Classification de textes anglais L2 par niveau de compétence langagière

Yannis Haralambous

Classification de textes anglais L2 par niveau de compétence langagière.

Autour d'une compétition d'apprentissage automatique en traitement automatique des langues

Yannis HARALAMBOUS

IMT Atlantique/ LabSTICC
Technopôle Brest Iroise, CS 83818
29238 Brest Cedex 3

yannis.haralambous@imt-atlantique.fr

RÉSUMÉ. Dans cet article, nous décrivons et discutons notre participation à une compétition d'apprentissage automatique dont l'objectif a été la classification de textes anglais produits par des apprenants, par niveau de compétence langagière.

MOTS-CLÉS : traitement automatique de la langue, apprentissage automatique, compétence langagière.

ABSTRACT. In this article, we describe and discuss our participation to a machine learning challenge whose goal was the prediction of linguistic competency of learners of English as a second language.

KEYWORDS: natural language processing, machine learning, language competency.

Introduction

À l'origine de cette communication se trouve une compétition dans le domaine de l'apprentissage automatique (*machine learning*) qui s'est déroulée entre le 28 mars et le 28 mai 2018. Nous allons décrire le corpus proposé et la classification demandée, les différentes tentatives de l'équipe de l'auteur et la solution donnée par le vainqueur de la compétition. Considérant cette compétition comme typique des calculs effectués en apprentissage automatique dans le domaine du traitement automatique de la langue, nous allons, en guise de conclusion, commenter ses enjeux et l'utilité des résultats obtenus.

1. La compétition

1.1. La conférence

La conférence CAp (Conférence sur l'Apprentissage automatique) est, comme le montre le site, le « rendez-vous annuel de la communauté francophone pour la présentation des résultats de recherche originaux, ainsi que l'échange et la diffusion d'expériences novatrices dans le domaine de l'apprentissage automatique ». L'édition 2018 de la conférence s'est déroulée à Rouen, sur 2 jours et demi (3 conférenciers invités, 23 exposés, deux sessions de poster avec 23 posters en tout).

Depuis 2017, la conférence CAp propose un concours. En 2017, il s'agissait d'identifier des entités nommées dans un corpus de 3 000 tweets français (l'apprentissage se faisant à partir d'un corpus de 3 000 tweets annotés). En 2018, le sujet proposé a été la classification des textes anglais L2 par niveau de compétence langagière, et c'est dans le cadre de cette compétition, intitulée *My Tailor is rich !*, que nous allons nous intéresser à cette communication. Notons qu'en 2019 il n'y a pas eu de concours proposé dans le cadre de la conférence CAp.

1.2. Le corpus

L'université de Cambridge — en collaboration avec la société privée *EF Education First*, société spécialisée dans les séjours linguistiques, fondée en 1965 et basée en Suisse — propose une interface en ligne payante pour apprendre l'anglais, du nom de *Englishtown* ou *English Live*. Les apprenants y saisissent des textes sur des thèmes prédéfinis : il y a 16 niveaux de compétence et dans chaque niveau l'apprenant doit écrire 8 textes. Ces textes, qui « offrent une variété de tâches réceptives et productives » (Huang *et al*., 2017) sont corrigés et commentés par des enseignants.

En 2013, le Département de linguistique théorique et appliquée de l'université de Cambridge décide de mettre les productions des apprenants d'*English Live* à dispo-

sition du public. Le corpus ainsi obtenu, appelé EFCAMDAT (*Education First Cambridge Open Language Database*), contient aujourd'hui 1 180 310 textes écrits par 174 743 apprenants de 198 nationalités différentes (dont 40,4 % sont des Brésiliens, 14 % des Chinois et 3,5 % des Français, la France étant la huitième nationalité par ordre d'effectif d'inscrits). En s'inscrivant (gratuitement) au site d'EFCAMDAT, on a accès à tous les textes, classés par niveau de compétence, par thème et par nationalité. Notons que les données personnelles des apprenants n'y figurent pas, mais que l'on a accès à leurs identifiants pour pouvoir suivre leur évolution. Pour chaque texte on dispose aussi de sa date de rédaction, de la note obtenue et des corrections effectuées par l'enseignant. Somme toute, il s'agit d'un corpus très riche, qui peut s'avérer très utile pour les personnes qui s'intéressent à l'apprentissage de l'anglais.

