EXTRACTION D'UN LEXIQUE BILINGUE(MANIP PYTHON)

OBJECTIVE

- Create a bilingual lexicon using a corpus of English and French texts

DEFINITIONOF A BILINGUAL LEXICON

- A bilingual lexicon is a dictionary that associates a word of a source language with the list of its possible translations in a target language.

cuisine	cooks, cook, cooking, kitchen, food...
comprendre	understand, understood, understanding, realize, ...
être	human, being, be, ...
économie	free-market, Economy, economics, market, economy, ...

METHODOLOGY

- The extraction will be *unsupervised* : the method used will take as input only a set of sentence pairs (a sentence in French and its translation in English).
- Such a corpus is called a "parallel corpus" and sentences that are translations of each other are called "parallel".

EXAMPLES OF PARALLEL CORPORA

Living on my own, I really miss my Mom's **cooking**.
Vivant seul, la **cuisine** de ma mère me manque.
She left the **kitchen** with the kettle boiling.
Elle quitta la **cuisine** avec la bouilloire.
Is there any coffee in the **kitchen**?
Y a-t-il encore du café dans la **cuisine**?
Cooking runs in my family.
La **cuisine** c'est de famille.
Both boys and girls should take **cooking** class in school.
Garçons et filles devraient suivre des cours de **cuisine** à l'école.

French-English

Πάω στο σπίτι μας.
Je vais chez nous (*lit.* dans notre maison).
Το σπίτι μου είναι μεγάλο.
Ma maison est grande
Το σπίτι της Έλλης είναι κοντά στην παραλία.
La maison d'Elli est à côté de la plage.
Με λένε 'Ελλη.
Je m'appelle Elli.
Ένα σπίτι του χωριού κάηκε
Une maison du village a brûlé.
Αγαπώ την Έλλη.
J'aime Elli.

French-Greek

« YOU SHALL KNOW A WORD BY THE COMPANY IT KEEPS » (FIRTH, J. R. 1957)

- The proposed approach is based on a linguistic theory, the *distributional hypothesis*: **the meaning of a word can be deduced directly from the context in which it appears**.
- The generalization of this hypothesis to the bilingual case can be formulated as follows:
  - a French word and an English word that *frequently* co-occur (appear in a pair of parallel sentences) have a high chance of being translations of each other.
- Thus, in the example of the corpus on the previous slide, it is natural to assume that *cuisine* can be translated either by *cooking* or by *kitchen* because these are the only words that appear in all translations.

MINI EXERCISE

- Considering the French-Greek corpus in the previous slide, can you give the Greek translations of the words *Elli*, *est* and *maison*.

NAÏVE APPROACH

- You will find on the github two files containing the novel *From the Earth to the Moon* in English (english.corpus) and in French (french.corpus). The documents have been pre-processed to :
  - be aligned at the sentence level: the i^{th} line of the French file is the translation of the i^{th} line of the English file.
  - segmented into words: this segmentation consists of separating some signs (e.g. « à␣l'école. » is rewritten as « à␣l'␣école␣. ») and grouping others («100 000 » is rewritten as « 100000 »).

NAÏVE APPROACH
CO-OCCURRENCE TABLE

- The extraction of a lexicon from this parallel corpus relies on the construction of a *co-occurrence table*.
- This table can be modeled by a python dictionary which associates to a French word (the 1st key), a second python dictionary whose keys are the set of English words co-occurring with the French word and whose values are the number of sentence pairs in which the French word and the English word both appear.

NAÏVE APPROACH
CO-OCCURRENCE TABLE

- For example, if the corpus consists of two sentences, this is what your python dictionary should look like :

doc nº 1	doc nº 2
la vache et le veau	the cow and the calf
le chien et le chat	the dog and the cat

```
cooc_table = {'chat': {'and': 1, 'the': 1, 'dog': 1, 'cat': 1},
              'chien': {'and': 1, 'the': 1, 'dog': 1, 'cat': 1},
              'et': {'and': 2, 'calf': 1, 'cat': 1, 'cow': 1, 'dog': 1, 'the': 2},
              'la': {'and': 1, 'the': 1, 'cow': 1, 'calf': 1},
              'le': {'and': 2, 'calf': 1, 'cat': 1, 'cow': 1, 'dog': 1, 'the': 2},
              'vache': {'and': 1, 'calf': 1, 'cow': 1, 'the': 1},
              'veau': {'and': 1, 'the': 1, 'cow': 1, 'calf': 1}}
```

INSTRUCTIONS

- Write a method **build_cooc_table(filepath_fr, filepath_en)** that takes two filenames describing a parallel corpus and returns the co-occurrence table of this corpus.
- Beware: we want to determine the *number of sentences* in which a word appears, and even if a word appears several times in the same sentence, it must be counted only once.
- => Remember to remove duplicates within a sentence before performing the counts.

INSTRUCTIONS

- Sort the cooc table by decreasing frequency and print it in an external file (1 pair of words per line).
- To sort the table, you will first have to transform it into a list of tuples (French word, English word, frequency).
- Finally, interpret the result obtained: does this method allow you to build a "good" lexicon?