

# EFCAMDAT

<https://philarion.mml.cam.ac.uk/>

---

## 概要

### ■ 規模

- ・ 830 万語
- ・ 100 万個の作文
- ・ 174,000 人
- ・ CEFR A1-C2

### ■ 付加情報

- ・ エラー
- ・ 品詞
- ・ 文法依存関係
- ・ 国籍

### ■ ユーザーマニュアル

[https://corpus.mml.cam.ac.uk/faq/EFCamDat-Intro\\_release2.pdf](https://corpus.mml.cam.ac.uk/faq/EFCamDat-Intro_release2.pdf)

## 利用

- ・ ユーザー登録をするだけ。無料。

### ■ データの選択

- ・ 「script」という単位
- ・ Teaching levels and units
  - ・ 1 から 16 のレベルを選んだあと
  - ・ そこに含まれるユニット（テーマ）を選ぶ
- ・ Learner nationalities
  - ・ エリアで選んだあと、
  - ・ 国を選ぶ

### ■ 検索パタンの指定（これをしなければ、データ全体をダウンロードすることになる）

- ・ [word=" 単語 "]
- ・ [pos=" 品詞 "]
- ・ [lemma=" レマ "]
- ・ 連続して複数の項目を指定することもできる
  - ・ [word="the"][pos="N"][word="of"]

## ■データのダウンロード

### 例

- ・日本人の書いたスクリプトは 21,374 個 ( 1,602,328 語 )
- ・ 3,441 人
- ・すべてのレベルから 126 ユニット
- ・ Raw script text のみ XMLformat でダウンロード
  - ・ 圧縮状態で 3.5MB
  - ・ 解凍して 13MB

## XML フォーマット

### ■データの修正

- ・ <selection id="32 の英数文字 "> の部分が、複数の selection がなければ不要。
  - ・ 不要な要素が入ったままだとエラーになる
- ・ <selection> タグの部分削除

### ■属性と要素

- ・ 項目の内容を、項目のタグ名と要素に分けて書けば話は単純
- ・ しかし、「内容」をタグの属性 (Attribute) として表記する方法もある

```
<learner>25</learner>  
<learner id="25"/>  
<learner id="25" age="22"/>
```

- ・ EFCAMDAT は両方の方式でデータが書かれている。
  - ・ xmlconvert で変換するときに手間がかかる

### ■データフレームに変換するスクリプト : xml2dfcamdat()

### 出典

Huang, Y., Murakami, A., Alexopoulou, T., & Korhonen, A. (2018). Dependency parsing of learner English. *International Journal of Corpus Linguistics*, 23(1), 28-54.

Geertzen, J. , Alexopoulou, T., & Korhonen, A. (2013). Automatic linguistic annotation of large scale L2 databases: The EF-Cambridge Open Language Database (EFCAMDAT). *Selected Proceedings of the 31st Second Language Research Forum (SLRF)*, Cascadilla Press, MA.