

Restitution de la Compétition CAp2018

Nicolas Ballier (pour le comité d'organisation de la Compétition CAp2018)

21/6/2018



PLAN

- ▶ Le contexte de la recherche (les corpus d'apprenants)
- ▶ Le jeu de données: le corpus et les caractéristiques
- ▶ Bilan de la compétition : évaluation des méthodes et des résultats
- ▶ Perspectives offertes par la recherche

Le contexte de la recherche : LCR

Corpus d'apprenants : Learner Corpus Research (Société savante, conférence, revues)

les corpus (depuis 1991 à Louvain):

corpus = a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language

a collection of texts assumed to be representative of a given language, dialect, or other subset of language, to be used for linguistic analysis (Francis 1992)

Louvain LCR2011, LCR2013: Bergen, LCR2015: Nijmegen, LCR2017: Bolzano;

The Learner Corpus Association:

<http://www.learnercorpusassociation.org/>

Corpus : EFCAMDAT (Cambridge)



UNIVERSITY OF
CAMBRIDGE

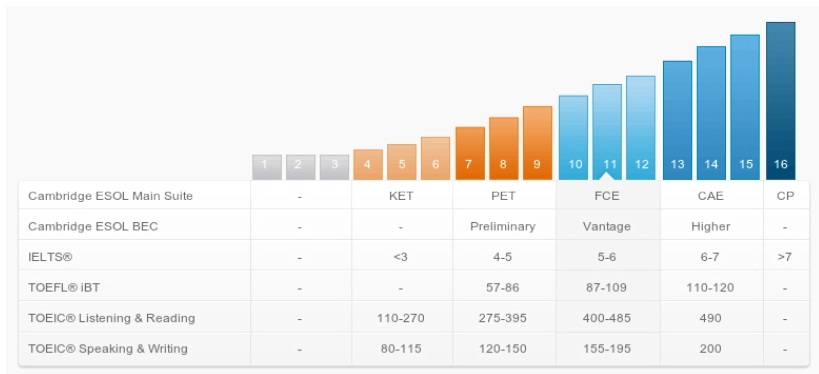
Department of Theoretical
and Applied Linguistics

!!

STANDARD DISCLAIMER !!



La variable à prédire : le niveau européen (CERL)

A1 -> C2 <https://corpus.mml.cam.ac.uk/efcamdat1>



EF-CAMbridge open language DATabase (Geertzen et al. 2013)

<https://corpus.mml.cam.ac.uk/efcamdat1>



UNIVERSITY OF
CAMBRIDGE
Department of Theoretical
and Applied Linguistics

EF-CAMbridge open language **DAT**abase

[Login / Register](#)

Introduction
Select scripts
Query corpus
Export data
Get access
FAQ

Login

Email:

Password:

[Forgotten your password?](#)

Register

When creating a new account for EF-CamDat, you will be asked to agree to the end user license and acknowledge

SELECTION DE VARIABLES DE LA BASE EFCAMDAT

- ▶ langue maternelle (nationalité. . .) unique
- ▶ topic (tâches et biais du corpus)
- ▶ la date
- ▶ le bug lié au copyright
- ▶ la variable 'text' : le petit bug (*mea culpa*) mais change très peu les résultats des meilleures méthodes !

CONTRIBUTION DE LA VARIABLE 'TEXT'

	With text variable		Without text variable	
	Score	Error	Score	Error
1. Balikasg	3.4856	1.41%	8.2455	2.43%
2. Terislepacom	7.2862	2.23%	11.8263	3.57%

AJOUT DE CARACTERISTIQUES

- ▶ Sous {koRpus}, package de R (repris dans {quanteda})
- ▶ complexité lexicale (indices)
- ▶ sophistication lexicale (vs. listes)
- ▶ lisibilité
- ▶ Le texte intégral

Quelques métriques de lisibilité (principaux indices)

Metric	Formula
TTR	V/N
MSTTR	V/N (fragments of n tokens)
MTLD	$V/\text{factors}$ (segments with the stabilization point of TTR)
MATTR	Mean of moving TTR (window technique)
MTLD-MA	Factors and window technique combined
Herdan's C	$\log V / \log N$
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)	$\log(\log V) / \log(\log N)$
Yule's K	$K = 10^4 \frac{[\sum_{m=1}^N f X^2] - N}{N^2}$
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

LA COMPETITION EN UN SCHEMA

Full text	length	words	syllab.	lex_cx	stx_cx ...	
All the world 's a stage, and all the men and women merely players. They have their exits and their entrances; And one man in his time plays many parts.	2	30	34	0.6	53	A1 <input type="checkbox"/> A2 <input type="checkbox"/> B1 <input type="checkbox"/> B2 <input type="checkbox"/> C1 <input type="checkbox"/> C2 <input checked="" type="checkbox"/>
Anyone who feels that if so many more students whom we haven't actually admitted are sitting in on the course than ones we have that the room had to be changed, then probably auditors will have to be excluded, is likely to agree that the curriculum needs revision.	1	48	61	0.43	29	A1 <input type="checkbox"/> A2 <input type="checkbox"/> B1 <input type="checkbox"/> B2 <input type="checkbox"/> C1 <input checked="" type="checkbox"/> C2 <input type="checkbox"/>
Spanish people is very friendly, I'm agree this. You can ask to my friends bob, the one I knew at a party last year.	2	25	28	0.31	5	A1 <input checked="" type="checkbox"/> A2 <input type="checkbox"/> B1 <input type="checkbox"/> B2 <input type="checkbox"/> C1 <input type="checkbox"/> C2 <input type="checkbox"/>

Input data

Output

Distribution des classes (corpus d'entraînement)

EFCAMDAT	CERL	effectif par classe
1-3	A1	11361
4-6	A2	7688
7-9	B1	5383
10-12	B2	2337
13-15	C1	491
16	C2	50

MATRICE DE COUT / COST MATRIX

- Classes & B1/B2 C1/C2

Estimated Reel	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

Cost matrix C .

