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)*

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

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

Ces niveaux balisent l'apprentissage des langues étrangères. 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.

`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 [m.eik michalke, 2017] et 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 du package cite les articles à l'origine des métriques proposées. Les différentes

Les métriques de lisibilité. Elles ont souvent été mises au point pour permettre de faire corrspondre à des niveux d'élèves.

1. sentences : nombre de phrases
2. words: nombre de mots
3. letters.all : nombre de lettres
4. syllables : nombre de syllabes
5. punct : nombre de signes de ponctuation
6. avg.sentc.length : nombre moyen de mots par phrase
7. avg.word.length : taille moyenne des mots en chaînes de caractères
8. avg.syll.word : nombre moyen de syllabes par mots
9. sntc.per.word : num 0.0687 0.0389 0.0556 0.0556 0.0588 ...
10. TTR : type to token ratio
11. ARI : index de lisibilité ARI

12. Bormuth: index de lisibilité Bormuth (fondé sur les tests de Clausure)
13. Coleman.C1
14. Coleman.C2
15. Coleman.C3
16. Coleman.C4
17. Coleman.Liau
18. Dale.Chall
19. Danielson.Bryan.DB1
20. Danielson.Bryan.DB2 :
21. Dickes.Steiwer
22. DRP
23. ELF
24. Farr.Jenkins.Paterson:
25. Flesch
26. Flesch.Kincaid :
27. FOG :
28. FORCAST :
29. Fucks :
30. Linsear.Write :
31. LIX :
32. nWS1
33. nWS2
34. nWS3
35. nWS4
36. RIX
37. SMOG
38. Spache
39. Strain

40. Traenkle.Bailer.TB1 : indice de lisibilité
41. Traenkle.Bailer.TB2 : indice de lisbilité
42. TRI (Kuntzsch's Text-Redundanz-Index) indice de lisibilité fondé sur une mesure de la redondance
43. Tuldava : num 4.17 4.83 4.32 4.51 4.29 4.53 4.94 3.77 3.83 3.79 ...
44. Wheeler.Smith : num 49.4 82.9 60 63 57.3 ...
45. text : int 30 9 23 36 48 46 2 3 8 44 ...
46. CTTR : num 6.35 6.06 6.8 6.43 5.64 6.11 5.52 4.86 6.5 5.3 ...
47. HD-D (vocd-D) : num 34.6 34.6 37.5 36.9 34.6 ...
48. Herdan's C : num 0.9 0.91 0.94 0.93 0.9 0.92 0.92 0.87 0.92 0.9 ...
49. Maas a : num 0.2 0.2 0.17 0.18 0.21 0.19 0.19 0.24 0.18 0.21 ...
50. Maas lgV0 : num 4.96 5.07 5.99 5.49 4.61 5.13 5.04 3.99 5.52 4.5 ...
51. MATTR : num 0.72 0.74 0.77 0.77 0.68 0.71 0.73 0.62 0.73 0.66 ...
52. MSTTR : num 0.7 0.76 0.78 0.81 0.64 0.73 0.73 0.6 0.78 0.63 ...
53. MTLD : num 70.7 107.1 161.9 141.1 70.7 ...
54. Root TTR : num 8.98 8.57 9.62 9.09 7.97 8.65 7.81 6.87 9.19 7.5 ...
55. Summer : num 0.88 0.89 0.92 0.9 0.87 0.89 0.89 0.83 0.9 0.86 ...
56. TTR.1 : num 0.59 0.64 0.72 0.68 0.58 0.64 0.68 0.51 0.68 0.6 ...
57. Traenkle.Bailer.TB2 : num 39.3 37 45.5 35.9 43.7 ...
58. TRI : num -24.22 -4.04 -9.88 -11.22 -30.28 ...
59. Tuldava : num 4.17 4.83 4.32 4.51 4.29 4.53 4.94 3.77 3.83 3.79 ...
60. Wheeler.Smith : num 49.4 82.9 60 63 57.3 ...
61. text : int 30 9 23 36 48 46 2 3 8 44 ...
62. CTTR : num 6.35 6.06 6.8 6.43 5.64 6.11 5.52 4.86 6.5 5.3 ...
63. HD-D (vocd-D) : num 34.6 34.6 37.5 36.9 34.6 ...
64. Herdan's C : num 0.9 0.91 0.94 0.93 0.9 0.92 0.92 0.87 0.92 0.9 ...
65. Maas a : num 0.2 0.2 0.17 0.18 0.21 0.19 0.19 0.24 0.18 0.21 ...
66. Maas lgV0 : num 4.96 5.07 5.99 5.49 4.61 5.13 5.04 3.99 5.52 4.5 ...

67. MATTR : num 0.72 0.74 0.77 0.77 0.68 0.71 0.73 0.62 0.73 0.66 ...
68. MSTTR : num 0.7 0.76 0.78 0.81 0.64 0.73 0.73 0.6 0.78 0.63 ...
69. MTLD : num 70.7 107.1 161.9 141.1 70.7 ...
70. Root TTR : num 8.98 8.57 9.62 9.09 7.97 8.65 7.81 6.87 9.19 7.5 ...
71. Summer : num 0.88 0.89 0.92 0.9 0.87 0.89 0.89 0.83 0.9 0.86 ...
72. TTR.1 : num 0.59 0.64 0.72 0.68 0.58 0.64 0.68 0.51 0.68 0.6 ...
73. Uber index : num 24.3 26.1 35.1 30.1 22 ...
74. Yule's K indice de diversité lexicale proposé par Yule
    : num 147 142 63 71 124 ... Lexical diversity: Yule's K
75. level niveau de référence à prédire (de A1 à C2)
76.

