# Assignment 2: Lexical semantics (graded part – version of October 25)

**General information**

- The graded part of Assignment 2 will earn you **30%** of your overall course grade. You will get a grade between 0 and 10.
- For the moment (October 26), there is only one exercise. More exercises will be added – we will keep updating this document.

**Exercise 1**. **To be done in pairs.** Deadline: **Wednesday November 3rd at 14:00**.

- Choose a language that you are reasonably proficient in and has a WordNet in NLTK. Choose a word that you like, for which the language's WordNet has at least 3 synsets for a given part of speech of the word. For instance: in Catalan, the verb *estudiar* (*to study*).
- Get 100 random samples of the word in context (around 50 words, or 1 paragraph) from a corpus of choice. Note: the corpus should be lemmatized and part-of-speech tagged. You can select it from NLTK or get it from an external source (for instance, the Universal Dependencies treebank, https://universaldependencies.org ).
- Each member of the pair **independently** annotates each sample with a sense (synset) from WordNet.
- Collate the results in a CSV file, with columns *name1*, *name2, sample*. For instance (invented data):

1) choose word 'estudiar' ('to study', in Catalan)
2) define that **s1** (sense 1) is *analyze.v.01;* **s2** *is study.v.03; etc.*
2) annotate data:

*gemma, lucas, sample*
*s1, s3, "blah estudiar blah blah..."*
*s2, s3, "blih blih estudies blih..."*
*s1, s1, "bloh bluh bleh estudiaré..."*
*…*

- Adapt the code we provide in file 'exercise1_assignment2_updated.py', that prints: the lemma of the word (and its translation into English); the number of synsets; for each synset, the synset name, its definition, and lemmas; and the contingency table of the annotations by each of you.
- Copy the resulting data to this document: https://docs.google.com/document/d/18HSH0TVJPAZj0qhtmNmy3v8x6TlAA5HbUtzQ4T7dK9I/edit?usp=sharing
- Upload your code and CSV file onto Aula Global, in a compressed file.