Appel à participation à la compétition « *my taylor is rich* » de CAp 2018

Prédiction du niveau en anglais à partir de production écrite d'apprenants

1 Annexes : précisions

1.1 Sur le cadre général

Ce cadre européen, grâce aux descripteurs de compétences qu'il présente pour chaque niveau, permet d'asseoir sur une base solide et objective la reconnaissance réciproque des qualifications en langue. L'étalonnage qu'il fournit permet d'élaborer des référentiels cohérents dans chaque langue et pour chaque niveau commun de l'échelle et aide les enseignants, les élèves, les concepteurs de cours et les organismes de certification à coordonner leurs efforts et à situer leurs productions les unes par rapport aux autres[1].

La recherche linguistique sur les productions, essentiellement écrites (compilées dans des corpus d'apprenants), s'inspire de plus en plus des techniques d'analyse de l'apprentissage automatique. Une partie de la communauté scientifique des corpus d'apprenants cherche à établir des traits critériés pour enrichir les descripteurs des niveaux du CERL.

1.2 Sur les classes

Comme le précise le site Eduscol pour le système scolaire français,« *l'échelle de compétence langagière globale fait apparaître trois niveaux généraux subdivisés en six niveaux communs (au sens de large consensus) :*

Niveau A : utilisateur élémentaire (= scolarité obligatoire), lui-même subdivisé en niveau introductif ou de découverte (A1) et intermédiaire ou usuel (A2).

Niveau B : utilisateur indépendant (=lycée), subdivisé en niveau seuil (B1) et avancé ou indépendant (B2). Il correspond à « une compétence opérationnelle limitée » (Wilkins) ou « une réponse appropriée dans des situations courantes » (Trim).

Niveau C : utilisateur expérimenté, subdivisé en C1 (autonome) et C2 (maîtrise). »

Ces niveaux balisent l'apprentissage des langues étrangères. Notamment, C2 ne doit pas être confondu avec la compétence langagière du locuteur natif. Celle-ci se situe au-delà et ne peut donc plus constituer le modèle idéal à partir duquel est évaluée la compétence en langue des élèves. La figure 1 illustre les équivalences entre es six niveaux communs (de A1 à C2) et d'autres mesures du niveau de langue.

[1] `eduscol.education.fr/cid45678/cadre-europeen-commun-de-reference-cecrl`

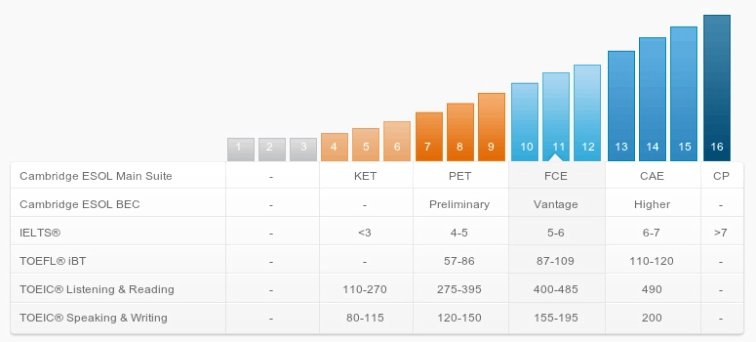

Figure 1: Illustration des équivalences entre les six niveaux communs (de A1 à C2) et d'autres mesures du niveau de langue. Cette illustration est empruntée à `https://corpus.mml.cam.ac.uk/efcamdat1/`.

1.3 Sur les caractéristiques

Les caractéristiques ont été calculées à l'aide du package koRpus de R [Michalke, 2017]. Elles comportent des indices de diversité lexicale (e.g type token ratio, HD-D/vocd-D, MTLD) et des métriques de lisibilité (Flesch-kincaid, SMOG, LIX, Dale-Chall). La documentation de koRpus cite les articles à l'origine des métriques proposées ainsi que les différentes formules de calcul. Elles sont rappelées dans la figure 2, qui indique comment les calculer à partir du nombre de mots, de leur type et des fréquences dérivées.