Les seize niveaux de compétence d'EFCAMDAT correspondent *grosso modo* aux six niveaux européens de compétence langagière, de la manière suivante : 1–3 (A1), 4–6 (A2), 7–9 (B1), 10–12 (B2), 13–15 (C1), 16 (C2). On voit que le niveau C2 est sous-représenté puisque seuls 8 textes sont rédigés par l'apprenant, une fois ce niveau atteint.

Notons que, vu qu'il s'agit d'un système d'apprentissage à distance, les apprenants rédigent les textes à domicile sans aucune surveillance. Ainsi, ils peuvent utiliser des ressources diverses et variées (des dictionnaires, des correcteurs orthographiques, etc.) et également se servir de fragments de textes provenant du Web (aucun système anti-plagiat n'est mentionné dans les descriptions d'*English Live*). Il n'y a donc aucune garantie que la production textuelle reflète le véritable niveau de maîtrise de la langue anglaise par l'apprenant.

1.3. Le concours

Les responsables du concours *My Tailor is rich !* de la conférence CAp 2018 ont sélectionné 40 975 textes à partir du corpus EFCAMDAT et en ont fourni les deux tiers aux participants au concours comme corpus d'apprentissage et un tiers comme corpus de test. Les deux corpus comportaient les textes ainsi que les 58 indices calculés par les organisateurs de la compétition :

- le nombre de phrases ;
- le nombre de mots ;
- le nombre de lettres ;
- le nombre de syllabes ;
- le nombre de signes de ponctuation ;
- le nombre moyen de mots par phrase ;
- la taille moyenne des mots en nombre de caractères ;
- le nombre moyen de syllabes par mots ;

- le nombre de mots par phrase ;
- le résultat d'une formule mettant en jeu le nombre de mots d'une seule syllabe ;
- ainsi qu'une variante mettant également en jeu le nombre mots divisé par le nombre de phrases ;
- l'indice ARI (*Automated Readability Index*). La formule de calcul utilise des coefficients du nombre de mots divisés par le nombre de syllabes et le nombre de prépositions ;
- l'indice Bormuth, qui donne une estimation du niveau scolaire nécessaire pour comprendre un texte. Il est fondé sur la liste des 3 000 mots les plus fréquents en anglais (liste Dale-Chall) ;
- l'indice de lisibilité proportionnel au nombre de lettres et au nombre de phrases (tous les 100 mots) ;
- l'indice de lisibilité (1995), qui reflète le degré de familiarité du lexique utilisé et qui repose sur la liste des 3 000 mots les plus fréquents en anglais (liste Dale-Chall) ;
- les résultats de deux formules de lisibilité qui reposent sur le nombre de caractères utilisés (espaces compris) ;
- un indice de lisibilité qui prend en compte des valeurs proportionnelles au nombre de mots et au nombre de caractères ;
- la mesure des degrés de lisibilité (*Degrees of Reading Power*) à partir de l'indice de Bormuth ;
- l'indice *Easy Listening Formula* : le nombre de mots polysyllabiques divisé par le nombre de phrases ;
- un indice proche de Flesch, mais où, tous les 100 mots, le nombre de mots monosyllabiques simplifie la prise en compte du nombre de syllabes, tous les 100 mots. On retranche à 206 835 les valeurs proportionnelles au nombre de mots divisés. L'indice est compris entre 100 (texte facile à comprendre) et 0 (texte très difficile) ;
- une métrique développée pour les besoins de l'armée américaine afin de proposer une conversion de la difficulté d'un texte en niveau scolaire nécessaire pour le lire ;
- un indice de lisibilité proposé dans les années 1950, censé représenter le nombre d'années d'études nécessaires à la compréhension d'un texte à la première lecture. Il prend en compte la proportion de mots de trois syllabes ou plus ;
- l'indice de lisibilité FORCAST collectivement mis au point avec des conscrits du Vietnam, qui repose sur la longueur des mots. On retranche à 20 le dixième des mots monosyllabiques (sur une fenêtre de 150 mots) ;
- une caractéristique stylistique proposée par W. Fucks (produit du nombre de caractères divisé par le nombre de mots et du nombre de mots divisé par le nombre de phrases) ;
- un indice qui prend en compte le nombre de mots de trois syllabes ou plus, le nombre de mots et le nombre de phrases. Proposé au départ pour l'analyse du suédois, cet indice prend en compte la proportion de mots composés de 7 lettres ou