Readability: Dale-Chall Readability Formula Readability: Automated Readability Index (ARI) index de lisibilité ARI Lexical diversity: Carroll's corrected TTR (CTTR)

1. sentences: Number of sentences.
2. words: Number of words.
3. letters: Named vector with total number of letters ("all") and possibly several entries
4. called "l<digit>", giving the number of words with <digit> letters.
5. all.chars: Number of all characters, including spaces.
6. syllables: Named vector with the number of syllables, simlar to letters, but entries are
7. called "s<digit>" (NA if hyphenation was skipped).
8. lttr.distrib: Distribution of letters: Absolute numbers, cumulative sum, inversed cumu-
9. lative sum, percent, cumulative percent, and inversed cumulative percent. syll.distrib: Distributionofsyllables(seelttr.distrib,NAifhyphenationwasskipped). syll.uniq.distrib: Distributionofuniquesyllables(seelttr.distrib,NAifhyphenation
10. was skipped).
11. punct: Number of punctuation characters.
12. conjunctions: Number of conjunctions.

13. prepositions: Number of prepositions.
14. pronouns: Number of pronouns.
15. foreign: Number of foreign words.
16. TTR: Type-token ratio.
17. avg.sentc.length: Average number of words per sentence.
18. avg.word.length: Average number of characters per word.
19. avg.syll.word: Average number of syllables per word (NA if hyphenation was skipped). sntc.per.word: Number of sentences per word.
20. sntc.per100: Number of sentences per 100 words.
21. lett.per100: Number of letters per 100 words.
22. syll.per100: Number of syllables per 100 words (NA if hyphenation was skipped). FOG.hard.words: Number of hard words, counted according to FOG (NULL if measure was
23. not computed).
24. Bormuth.NOL: Number of words not on the Bormuth word list (NULL if measure was not
25. computed).
26. Dale.Chall.NOL: Number of words not on the Dale-Chall word list (NULL if measure was not computed).
27. Harris.Jacobson.NOL: NumberofwordsnotontheHarris-Jacobsonwordlist(NULLifmea-
28. sure was not computed).
29. Spache.NOL: Number of words not on the Spache word list (NULL if measure was not com-
30. puted).
31.

Trois types de caractéristiques pourraient être disponibles :
- f1 : texte intégral (environ 160 mots)
- f2 : metriques de complexité lexicale (65 métriques dites d'annotation)

Table 1: Synthèse des principales métriques implémentées dans LCA (Lu 2012, Lu2014).

Métriques	Signification
Sentence	Le nombre de segments séparés par un point dans chaque transcription
Wordtypes, swordtypes	Trois indicateurs dénombrant le nombre de mots en fonction de leur catégorie grammaticale
Lextype, slextypes	Indicateurs du nombre de lemmes lexicaux
Wordtokens, swordtokens	Indicateurs du nombre de tokens liés aux mots différents
lextokens, slextokens	Indicateurs du nombre de tokens lexicaux liés aux lemmes différents
Ld	Indice de densité lexicale (= nombre de mots lexicaux et non grammaticaux/nombre de tokens)
ls1, ls2	Proportions de lexèmes parmi les 2000 mots les plus fréquents (Laufer & Nation 1994)
vs1, vs2, cvs1	Ratio de verbes ne figurant parmi les 20 ou 200 verbes les plus fréquents dans un français / nombre de verbes utilisés (et ses variantes, cf. Wolfe-Quintero et al. 1998)
ndw, ndwz, ndwerz, ndwesz	Indices fondés sur le nombre de mots différents pris dans des échantillons de 50 items.
ttr, msttr, cttr, rttr, logttr	Indices fondés sur le ratio entre le nombre de mots différents et le nombre total de mots (type-to-token ratio et ses avatars normalisés ou transformés :TTR, MSTTR, CTTR, RTTR, et LogTTR).
lv, vv1, svv1, cvv1, vv2	Indices de variation concernant le lexique et les formes verbales
nv, adjv, advv, modv	Indices de variation concernant les noms, adjectifs, adverbes et modificateurs (englobant adjectifs et adverbes)

En voici une description détaillée. Il y a 65 variables dans le prototype transmis. Le nombre de phrases, de mots, de lettres, de syllabes, les Type-to-Token ratio et les mesures dérivées. La lisibilité "désigne le degré de facilité avec un texte peut être lu" (François, 2011). Les indices proposés sont souvent de la forme d'une régresion linéaire mettant en jeu le nombre de mots et la longueur des mots utilisés. La sophistication lexicale mesure la richesse du lexique à l'aide d'inventaires de références.

L'analyseur de complexité lexicale développé par Xiaofei Lu (Lu 2012) produit ce type de métriques. Ce *Lexical Complexity Analyzer* est décrit dans plusieurs travaux de Xiofei Lu (Lu 2013, Lu 2014) et à été mis au point pour l'anglais. Il suppose un choix de variété de référence, britannique ou américaine, les fréquences lexicales de référence ayant été calculées (respectivement) à partir du British National Corpus ou de l'American National Corpus. L'analyseur LCA suppose en entrée un format de données lemmatisées (au format TreeTagger) et produit 24 mesures. Nous reportons par commodité ici les principaux indices et renvoyons à la lecture de (Lu 2012) pour une discussion critique des différents indices. Les mesures portent sur la densité lexicale, la variation lexicale et l'élaboration lexicale (sophistication). Les mesures de densité lexicale portent sur les types, les*tokens* et sur le ratio de certaines catégories. La variation lexicale est essentiellement comptabilisée à partir du type-to-token ratio et de ses variantes. Les mesures d'élaboration lexicale reposent sur des proportions des unités lexicales plus élaborées.

References

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