Metric	Formula
TTR	V/N
MSTTR	V/N (fragments of n tokens)
MTLD	V/factors (segments with the stabilization point of TTR)
MATTR	Mean of moving TTR (window technique)
MTLD-MA	Factors and window technique combined
Herdan's C	logV / logN
Guiraud's RTTR	$V/\sqrt{N}$
Carrol's CTTR	$V/2\sqrt{N}$
Uber Index (U)	$(\log N)^2/\log N - \log V$
Summer's Index (S)	log(logV)/log(logN)
Yule's K	$K = 10^4 \frac{[\sum_{m=1}^{N} f \quad X^2] - N}{N^2}$
Maas a	$a^2 = (\log N - \log V) / \log N^2$
Maas log	$\log V_0 = \log V / \sqrt{1 - \frac{\log V^2}{\log N}}$
HDD-D	For each type, the probability of finding any of its tokens in a random sample of 42 words taken from the same text

Figure 2: Tableau indiquant la définition de différentes métriques. Dans le tableau, N désigne le nombre de tokens (tous les mots du textes) V le nombre de formes lematisées différentes, X le vecteur de fréquence de chaque type contenu dans un texte et f le vecteur des fréquences de chaque X (adapté de [Lissón et al., 2018]).

Des caractéristiques indiquant la lisibilité du texte ont été ajoutées. La lisibilité « désigne le degré de facilité avec un texte peut être lu » [François, 2011]. Les métriques de lisibilité évaluent la difficulté de lecture d'un texte. Elles ont souvent été mises au point pour permettre de faire correspondre à des textes des niveaux d'élèves dans le système américain (école primaire jusqu'au niveau 5, secondaire jusqu'au niveau 12, université au-delà). Les indices proposés sont souvent issu de l'estimation des paramètres d'un modèle de régression linéaire mettant en jeu des fonctions du nombre de mots et de la longueur des mots utilisés. La complexité lexicale repose essentiellement sur la longueur des mots utilisés. La sophistication lexicale mesure la richesse du lexique à l'aide d'inventaires de références.

Voici la liste des caractéristiques proposée dans l'ordre ou elles sont proposées dans le fichier `train_cap2018.cvs`, associées à leur position dans le fichier. Le détail des formules de calcul de certaines caractéristiques est précisée dans la figure 2 :

2. sentences : nombre de phrases.
3. words: nombre de mots.
4. letters.all : nombre de lettres.
5. syllables : nombre de syllabes.
6. punct : nombre de signes de ponctuation.
7. avg.sentc.length : nombre moyen de mots par phrase.
8. avg.word.length : taille moyenne des mots en nombre de caractères.
9. avg.syll.word : nombre moyen de syllabes par mots.
10. sntc.per.word : nombre de phrases par mot.
11. TTR : mesure brute du *type to token ratio*.
12. ARI : index de lisibilité (*Automated Readability Index*). La formule de calcul met en jeu des coefficients du nombre de mots divisés par le nombre de syllabes et le nombre de prépositions.
13. Bormuth: index de lisibilité 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).
14. Coleman.C1: formule mettant en jeu le nombre de mots d'une seule syllabe.
15. Coleman.C2: variante mettant en jeu également en jeu le nombre mots divisé par le nombre de phrases.
16. Coleman.C3 variante prenant également en compte la proportion de pronoms.
17. Coleman.C4 variante prenant également en compte la proportion de mots qui sont des prépositions.
18. Coleman.Liau : indice de lisibilité proportionnel aux nombre de lettres et au nombre de phrases (tous les 100 mots).
19. Dale.Chall : 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).
20. Danielson.Bryan.DB1 & Danielson.Bryan.DB2 : deux formules de lisibilité qui reposent sur le nombre de caractères utilisés (espaces comprises).
21. Dickes.Steiwer: indice de lisibilité qui prend en compte des valeurs proportionnelles au nombre de mots, au nombre de caractères.
22. DRP: mesure des degrés de lisibilité (*Degrees of Reading Power*) à partir de l'indice de Bormuth.
23. ELF : (*Easy Listening Formula*): le nombre de mots polysyllabiques divisé par le nombre de phrases.
24. Farr.Jenkins.Paterson: indice proche de Flesch, mais où le nombre de mots monosyllabiques tous les 100 mots simplifie la prise en compte du nombre de syllabes tous les 100 mots.
25. Flesch : métrique sensible à la langue utilisées. Ce sont les valeurs pour l'anglais qui ont été

utilisées. On retranche à 206,835 des valeurs proportionnelles au nombre de mots divisés. L'indice est compris entre 100 (texte facile à comprendre) et 0 (texte très difficile).

