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

# Machine learning level prediction competition in conjunction with CAp 2018.

We are pleased to announce the machine learning level prediction competition in conjunction with CAp 2018.

## 1 Objectives

The CAp 2018 conference is hosting the following machine learning competition. The Common European Framework of Reference for Languages (CERL) maps linguistic competence in a foreign language onto six reference levels, described to be shared by European countries: A1, A2, B1, B2, C1 and C2.

The goal of this competition is to achieve, by learning, a system to predict the level of competence of a learner, from one of these written productions comprising between 20 and 300 words and a set of characteristics calculated from this text.

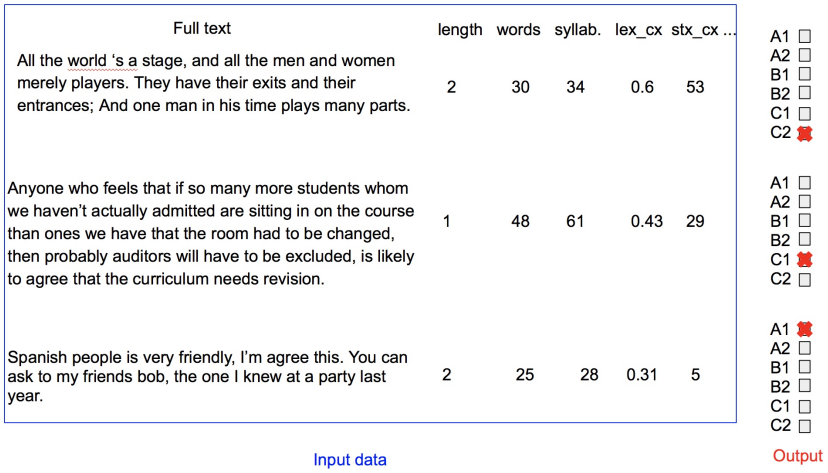


Illustration of the competition task.

# 2 Data set description

To access data competitors have to:

1. register at:
   `https://corpus.mml.cam.ac.uk/efcamdat2/public_html/explore/`

2. after registering, the competitors should then send an email to: `efcamdat.team@gmail.com` with the subject:

   *Request for CAp2018 Shared Task Data*

After we've confirmed that you have registered, we will send out an email with log-in details to the shared task folder within 24 hours.

Data comes from an extraction of the database published by Cambridge and Education First [Geertzen et al., 2013].

Disclaimer: learning and test data was selected and manipulated independently of the participation of the Cambridge and Education First research teams.

The proposed dataset includes 27,310 text examples written by learners and a set of associated characteristics including: `https://corpus.mml.cam.ac.uk/efcamdat1/`.

- lexical complexity metrics,
- readability metrics,
- the prediction variable (the linguistic competence of the person)

## 2.1 Features

Fifty-nine feature variables are available. The first (called it fulltext) is a text: the full text produced by the person to be assessed (with an average of 70 words).

The other 58 variables are metrics calculated from the text. They describe the degree of vocabulary sophistication and the complexity of the text. Among these metrics we find: the number of sentences, words, letters, syllables, the it type-to-token ratio (and derived measures), readability measures calculated from the correlation between the number of words and length of words used and lexical sophistication, which measures the richness of the lexicon using reference inventories. A precise description of the characteristics is available in Section 7.

Two important points regarding features:

1. For the competition, participants can use all or part of the available variables; in particular, it is not mandatory to use the texts.

2. A special jury prize may be awarded to a successful solution that has selected the best features.

### 2.2 Training labels

Classes to be predicted are the 6 reference levels of the CERL (the last variable of the file called it level1). The data proposed on the site are of 16 different levels. Here is how the conversion between the levels estimated by EFCAMDAT and the CERL and the numbers of each class in the learning set is:

| 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 |

The test data has the same proportion by class as the training data.

## 3 Evaluation

The performance measure used will be:

$$E = \frac{1}{n} \sum_{i=1}^{6} \sum_{j=1}^{6} C_{ij} N_{ij}$$

where $N$ is the confusion matrix of general term $N_{ij}$ counting the number of people of class $i$ classified as $j$, $n$ the size of the training set and $C$ the cost matrix defined here after.

| Reel \ Estimated | 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 $C$.

This cost matrix has been calculated as a weighted cross-entropy measure between classes. The probabilities taken into account are the probabilities of appearance of the classes as they appear in the learning and test samples. The weights were given by domain experts to take into account the importance of each class.

# 4 Important dates

- 28 March: call for participation
- 28 April: release of the test set
- 28 May : deadline submission

# 5 Prizes

NVIDIA will give GPU graphics cards (2 or 3) to the top 2 or 3 contributions.

# 6 Organizing Committee

- Nicolas Ballier (CLILLAC-ARP, Université Paris Diderot)
- Stéphane Canu (LITIS, INSA Rouen Normandie)
- Thomas Gaillat (Insight Centre for Data Analytics NUIG, Irland)
- Gilles Gasso (LITIS, INSA Rouen Normandie)
- Caroline Petitjean (LITIS, Université Rouen Normandie)
- Alain Rakotomamonjy (LITIS, Université Rouen Normandie)

# References

[Geertzen et al., 2013] Geertzen, J., Alexopoulou, T., and Korhonen, A. (2013). Automatic linguistic annotation of large scale l2 databases: The ef-cambridge open language database (efcamdat). In *Proceedings of the 31st Second Language Research Forum. Somerville, MA: Cascadilla Proceedings Project.*