plus. Les textes ayant un indice inférieur à 25 sont censés être faciles à lire, ceux « normaux » ont un indice autour de 40 et ceux considérés comme difficiles au-delà de 50 ;
- des indices proposés dans les années 1980 pour l'allemand (*Neue Wiener Sachtextformeln*), qui prennent en compte, dans des proportions variables, les mots de trois syllabes ou plus et les mots de six lettres ou plus ;
- une adaptation pour l'anglais de l'indice LIX. Il s'agit du nombre des mots de six lettres ou plus divisé par le nombre de phrases ;
- la *Simple Measure of Gobbledygook* (SMOG). Indice de lisibilité fondé sur la racine carrée du nombre de mots polysyllabiques, calculés au début, au milieu, début et fin de texte ;
- un indice de lisibilité fondé sur le nombre des mots du texte qui ne figurent pas dans la liste de mots de références de Spache ;
- un indice de lisibilité des médias mis au point en 2006 qui divise le nombre de syllabes des trois premières phrases par dix ;
- des indices de lisibilité prenant en compte la proportion de prépositions (*Traenkle.Bailer.TB1*) et de conjonctions (*Traenkle.Bailer.TB2*) ;
- un indice de diversité lexicale fondé sur la probabilité de retrouver un mot dans une fenêtre de 42 mots ;
- l'indice C de Herdan ;
- les indices de complexité lexicale proposés en 1972 qui mettent en jeu les logarithmes des types et des tokens ;
- la moyenne mobile du *type to token ratio*. (*Moving Average of TTR*) calculée par le biais d'une fenêtre mobile. Si le texte comporte moins de 400 mots, le MATTR ne peut pas être calculé et renvoie NA.

Ces indicateurs sont destinés, dans leur écrasante majorité, à refléter la « lisibilité » d'un texte, c'est-à-dire le niveau de compétence en langue requis pour sa lecture et sa compréhension. Il s'ensuit donc que ces indices n'ont pas été créés pour être appliqués à des textes d'apprenants comportant des erreurs et des imperfections, mais à des textes littéraires ou éducatifs rédigés par des personnes maîtrisant parfaitement la langue et souhaitant atteindre différentes populations (élèves des différents échelons de l'éducation, étrangers, personnes atteintes de troubles, etc.).

Le but du concours était non pas la prédiction de la compétence langagière sur 16 niveaux spécifique à *English Live*, mais la prédiction sur les 6 niveaux de l'Union européenne (A1–C2) décrits dans le Cadre européen commun de référence pour les langues (CECR, https://www.coe.int/fr/web/common-european-framework-reference-languages). Pour évaluer les résultats des participants au concours, les organisateurs ont établi une formule de calcul de score, basée sur le tableau suivant :

	A1	A2	B1	B2	C1	C2
A1	0	1	2	3	4	6
A2	1	0	1	4	5	8
B1	3	2	0	3	5	8
B2	10	7	5	0	2	7
C1	20	16	12	4	0	8
C2	44	38	32	19	13	0

Les lignes correspondent à la réalité et les colonnes à des prédictions. Ainsi, lorsqu'un niveau est bien prédit, le score affecté est nul. Lorsqu'un niveau A1 est prédit en tant que C2, le score accordé à cette prédiction est de 6. Inversement, lorsqu'un niveau C2 est prédit en tant que A1, le score accordé est de 44. On voit donc que la sous-estimation du niveau langagier est bien plus sévèrement punie que sa surestimation. Le score final de chaque participant est calculé à partir de la moyenne des scores obtenus dans toutes les prédictions.

2. Nos tentatives de solution

Notre équipe (deux membres de l'équipe DECIDE de l'IMT Atlantique, à savoir Philippe Lenca et l'auteur) a tenté un certain nombre d'approches, qui lui a valu la dixième place dans le palmarès du concours. En voici une description rapide.

2.1. Le niveau orthotypographique

Nous avons considéré qu'un auteur expérimenté de l'anglais sera à l'aise avec l'orthotypographie de la langue et, en particulier, avec l'utilisation de la ponctuation. Ainsi nous avons établi un score de « mauvaise utilisation de la ponctuation » quand, par exemple, une virgule ou un point étaient précédés d'un espace ou n'en étaient pas suivis. Cette étude aurait pu se combiner avec une corrélation sur l'origine des apprenants, puisque les différentes écritures ont différentes conventions de ponctuation, mais l'information sur l'origine des apprenants ne faisait pas partie des données fournies dans le cadre du concours.

