"*My taylor is rich*"
CAp 2018 Competition: Call for Participation

Machine learning level prediction competition in conjunction with CAp 2018.

1 Appendix: precisions and details

1.1 Global framework

The Common European Framework of Reference for Languages (CERL), with its competency descriptors for each level, provides a solid and objective basis for the mutual recognition of language qualifications. The calibration it provides helps to develop consistent benchmarks in each language and for each common level of competency to helps teachers, students, course designers and certification bodies coordinate their efforts and evaluate their productions in relation to each other[1].

Linguistic research on productions, essentially written (compiled in corpora of learners), is increasingly inspired by automatic learning analysis techniques. Part of the scientific community of learners seeks to establish new relevant features to enrich the descriptors of CERL levels.

1.2 About the classes

As specified on the ISF web site, *"foreign language proficiency is measured on a six level scale from A1, for beginners, up to C2, for those who master the language. The framework and its six level scale of progression is a widely accepted a standard not only in Europe but also increasingly worldwide"*[2]:

Level A : basic language skills, A1 – Breakthrough (or beginner) and A2 – Waystage (or elementary),

Level B : independent use of language, B1 – Threshold (or intermediate) and B2 – Vantage (or upper intermediate),

Level C : proficient use of language. C1 – Effective Operational Proficiency (or advanced) and C2 Mastery (or proficiency) can understand with ease virtually everything heard or read.

These levels guide the learning of foreign languages. In particular, C2 should not be confused with the native speaker's language proficiency. This is beyond and can therefore no longer be the ideal model from which students' language proficiency is assessed. Figure 1 illustrates the equivalences between the six common levels (A1 to C2) and other measures of language level.

[1] `https://www.coe.int/t/dg4/linguistic//Cadre1_en.asp`
[2] `https://www.lsf-france.com/info/cefr-language-levels/`

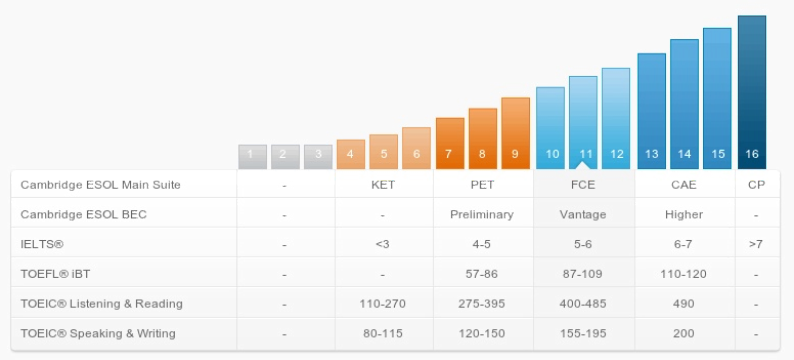

Figure 1: Illustration of equivalences between the six CERL common levels (A1 to C2) and other measures of language level. This illustration is borrowed from `https://corpus.mml.cam.ac.uk/efcamdat1/`.

1.3 Features

Features were computed with the R package koRpus [Michalke, 2017] and include indices of lexical diversity (e.g type-to-token ratio, HD-D/vocd-D, MTLD) and readability metrics (Flesch-kincaid, SMOG, LIX, Dale-Chall). Readability corresponds to the ease with which you can read a text. It often corresponds to a grade in the US school system (primary school up to 5th grade, Middle School until 8th grade, highschool up to 12th grade).

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	$a^2 = (logN - \log V) / \log N^2$
Maas log	$\log V_0 = logV/\sqrt{1 - \frac{\log V^2}{\log N}}$
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: definition of some metrics used as feartures. In the table, N denotes the number of words, V the number of type of words, X the frequency vector for each type and f the frequency vector for each X (adapted from [Lissón et al., 2018]).

Lexical complexity is essentially computed as a correlate of word length, be it computed on the basis of number of syllables or number of letters. Lexical sophistication is often measured in relation to reference inventories of frequent words. Lexical variation can be computed with the type-to-token ratio (TTR, number of different "words" in the text divided by total number of words. As TTR is sensitive to the size of the text (TTR decreases when the size of the text increases), several derived metrics have been proposed to smooth this effect. You will find hereafter the list of available features in the dataset together with their position in the `.csv` file. Details for some formulas are given Figure 2.

