Exercise: Text Analysis with TXM

(Designed by Ulrike Henny, translated by Christof Schöch, 2016)

1. Load a corpus

From the USB drive, please select one of the following tagged corpora:

- doyle.txm: 12 novels by Arthur Conan Doyle (Christof Schöch).
- powiesc.txm: 75 Polish nineteenth and early twentieth-century novels (Jan Rybicki).
- roman19.txm: 36 French nineteenth-century novels (Christof Schöch & Stefanie Popp).
- nohispam.txm: 24 hispanoamerican nineteenth-century novels (Ulrike Henny).

Load the corpus (File > Load) and use it for the following exercises. Remember that each language has its own POS tagset.

2. Word forms, lemmata and part-of-speech

- Which are the 5 most frequent and 5 least frequent word forms in the corpus?
- Which are the 3 most frequent and 3 least frequent part of speech in the corpus?
- How many tokens belonging to the part of speech „verb“ are there in the corpus?
- How many times does the word form „street“ (or for the Polish corpus: „ulica“ / for the French corpus: „rue“ / for the Spanish corpus: „calle“) appear?
- Please create an index for the lemma „street“ (or: „ulica“ / „rue“ / „calle“)
  - How many tokens are there?
  - How many different word forms are present?
- Search for word forms which contain the string „street“ (or: „ulica“ / „rue“ / „calle“). Ask TXM to show you not just the word form, but also the lemma.
- In which text(s) does the word form „side-street“ (or: „aleja“ / „ruelle“ / „callejero“) appear?

3. Advanced searches

- Search for the lemma „work“ (or: „pracować” / „travailler“ / „trabajar“) followed by one or several adverbs.
- Search for the verbs which appear in a distance of 0 to 3 words to the lemma „work“ (or: „pracować” /„travailler“ / „trabajar“).
- Search for adverbs which are preceded by the lemma „live” or „die” (or: „żyć” or „umrzeć” / „vivre” or „mourir” / „vivir” or „morir”).

- Search for words which contain a sequence of 4 vowels (for Spanish or English) respectively a sequence of 5 vowels (for French and Polish). How many are there in your corpus?

4. Concordances and Cooccurrences

- Create a concordance for the word form „Londoners“ (or: „krakowiak“ / „parisienne“ / „mexicana“). Have a look at one result in its immediate and in its wider context.
- Create a cooccurrence analysis for the word „city“ (or: „miasto“ / „ville“ / „ciudad“). Which other noun has the highest cooccurrence score? Which other adjective? Repeat the same for the word „country“ (or: „kraj“ / „campagne“ / „campo“). Compare the resulting top nouns and top adjectives for the two queries.

5. Comparative queries

Progression

- Create a progression for the two lemmata „joy“ and „fear“ (or: „radość” and „strach” / „joie“ and „peur“ / „alegría and „miedo“). Find out how to visualize two queries at a time. What do you observe with regard to the relation between the two lemmata? Which authors have a particular pattern of usage of the two terms?
- Create a similar progression for interjections. (Find out which POS-tag is used for interjections in the language you are using.)

Partions and Specificities

- Partition your corpus according to decades (or subgenres, if they are available).
    - Find out in which decades which of the terms „city“ and „country“ (or: „miasto“ and „kraj“ / „ville“ and „campagne“ / „ciudad“ and „campo“) is more frequent than the other.
    - Find out in which decade interjections where used the most. Why is this particularly difficult?
- Partition your corpus according to the novels' authors.
    - Select two authors and find out which lemmata are typical, i.e. overrepresented or specific for them using the Specificity function.
    - Find out which part-of-speech are typical, i.e. overrepresented or specific to these two authors, respectively.