BILAN DE LA COMPETITION

- ▶ un mois pour le corpus d'entraînement (2/3)
- ▶ un mois pour le corpus de test (1/3)
- ▶ une quinzaine de participants (Colombie. . .)
- ▶ une quarantaine de soumissions
- ▶ 2,5 soumissions / équipe
- ▶ 2 vainqueurs

LE TABLEAU D'HONNEUR

Rank	Team	Score
1.	Balikasg	3.4856
2.	Terislepacom	7.2862
3.	ICSI	8.5970
4.	Caoutchouc	9.4537
5.	ACNK	10.4350
6.	reciTAL team	11.2112
7.	TAU	12.8222
8.	Capitaine-Ad-Hoc	14.1696
9.	Chamlia	17.5161
10.	Haralambous + Lenca Team	17.9774
11.	MB	31.6637
12.	Rufino	33.4285
13.	Team UTC	40.7879
14.	Limsi	41.5202

évaluation des méthodes et des résultats

	Team	Score	Method
1.	Balikasg	3.4856	dedicated features, POS tags, bigram + stratified CV + gradient boosted trees
2.	Terislepacom	7.2862	gradient boosting tree
3.	ICSI	8.5970	BoW (60000 features) + H20
4.	Caoutchouc	9.4537	BoW (22000 features) + numerous classifiers tested (best: log reg)
5.	ACNK	10.4350	NN (LTSM)
6.	reciTAL team	11.2112	n-grams + POS tags + text, num features, feature selection + XGBoost
7.	TAU	12.8222	fastText + PCA on num features + Random forest
8.	Capit Ad-Hoc	14.1696	doc2vec fastText Glove 6B 50D + NN
9.	Chamlia	17.5161	Recurrent Bi-directional network with Attention
10.	Haral-Lenca	17.9774	linguistic indicators, graph2vec NN
12.	Rufino	33.4285	n-grams + token-level features + 58 num features -> 13 feature selected + KNN regressor
13.	Team UTC	40.7879	56 numerical features + XGBoost

packages : grammar-check fastText language_check H20 NLTK spacy

LES METHODES

- ▶ DEUX STRATEGIES : les features seuls / le texte
- ▶ représentations (BOW > Word2Vec)
- ▶ méthodes (forêts aléatoires > RNN ??)
- ▶ prime aux caractéristiques spécifiques (*difficult words*) : python

baseline des valeurs numériques : score 40 (SVM) / Team UTC

5037	516	103	17	8	0
438	3024	345	34	1	2
81	312	2139	155	4	1
19	47	189	873	39	1
2	6	13	53	171	1
0	2	2	5	13	3

LA MATRICE DE CONFUSION DU VAINQUEUR

Perte sur les C2 sans la variable 'text'

Balikasg with text variable

5622	45	11	1	2	0
19	3776	45	1	0	3
3	10	2669	4	5	1
0	3	5	1138	20	2
0	0	2	4	239	1
0	0	0	1	5	19

Balikasg without text variable

5606	59	11	4	1	0
48	3734	60	2	0	0
4	28	2633	22	5	0
2	5	22	1129	9	1
0	1	9	20	215	1
0	2	2	5	9	7

TECHNIQUES FONDEES SUR LE TEXTE

- ▶ les expressions fréquentes
- ▶ “difficultWords” : (textstat 1 python package),
- ▶ number of misspelled words using the publicly available aspell dictionary,2,
- ▶ “duplicates”, the number of duplicate words in each essay,
- ▶ “total vs unique words”, the ratio of words to unique words for each essay,
- ▶ “duplicates”, the number of duplicate words in each essay,
- ▶ “max consecutive nouns”, the number of nouns in the biggest noun phrase,
- ▶ “lower idf than avg”, the number of words in the essay that have lower idf values
- ▶

perspectives offertes par la recherche

- ▶ première compétition pour LCR (learnerathon)
- ▶ risque de surapprentissage (textes très courts, manque de données évaluées)
- ▶ idées pour une autre compétition : les indices de complexité syntaxiques
- ▶ autres caractéristiques : 400 disponibles via : <http://www.ctapweb.com/> Chen, X.B., Meurers, D. (2016). CTAP: A Web-Based Tool Supporting Automatic Complexity Analysis. In Proceedings of The Workshop on Computational Linguistics for Linguistic Complexity. Osaka, Japan. The International Committee on Computational Linguistics

D'autres corpus : MERLIN (allemand, tchèque, italien)

REMERCIEMENTS / BIG UP FOR

- ▶ CAMBRIDGE Team : Dora Alexopoulou , Carlos Balhana
- ▶ Nicolas Ballier (CLILLAC-ARP, Université Paris Diderot)
- ▶ Stéphane Canu (LITIS, INSA Rouen Normandie)
- ▶ Thomas Gaillat (Insight Centre for Data Analytics NUIG, Ireland)
- ▶ Gilles Gasso (LITIS, INSA Rouen Normandie)
- ▶ Caroline Petitjean (LITIS, Université Rouen Normandie)
- ▶ Alain Rakotomamonjy (LITIS, Université Rouen Normandie)

MERCI AU SPONSOR, A CAMBRIDGE ET AUX
PARTICIPANTS



UNIVERSITY OF
CAMBRIDGE

Department of Theoretical
and Applied Linguistics

