**Learner Corpus Research Summer School**

**Workshop: Automatic annotation of learner corpus data**

July 4th, 2023, Louvain-la-Neuve

ACTIVITY 1. POS-tagging

1. Inspect the texts in the folder ‘icle\_texts’.

***Question 1. What do you notice about them? What pre-processing steps (if any) might be necessary before using automatic annotation tools?***

1. Open the file *icle.xlsx*.
2. Use the SUBSTITUTE() function to remove all instances of <\* and <\*\* from the texts.

*Note: You may need to do this in multiple steps.*

1. Open the [CLAWS tagger](http://ucrel-api.lancaster.ac.uk/claws/free.html).
2. Copy the cleaned text for file BRFF1065.txt into the input box.
3. Select the options for the C5 tagset and horizontal output style.
4. Click ‘Tag text now’

***Question 2: How many tokens have been tagged?***

***Question 3. What is the tag for the base form of a lexical verb in the C5 tagset?***

1. Return to the previous page and tag the text using the C7 tagset.

***Question 4. What is the tag for the base form of a lexical verb in the C7 tagset?***

1. Return to the previous page and select the vertical output style.

***Question 5. Look at the annotation for the word ‘lives’. What do you think is meant by ‘[VVZ/86] NN2/14’?***

1. Now open the online version of [TreeTagger](https://cental.uclouvain.be/treetagger/)
2. Copy the cleaned text for file BRFF1065.txt into the input box.
3. Select ‘English’ as the language of the text.
4. Click on ‘Process’.
5. Download the resulting output file and open it in Excel.

***Question 6. What does the code ‘@card@’ mean?***

1. Count the number of adjectives in this text. In a blank cell, type the following formula: =COUNTIF(B2:B576, “JJ”).

***Question 7.How many common nouns are used in this text?***

*Hint:* You can find the description of the tagset [here](https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/Penn-Treebank-Tagset.pdf).