**Topic 1**

**Introduction**

**1. Types of corpora**

* Monolingual or multilingual?
* For multilingual: comparable or parallel?
* Native speakers, L2 speakers or language learners?
* General or language for specific purposes?
* Spoken or written?

**Exercise**

Which corpora do you know or have already used? How would you describe them?

**2. Main concepts**

**token:** a string of alphanumeric characters between two spaces

**type:** unique token

**Type/Token Ratio:** the number of types divided by the number of tokens

**lemma:** the dictionary form of a word (e.g. go for *went*, book for books, etc.)

**wordform:** the form in which a word appears in the text (e.g. *went*, *books*, *nicer*)

**POS tag:** part-of-speech tag (e.g. NOUN, VERB, ADJ, PROPN, NN1)

**frequency list:** a list of types (lemmas or wordforms) with their frequencies in a corpus

**dependency:** the relationship between a syntactic head and its dependent element, e.g. noun and determiner, predicate and object, adjective and adverbial modifier, etc.

**n-gram:** n words that appear together in a text, e.g. *I went, went to, to Paris* (2-grams); *I went to*, *went to Paris* (3-grams); *I went to Paris* (4-gram)

**key words:** words that are distinctive of a particular register, text type, author, etc. These are words that are more frequent in a corpus in comparison with a reference corpus (‘the default’) than one would expect by chance (e.g. *verb*, *phoneme*, *accusative* in a corpus of linguistic publications).

**collocation:** two or more words that occur together more frequently than one would expect based on their individual frequencies, e.g. *Prime Minister*, *America first*, *civil war*.

**3. Levels of annotation**

- Raw text (zero level)

- Identification of sentences.

- Tokenization – segmentation of sentences into tokens (words or sentences), separation of punctuation marks from words

- Lemmata/lemmas

- Parts of Speech (e.g. noun, adjective, preposition). Depends on the POS tagset.

- Morphological tags (e.g. for nouns: singular or plural, for verbs: tense, mood, person, etc.)

- Syntactic roles and dependencies (e.g. a predicate and its object).

- Prosody, pauses, overlap, etc.

- Semantic and pragmatic annotation (rarely)

- Additional information (alignment, timing)

**Exercise**

Have a look at the corpora samples on GitHub. Which types of annotation do you see?

**4. Important corpora**

Need to pay to use offline:

* COCA, COHA and other corpora by Mark Davies

Freely downloadable or upon request (with registration):

* Universal Dependencies
* British National Corpus
* Santa Barbara Spoken Corpus of American English
* COW corpora (Web-based, in different languages)
* ICE family corpora (many varieties of English)
* Leipzig Corpora Collection
* MultiCast (but difficult to restore the text)
* OPUS corpus