseological development: A replication study. In de Haan, P., de Vries, R. & van Vuuren, S. (eds.), *Language, Learners and Levels: Progression and Variation*. Louvain-la-Neuve, Presses Universitaires de Louvain, 387-410.

Granger, S., Dagneaux, E. & Meunier, F. (2002). *International corpus of learner English*. Louvain-la-Neuve, Presses universitaires de Louvain.

Gyllstad, H. (2007). Testing English collocations: developing tests for use with advanced Swedish learners. Unpublished doctoral dissertation. Lund University.

Herrington, A. & Moran, C. (2014). Writing to a machine is not writing at all. In Elliot, N. & Perelman, L. (eds), *Writing Assessment in the 21st Century: Essays in Honor of Edward M. White*, New-York, Hampton Press, 425-438.

Kyle, K. & Crossley, S. A. (2015). Automatically assessing lexical sophistication: indices, tools, findings, and application. *TESOL Quarterly*, 49, 757-786.

Kyle, K., Crossley, S. & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): version 2.0. *Behavioral Research*, 50, 1030-1046.

Legallois, D. & Tutin, A. (2013). Présentation : vers une extension du domaine de la phraséologie. *Langages*, 189, 3-25.

Lenko-Szymanska, A. & Wolk, A. (2016). 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.

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

Myles, J. (2002). Second language writing and research: the writing process and error analysis in student texts. *Teaching English as a Second or Foreign Language*, 6, 1-16.

Pawley, A. & Syder, F. H. (1983). Two puzzles for linguistic theory : nativelike selection and nativelike fluency. In Richards, J. C. & Schmidt, R. W. (eds), *Language and Communication*, London, Longman, 191-226.

Perelman, L. C. (2014). When “the state of the art” is counting words. *Assessing Writing*, 21, 104-111.

Ramineni, C. & Williamson, D. M. (2013). Automated essay scoring : psychometric guidelines and practices. *Assessing Writing*, 18(1), 25-39.

Shermis, M. D. & Hammer, B. (2013). Contrasting state-of-the-art automated scoring of essays. In Shermis, M. D. & Burstein, J. (eds.), *Handbook of Automated Essay Evaluation*, New York, Routledge, 313-346.

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

Somasundaran, S., Lee, C. M., Chodorow M. *& al.* (2015). Automated scoring of picture-based story narration. *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, 42-48.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87, 245-251.

Thewissen, J. (2013). Capturing L2 accuracy developmental patterns: insights from an error-tagged EFL learner corpus. *Modern Language Journal*, 97, 77-101.

Verspoor, M., Schmid, M. S. & Xu, X. (2012). A dynamic usage based perspective on L2 writing. *Journal of Second Language Writing*, 21, 239-263.

Vojak, C., Kline, S., Cope, B. *& al.* (2011). New spaces and old places: an analysis of writing assessment software. *Computers and Composition*, 28, 97-111.

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

Wołk, K., Wołk, A. & Marasek, K. (2017). Unsupervised tool for quantification of progress in L2 English phraseological. *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*, 383–388.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual Review of Applied Linguistics*, 32, 231-254.

Yannakoudakis, H., Briscoe, T. & Medlock, B. (2011). A new dataset and method for automatically grading ESOL texts. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 180-189.