Concordances and frequency lists for your words

Serge Sharoff

Serge Sharoff

The purpose of this session is to give you a feel of several online interfaces. You are not expected to analyse the concordance lines at this stage. The **learning objective** is to gain confidence in navigating in each of the interfaces and in obtaining results for your queries.

1 Interfaces

The following interfaces will be introduced in this session:

1. SketchEngine (SE): `https://app.sketchengine.eu/`
2. Aranea: `http://aranea.juls.savba.sk`
3. BYU interface: `https://www.english-corpora.org/`
4. **Parallel:** Opus: `http://opus.nlpl.eu/`
5. **Parallel:** Linguee: `http://www.linguee.com/`

Keep them open in separate browser windows.

2 Corpora

The exercises today are for English. Please use the following corpora (available through some of the online interfaces):

1. BNC, the British National Corpus (Leeds CQP, IntelliText, SE, BYU);
2. ukWac, snapshot of `.uk` TLD (Leeds CQP, IntelliText, SE);
3. Europarl (IntelliText, SE, Opus);
4. a collection of parallel texts from Linguee

3 Exercise 1: Simple queries

Select a corpus and make a **concordance** for:
ill-gotten
gotten
espouse

Find out the frequencies of such words in each of the corpora in each of the interfaces (if they provide this info).

4 Exercise 2: Morphological information

Check if you understand the following terminology:

A **paradigm** is a set of all possible **word forms**, which share the same lexical meaning but differ morphologically:

be *am, are, is, was, were, been, being;*

tall *taller; tallest* (comparative adjective)

play *plays; played; playing* (verb) or *play; plays* (noun)

A **lemma** (or a "**base form**") is a "canonical" "dictionary" form, in bold in these lists.

Part of Speech (PoS) is the functional class of a word form in a sentence, usually expressed by a code (*noun*=NN; plural noun=NNS; adjective=JJ; base verb=VV; verb with the *ing* ending=VVG).

in **SketchEngine/Aranea** choose the appropriate query field;
in **Opus** replace the keyword `word` with the keyword `lem`.

1. Take a look at the outputs of lemma vs word form queries for:
   go, get, impinge, tall
   Note: to find examples of relatively rare forms, e.g., *gotten* or *tallest*, you need to search for them separately.

2. Find examples of:
   utter (VERB/ADJECTIVE)
   water (NOUN/VERB)
   waters (NOUN/VERB)

The meanings of many words vary depending on their functional classes (PoS tags). The easiest option is with the Query builder in IntelliText. As for the other interfaces, we will explore the magic of regular expressions later.

5 Exercise 3: Phrases and lexical patterns

Find differences in the frequencies of:
naughty little vs *little naughty*
as if it were vs. *as if it was*
translate into vs. *translate to*

6 Exercise 4: Lexicogrammatical patterns

Some verbs in English are followed by other verbs in the *ing* form, while some are followed by verbs in the form with *to*. Detect which are which on the basis of corpus data:

arrange, mean, plan, refuse, manage, postpone

7 Exercise 5: Parallel corpora

Use EuroParl in IntelliText, SE, OPUS if you deal with an EU language (choose an appropriate language code under *Alignments* in Opus; In IntelliText the parallel corpora are under the special rubric in the Language Selection menu). Use the UN corpus for Arabic, Chinese and Russian. For Japanese, use OpenSubtitles in the OPUS interface. Linguee offers large parallel corpora of unspecified origin.

Investigate translations from English into your languages for identical words with different functional classes, e.g.
utter (VERB/ADJECTIVE)
water (NOUN/VERB)
waters (NOUN/VERB)

In the OPUS interface use `lem` for lemmas and `tnt` for PoS (the names are listed under the positional annotation field).