2.2. Le niveau orthographique

Nous avons détecté les « coquilles » (mots qui n'apparaissent pas dans le dictionnaire des formes standard de l'anglais ou dans Wikipédia/WordNet en tant qu'entités nommées). Pour ces « coquilles », nous avons affecté des scores différents selon

la distance entre la forme utilisée et une potentielle forme « correcte », la distance pouvant être celle de Levenshtein ou alors la distance (euclidienne) des touches sur le clavier (pour le calcul de cette *distance de clavier* l'information sur l'origine des apprenants serait de nouveau utile, pour tenir compte des différences entre claviers). Enfin, nous avons affecté un score plus important à la récidive : lorsque la même erreur se retrouvait plusieurs fois dans le même texte, le score se trouvait multiplié par un certain facteur.

Nous avons traité à part deux catégories d'erreurs qui nous ont paru caractéristiques : (1) la gestion de la capitalisation, et plus particulièrement celle du pronom personnel *I*, celle des adjectifs nationaux (*English*, *French*, etc.), et des jours de la semaine et des mois (*Monday*, *January*, etc.) ; (2) les erreurs d'élision : **I'am* (au lieu de *I'm*), **I'ts* (au lieu de *It's*), **I'have* (au lieu de *I've*), etc.

2.3. Les erreurs de morphologie

Nous avons compté et pondéré les erreurs d'accord (par exemple, **two friend*), les erreurs de conjugaison (**I sayed* au lieu de *I said*), la confusion entre le participe passé et le gérondif (**I seen him*, au lieu de *I saw him*) et les erreurs de préposition (**I go at the theater*, au lieu de *I go to the theater*), etc.

À l'aide de l'ouvrage très intéressant de Swan & Smith (2001), nous avons répertorié un certain nombre d'erreurs fréquemment commises par différentes nationalités d'apprenants. En particulier, parmi celles-ci, pour les pays les plus représentés parmi les apprenants, figurent les erreurs de *verbe à particule*. Nous avons établi la distribution probabiliste des correspondances entre verbe et particule séparable (par exemple, *add on* = ajouter est plus rare que *add up* = additionner) et nous avons établi un score pénalisant les utilisations inexistantes de verbe à particule tout en gratifiant l'utilisation de verbe à particule plus rare (dont la connaissance présuppose une meilleure compétence linguistique).

2.4. Le niveau lexical

Nous nous sommes posés la question du décider comment établir la probabilité qu'un mot ou un terme (simple ou composé) soit utilisé dans un contexte de langue « simple ».

Une première tentative a été de nous baser sur une comparaison lexicale et terminologique des articles du corpus *Simple English Wikipedia* (SEW) et des articles correspondants du Wikipédia anglais standard. Dans le premier cas, il s'agit d'une encyclopédie Wikipédia rédigée dans un anglais simple, à destination des apprenants, ou des personnes atteintes de troubles. Le corpus SEW est basé (ou du moins est censé être basé) sur un langage contrôlé du nom de *Basic English* (où Basic est l'acronyme

de *Business Academic Scientific International Commercial*) qui comporte un vocabulaire de 850 mots (Ogden, 1930) et un nombre de règles syntaxiques très limité. Dans la classification PENS de Kuhn (2014), le langage *Basic English* est décrit comme étant de classe $P^2E^5N^5S^1$ (langage avec degré d'ambiguïté considérablement inférieur à celui des langages naturels, avec expressivité maximale, naturalité maximale et ayant la complexité d'un langage naturel) avec les propriétés C (orienté humain) et W (destiné à la modalité écrite).

La majorité des 143 060 articles du corpus simplifié provient d'une réécriture des articles du Wikipédia standard. Ainsi, en comparant les deux, on peut établir un indicateur de la « simplicité » de chaque mot ou terme. En guise d'exemple, voici le premier paragraphe de la page consacrée à Cluj-Napoca dans les deux corpus (c'est nous qui soulignons) :

English Wikipedia	Simple English Wikipedia
Cluj-Napoca is the fourth *most populous* city in Romania, and the *seat* of Cluj County in the northwestern part of the country. Geographically, it is roughly equidistant from Bucharest, Budapest and Belgrade. As of 2011, 324,576 *inhabitants lived within the city limits*, marking a slight increase from the figure recorded at the 2002 census.	Cluj-Napoca is the third *biggest* city in Romania, and is the *capital* city of Cluj County, in the north-western part of Transylvania. Bucharest is about 330 kilometers away from Cluj-Napoca. About 330,000 *people live in the city*.