26. Flesch.Kincaid : métrique développée pour les besoins de l'armée américaine pour proposer une conversion de la difficulté d'un texte en niveau scolaire nécessaire pour le lire.
27. FOG : indice de lisibilité proposé dans les années cinquante. Il est 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.
28. FORCAST : indice de lisibilité collectivement (FORCAST = Patrick FORd, John CAylor and Thomas STicht) 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).
29. Fucks : 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).
30. Linsear.Write : indic qui prend en compte le nombre de mots de trois syllabes ou plus, le nombre de mots et le nombre de phrases.
31. LIX : proposé au départ pour l'analyse du suédois, cet indice prend en compte la proportion de mots de 7 lettres ou plus. Les textes qui ont un indice inférieur à 25 sont censés être facile à lire, sont « normaux » autour de 40 et considérés comme difficiles au-delà de 50.
32. nWS1 à nWS4 : Indices proposés dans les années quatre-vingts 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.
36. RIX : adaptation pour l'anglais de l'indice LIX. Il s'agit du nombre de mots de six lettres ou plus divisé par le nombre de phrases.
37. SMOG : *Simple Measure of Gobbledygook* (SMOG). Indice de lisibilité fondé sur la racine carré du nombre de mots polysyllabiques, calculés au début, au milieu, début et fin de texte.
38. Spache : indice de lisibilité fondé sur le nombre de mots du texte qui ne figurent pas dans la liste de mots de références de Spache.
39. Strain : indice de lisibilité des médias mis au point en 2006 qui divise le nombe de syllabes des trois premières phrases par dix.
40. Traenkle.Bailer.TB1 & Traenkle.Bailer.TB2 : indices de lisibilité prenant en compte la proportion de prépositions (Traenkle.Bailer.TB1) et de conjonctions (Traenkle.Bailer.TB2).
42. TRI (Kuntzsch's Text-Redundanz-Index) indice de lisibilité fondé sur une mesure de la redondance. Proposé au départ pour l'analyse des jouranaux allemands, il met en jeu le nombre de signes de ponctuation et le nombre de mots étrangers.
43. Tuldava: indice réputé ne pas être sensible à la langue analysée, il met en jeu le logarithme du nombre de mots divisé par le nombre de phrases.
44. Wheeler.Smith: indice de lisibilité proposé dans les années cinquante, qui met en jeu le nombre de deux syllabes ou plus.
45. text : numéro du texte (arbitraire)
46. CTTR : algorithme proposé par Carroll pour corriger le TTR.
47. HD-D (vocd-D): indice de diversité lexicale fondé sur la probabilité de retrouver un mot dans une fenêtre de 42 mots.
48. Herdan's C : indice C de Herdan, (*cf* figure 2).
49. Maas & lgV0 : indices de complexité lexicale proposés en 1972 qui mettent en jeu les logarithmes des types et des tokens, (*cf* figure 2).
51. MATTR: 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.

52. MSTTR (Mean Segmental Type-Token Ratio): moyenne des TTR sur les segments considérés.
53. MTLD (Measure of Textual Lexical Diversity): mesure correctrice du TTR.
54. Root TTR : racine carrée du TTR.
55. Summer: indice lexical, (*cf* figure 2).
56. TTR.1 : *type to token ratio* (arrondi à deux chiffres après la virgule). C'est la même variable que la variable 11 TTR, mais arrondie.
57. Uber : indice proposé en 1978, (*cf* figure 2).
58. Yule's K : indice de diversité lexicale proposé par Yule en 1944 , (*cf* figure 2).
59. level: niveau de référence, variables à prédire (de A1 à C2).

References

[François, 2011] François, T. (2011). *Les apports du traitement automatique du langage à la lisibilité du français langue étrangère.* PhD thesis, Ph. D. thesis, Université Catholique de Louvain. Thesis Supervisors: Cédrick Fairon and Anne Catherine Simon.

[Lissón et al., 2018] Lissón, P., Ballier, N., and Linguistics, E. (2018). Investigating learners' progression in French as a Foreign Language: vocabulary growth and lexical diversity. CUNY Student Research Day. Poster.

[Michalke, 2017] Michalke, M. (2017). *koRpus: An R Package for Text Analysis.* (Version 0.10-2).