2. sentences: Number of sentences.
3. words: Number of words.
4. letters: Named vector with total number of letters ("all") and possibly several entries
5. punct: Number of punctuation characters.
6. avg.sentc.length : Average sentence length (number of words per sentence)
7. avg.word.length : Average word length (number of characters per word)
8. avg.syll.word : Average number of syllables per word
9. sntc.per.word : Number of sentences per word.
10. TTR : type to token ratio
11. ARI : Automated Readability Index. it takes into account the number of tokens words divided by the number of syllables and the number of prepositions in the text.
12. Bormuth: Bormuth readability index. it gives an estimation of the grade required to understand the text. The computation is based on the most frequent 3 000 words in English (Dale-Chall list).
13. Coleman.C1: Readability Formulas, taking into account monosyllabic words
14. Coleman.C2: variant also taking into account the number of words divided by the number of sentences
15. Coleman.C3 variant also taking into account the proportion of pronouns among words
16. Coleman.C4 variant also taking into account the proportion of prepositions among words
17. Coleman.Liau : readability index, proportional to the number of letters and sentences (every 100 words)
18. Dale.Chall : readability index (1995), which reflects the degree of familiarity of the lexicon, compared to the 3,000 most frequents words in English (Dale-Chall list).
19. Danielson.Bryan.DB1 & Danielson.Bryan.DB2 : two readability formulas based on the number of characters (space included).
20. Dickes.Steiwer: readability for German that takes into account values proportional to the number of words, characters and TTR
21. DRP: (*Degrees of Reading Power*) measure readibility from Bormuth index.
22. ELF : (*Easy Listening Formula*): number of polysyllabic words divided by the number of sentences.
23. Farr.Jenkins.Paterson: a simplified version of Flesch, where the number of one syllable words per 100 words replaces the number of syllables per 100 words.
24. Flesch : the English values for this language-dependent metrics have been used. The index takes into account the number of syllables. It ranges between 100 (east texts) and 0 (very difficult text).
25. Flesch.Kincaid : this metric was developed with Vietnam draftees to assess the US school grade corresponding to the difficulty level of a text.
26. FOG : readability index suggested in the 1950's. It measures the number of years of study (school grade) required to understand a text on its first reading. It takes into account the number of words per sentence and the proportion of words with three syllables or more.
27. FORCAST : (FORCAST = Patrick FORd, John CAylor and Thomas STicht) a method

implemented with Vietnam draftees, which is based on word length.

28. Fucks : a stylistic feature proposed by W. Fucks. The number of characters divided by the number of words is multiplied by the number of words divided by the number of sentences.
29. Linsear.Write : readability index that takes into account the number of words of three syllables or more, the number of words and the number of sentences.
30. LIX : this readability index was first proposed for Swedish, it takes into account the proportion of words of seven letters or more. Texts with a 25 index are supposed to be easy to read, "normal" texts are around 40 and texts above 50 are considered to be difficult to read.
31. nWS1 to nWS4 : these readability indices proposed in the 80's for the analysis of German (Neue Wiener Sachtextformeln), take into account -in variable proportions- words of three syllables or more and words of six letters or more.
36. RIX : adaptation for English of the LIX index. It takes into account the number of six letters or more divided by the number of sentences.
37. SMOG : *Simple Measure of Gobbledygook* (SMOG). Readability Index based on the square root of the number of polysyllabic words computed at the beginning, middle and end of the text.
38. Spache : readability index based on the number of words of a text that is not in Spache reference inventory of words.
39. Strain : readability index for medias proposed in 2006, which takes into account the number of syllables.
40. Traenkle.Bailer.TB1 & Traenkle.Bailer.TB2 : readability indices taking into account the proportion of prepositions (Traenkle.Bailer.TB1) and conjunctions (Traenkle.Bailer.TB2).
42. TRI (Kuntzsch's Text-Redundanz-Index) readability index initially suggested for German newspapers, it takes into account the number of punctuation symbols and foreign words.
43. Tuldava: a supposedly language-independent readability index that takes into account the logarithm of the number of words divided by the number of sentences.
44. Wheeler.Smith: readability index proposed in the 1650s that takes into account words of two syllables ore more.
45. CTTR : algorithm proposed by Carroll to smooth TTR.
46. HD-D (vocd-D): lexical diversity index based on the likelihood to find a given word in a 42 word window.
47. Herdan's C : log(V) / log(N), where V is the number of types and N the number of tokens.
48. Maas & lgV0 : indices of lexical complexity suggested in 1972, which take into account logarithms of types and tokens
51. MATTR: (*Moving Average of TTR*), computed by means of a mobile window. Returns "NA" if the text has less than 400 words.
52. MSTTR (Mean Segmental Type-Token Ratio): averages TTR over several segments.
53. MTLD (Measure of Textual Lexical Diversity): corrected measure of the TTR
54. Root TTR : rotted square TTR
55. Summer: lexical diversity index,
56. TTR.1 : rounded Type-to-Token ratio (up to a certain point redundant with variable 11)
57. Sumer': lexical index defined Figure 2.
58. Yule's K : lexical diversity index proposed by Yule in 1944.
59. level: the European level (from A1 to C2) of the learner that we try to predict

References

[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).