À noter dans cet exemple le traduction de *most populous* par *biggest*, de *seat* par *capital city*, de *inhabitants lived within the city limits* par *people live in the city*, etc. De même, les informations *it is roughly equidistant from Bucharest* et *marking a slight decrease from the figure recorded at the 2002 census* ont été retirées du texte simplifié.

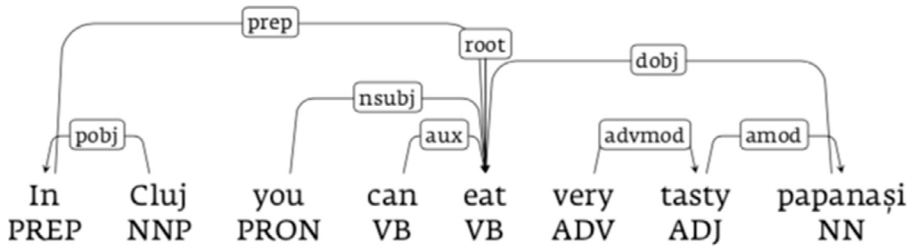
Pour une deuxième tentative, nous avons utilisé les *propriétés psycholinguistiques des mots*, telles que définies par (Gilhooly & Logie, 1980) et implémentées par (Kyle & Crossley, 2014) dans l'outil TAALES (*Tool for the Automatic Analysis of Lexical Sophistication*) :

- la *concrétude* : plus un objet est concret, plus il peut être considéré comme simple (exemple donné dans les articles cités : *apple* est plus concret que *infinity*) ;
- la *familiarité* : un mot est plus simple quand il est utilisé dans le langage de tous les jours (exemple donné : breakfast est plus familier que egress, ce dernier mot signifiant « sortie ») ;
- l'*imageabilité* (facilité de se forger une image, *imageability*) : un mot est plus simple quand on peut le décrire par une image : *beach* est plus imageable que *philology* ;
- l'existence d'un voisinage sémantique large : *beautiful* a un plus large voisinage que *adze* (qui est le terme anglais pour « aissette », un outil de couvreur) ;
- l'âge d'acquisition du terme, selon le programme éducatif américain ou britannique (ainsi, le mot *apple* est appris beaucoup plus tôt que *calculus*).

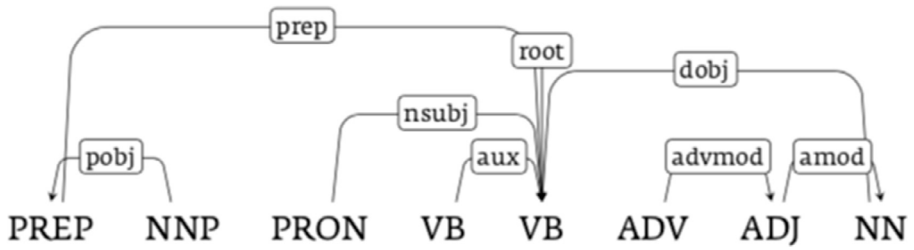
2.5. La syntaxe

La syntaxe est indiscutablement un indicateur potentiel de la compétence langagière de l'apprenant, surtout si la langue maternelle utilise des règles syntaxiques différentes de celles de l'anglais. Parmi les défaillances en matière de syntaxe on peut considérer aussi bien les particularités de la syntaxe de l'anglais qui ont été mal acquises que l'influence pernicieuse de la syntaxe de la langue maternelle de l'apprenant. Pour donner un exemple du deuxième cas, il suffit de considérer la production effectivement observée dans le corpus : **Would you confirming me that you've well been receive this email ?* d'un apprenant de niveau A1 dont on peut raisonnablement supposer qu'il est d'origine francophone puisque cette phrase reprend par mimétisme la syntaxe de la phrase française « pourriez-vous me confirmer que vous avez bien reçu ce mail ? ».

Mais comment évaluer la compétence syntaxique en faisant abstraction des lexèmes utilisés ? À cette fin, nous avons choisi d'utiliser la syntaxe par dépendances :



La phrase *In Cluj you can eat very tasty papanași* est ici analysée selon la syntaxe par dépendances, le verbe *eat* étant la tête de la phrase. Dans cet arbre, on représente des dépendances syntaxiques annotées par leur nature (sujet, COD, auxiliaire, etc.). De même, chaque mot est annoté par sa partie de discours (nom, nom propre, verbe, adverbe, adjectif, etc.). Nous avons analysé syntaxiquement la totalité du corpus, et avons ainsi établi un corpus d'arbres avec les natures de dépendance syntaxique et les parties du discours (mais sans les mots) :



Nous avons également établi un deuxième corpus dans lequel nous avons retiré les parties du discours. À partir de ces deux corpus d'arbres décrivant les phrases des productions, nous avons créé un plongement de graphes en utilisant l'approche *graph2vec* (Narayanan *et al*., 2017). Ce plongement nous a permis de calculer la similarité des graphes en tant que distance euclidienne dans un espace vectoriel de dimension 128.

2.6. Le réseau de neurones résultant

Nous avons combiné tous les indicateurs décrits ci-dessus (§§ 2.1–2.5) dans un réseau de neurones *feed-forward* à deux entrées (données textuelles et indicateurs d'un côté, arbres syntaxiques de l'autre) et à quatre couches. Nous avons donc utilisé 675 dimensions en entrée pour en sortir 6 (la probabilité de classification dans chacun des niveaux A1–C2).

Ce réseau nous a permis d'atteindre une précision de 89,86 % sur la prédiction de la compétence langagière (contre 98,2 % pour le gagnant du concours cf. ci-dessous).

3. La solution gagnante

Le concours a été gagné par Georgios Balikas (*Kelkoo Group*, Grenoble). Dans un article décrivant sa méthode, (Balikas, 2018) nous apprend qu'il a utilisé des modèles statistiques de langue, des clusters de mots et des topiques calculées par la méthode LDA (allocation latente de Dirichlet). Ces techniques lui ont servi à classer les textes par leurs thématiques. Indépendamment de la qualité du travail de M. Balikas, qui est excellente et mérite amplement le prix gagné, nous pouvons nous poser un certain nombre de questions liées au concours. Pour commencer, pourquoi une recherche de thématique serait-elle pertinente lorsqu'il s'agit d'évaluation de niveau de compétence langagière ? Certaines thématiques seraient-elles plus complexes et donc plus aptes à être abordées par des locuteurs expérimentés ?

On pourrait faire cette hypothèse s'il s'agissait d'apprenants jeunes, dont le domaine de connaissances (et donc aussi de compétences langagières) serait corrélé à leur âge. Mais dans notre cas, il s'agit d'apprenants adultes, de tous âges, de tous niveaux d'éducation et de toutes les catégories socioprofessionnelles.

Comment cela se fait-il donc que les thématiques des textes ont-elles joué un rôle prépondérant dans la solution gagnante ?

La raison est toute simple (et M. Balikas en parle lui-même dans la conclusion de son article) : les textes du corpus EFCAMDAT *proviennent d'exercices d'écriture dont les thématiques sont prédéfinies et dépendent strictement du niveau de compétences affecté à chaque apprenant*. Ainsi, le système de M. Balikas a, en réalité, prédit l'exercice pédagogique qui a été à l'origine de chaque texte, et la correspondance entre celui-ci et le

niveau de compétence langagière est sans ambiguïté (modulo les éventuels rapprochements de thématiques dans les différents niveaux).

On peut donc porter un jugement favorable à la solution de M. Balikas en disant qu'elle utilise un aspect du corpus que nous (et sans doute aussi d'autres perdants du concours) n'avons pas entrevu. Il a agi en tant que *data scientist* et a identifié une tendance des données, une « pépite de connaissance » (pour utiliser le vocabulaire de la fouille de données) que nous autres avons ignorée. Mais on peut aussi dire que, de notre côté, nous avons voulu résoudre un problème plus général : la classification de textes *quelconques* par niveau de compétence langagière, en nous basant accessoirement sur EFCAMDAT en tant que corpus d'apprentissage. Nous avons agi en tant que linguistes désireux de résoudre un problème plus général et afférent à l'apprentissage d'une langue, quel que soit le corpus de productions des apprenants. Il serait d'ailleurs intéressant de comparer notre méthode avec celle de M. Balikas sur un corpus de textes de thématiques quelconques (même si le biais thématique existera toujours puisqu'on peut difficilement imaginer une approche pédagogique où toutes les rédactions sont à thème libre).

Ce qui nous ramène à la question plus générale : dans quelle mesure le concours *My Tailor is rich !* peut-il être défini comme étant une « prédiction du niveau en anglais à partir de production écrite d'apprenants » tel qu'il est le cas dans le document descriptif (Ballier *et al*., 2018) ?

À bien y réfléchir, la définition donnée n'est pas trompeuse puisque le degré de généralité de la solution souhaitée n'est explicité : on pourrait tout au plus parler de « péché d'omission ». Une définition plus proche de la réalité (mais moins intéressante sur le plan de la communication) aurait pu être « prédiction du niveau de compétence attribué à l'apprenant dans le cas spécifique du corpus EFCAMDAT ».

Discussion

Ce concours est un exemple caractéristique de la tendance des méthodes d'apprentissage automatique (et plus particulièrement de *deep learning*, c'est-à-dire d'apprentissage automatique faisant intervenir les réseaux de neurones profonds) appliqué au traitement automatique de la langue. On y considère les données linguistiques comme des amas de données hétéroclites de toutes sortes et on essaie, par le pouvoir du calcul, de faire des prédictions qui, dans le cas de l'être humain, nécessiteraient des connaissances linguistiques et culturelles.

Un illustre exemple de cette tendance est la méthode de classification de textes décrite dans (Zhang *et al*., 2016) : à l'aide d'un réseau de neurones convolutionnel à 27 couches, les auteurs réussissent à classifier des textes sans recourir à la moindre ressource linguistique ! En effet, chaque texte est codé comme une suite de caractères, alors que les notions de mot, de partie de discours, de morphologie, de syn-

taxe, de sémantique sont totalement absentes (les partisans du *deep learning* diront que toutes ces notions sont potentiellement re-inventées par le réseau de neurones s'il en a besoin pour sa tâche).

Que nous apprend cette approche sur la manière de classifier des textes ? Rien. Il en est de même pour l'approche qui consiste à dégager les thématiques de textes afin de les classer par niveau de compétence langagière : elle ne nous apprend rien sur la compétence langagière en soi, tout au plus nous informe-t-elle sur les choix de thématique des exercices de rédaction, mais cette information est déjà disponible sur le site d'*English Live*. Pourtant, ces méthodes sont suffisamment efficaces pour gagner des concours et leur efficacité est proportionnelle à la puissance de calcul des machines dont on se sert pour les appliquer.

Cette tendance actuelle de la discipline de traitement automatique de la langue, présage-t-elle une évolution de celle-ci ? S'orientera-t-elle vers la résolution de problèmes pratiques (le domaine n'en manque pas : la traduction automatique en temps réel, la recherche intelligente, la création d'agents de discussion autonomes, etc.) de manière de plus en plus efficace mais sans support théorique, ou arrivera-t-elle à combiner et à faire interagir l'étude théorique de l'objet langue et l'utilisation de méthodes opaques (mais efficaces) comme les réseaux de neurones ? L'avenir nous le dira.

Bibliographie

Balikas, G. (2018). Lexical Bias in Essay Level Prediction. https://arxiv.org/pdf/1809.08935.pdf

Ballier, N. (2018). Appel à participation à la compétition « My tailor is rich » de CAp 2018. Prédiction du niveau en anglais à partir de production écrite d'apprenants. http://cap2018.litislab.fr/competition_fr.pdf

Gilhooly, K. J. & Logie, R. H. (1980). Age of acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation* 12, 395–427.

Haralambous Y. & Lenca, P. (2018). Response to the « My Tailor is rich ! » challenge. https://perma.cc/NED3-GFUH

Huang, Y. *et al.* (2017). The EF Cambridge Open Language Database. https://corpus.mml.cam.ac.uk/efcamdat2/public_html/EFCamDat-Intro_release2.pdf

Kyle, K. & Crossley, S. A. (2014). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol* 49, 757–786.

Kuhn, T. (2014). A Survey and Classification of Controlled Natural Languages. *Computational Linguistics* 40 (1), 121–170.

Narayanan, A. *et al*. (2017). Graph2vec: Learning distributed representations of graphs. https://arxiv.org/pdf/1707.05005

Ogden, C. K. (1930). *Basic English: A general introduction with rules and grammar*. London : Paul Treber & Co.

Swan, M. & Smith, B. (2001). *Learner's English: A teacher's guide to interference and other problems*. Cambridge: Cambridge University Press.

Zhang, X., Zhao, J. & Le Cun, Y. (2016). Character-level convolutional networks for text classification. https://arxiv.org/pdf/1509.01626