**Introduction to Corpus Linguistics**

WiSe 2018-2019

by Dr. Natalia Levshina ©

**Session 4. Words in context**

Please don’t forget: next time (4December) there is no class.

The goals of our today’s session is to learn

a) how to find words (wordforms or lemmata) with particular properties in a UD corpus, parsed with the help of udpipe;

b) how to extract the sentences where these words (or features) occur;

c) how to extract sequences of words (or other features).

We’ll be using a corpus of Donal Trumps’ recent tweets downloaded from <http://www.trumptwitterarchive.com/archive>

The corpus is available from GitHub. Its name is trump.txt. Every line is a tweet. This is what the first lines look like.

RT @GOP: “Tomorrow the voters of this state will cast their ballots in one of the most important Senate elections of your lives—of all of…

RT @GOP: “Only with a strong Senate GOP majority can we defend your tax cuts defend your Second Amendment protect your Medicare and Socia…

RT @GOP: “I need the great people of Mississippi to send a message to… the radical Democrats by electing @cindyhydesmith.” -@realDonaldTrum…

....starts today election is on December 4th. @VoteBradRaff is tough on Crime and Borders Loves our Military and Vets. He will be great for jobs!

Brad Raffensperger will be a fantastic Secretary of State for Georgia - will work closely with @BrianKempGA. It is really important that you get out and vote for Brad - early voting....

Save the corpus locally.

1. Open the corpus in R and parse it with the help of *udpipe*.

**trump <- scan(file = file.choose(), what = "character", sep = "\n", comment.char = "#", encoding = "UTF-8")**

**trump[11]**

[1] "Mexico should move the flag waving Migrants many of whom are stone cold criminals back to their countries. Do it by plane do it by bus do it anyway you want but they are NOT coming into the U.S.A. We will close the Border permanently if need be. Congress fund the WALL!"

**library(udpipe)**

#If you have already downloaded the language model:

**ud\_eng <- udpipe\_load\_model("english-ud-2.0-170801.udpipe")**

#If you don’t have the file English-ud-2.0-170801.udpipe in your current working R directory, check Session 2 and 3 for R code.

Now we’ll parse the corpus. This might take a while:

**trump\_ud <- udpipe(x = trump, object = ud\_eng)**

**head(trump\_ud)**

**dim(trump\_ud)**

[1] 115502 17

The rows are tokens (words, punctuation marks, etc.). What are all these columns?

**colnames(trump\_ud)**

[1] "doc\_id" "paragraph\_id" "sentence\_id"

[4] "sentence" "start" "end"

[7] "term\_id" "token\_id" "token"

[10] "lemma" "upos" "xpos"

[13] "feats" "head\_token\_id" "dep\_rel"

[16] "deps" "misc"

The following columns are important for us today:

* sentence – the sentence text
* token – the word form (or punctuation mark)
* lemma
* upos – part of speech
* dep\_rel – syntactic dependency (subject, object, etc.)

For example, let’s take the 100th word:

**trump\_ud$sentence[100]**

[1] "@VoteBradRaff is tough on Crime and Borders Loves our Military and Vets."

**trump\_ud$token[100]**

[1] "tough"

**trump\_ud$lemma[100]**

[1] "tough"

**trump\_ud$upos[100]**

[1] "ADJ"

**trump\_ud$dep\_rel[100]**

[1] "root"

**2. How to search for wordforms and lemmata**

Let’s search for the word ‘America’ by tokens (wordforms).

**trump\_ud[trump\_ud$token == "America",]**

[output omitted]

That’s too much information. Let us select only the sentences:

**trump\_ud$sentence[trump\_ud$token == "America"]**

**…**

**length(trump\_ud$sentence[trump\_ud$token == "America"])**

[1] 143

There are 143 occurrences of this wordform.

**Exercise**

Find the sentences with the words China and Russia.

Sometimes it makes more sense to search by lemmata – in case if there are different morphological forms, e.g. immigrant and immigrants.

**trump\_ud$sentence[trump\_ud$lemma == "migrant"]**

1] "Migrants at the Southern Border will not be allowed into the United States until their claims are individually approved in court."

[2] "The Mayor of Tijuana Mexico just stated that “the City is ill-prepared to handle this many migrants the backlog could last 6 months.”"

[3] NA

[4] "The migrant ‘caravan’ that is openly defying our border shows how weak &amp; ineffective U.S. immigration laws are."

[5] "Muslim migrant beats up Dutch boy on crutches!"

**Exercise**

Find the sentences with the word *criminal*.

We can search for synonyms or other groups of words:

**trump\_ud$sentence[trump\_ud$lemma %in% c("immigrant", "migrant", "alien")]**

…

**Exercise**

Find the sentences with the words *criminal* and *illegal*.

Let us now combine the information from different columns. For example, lemma and part of speech (upos). But what parts of speech do we have there? To check that, we can use the following code:

**summary(factor(trump\_ud$upos))**

ADJ ADP ADV AUX CCONJ DET INTJ NOUN NUM

8041 11004 5930 7160 3732 8253 270 17784 1336

PART PRON PROPN PUNCT SCONJ SYM VERB X

2578 8759 14644 11422 1556 465 12010 558

The word *criminal* can be a noun and an adjective. Let us find the contexts with each:

**trump\_ud$sentence[trump\_ud$lemma == "criminal"&trump\_ud$upos == "ADJ"]**

**…**

**trump\_ud$sentence[trump\_ud$lemma == "criminal"&trump\_ud$upos == "NOUN"]**

**…**

**Exercise**

Find the sentences with the word *right* in different functions. Which part of speech prevails?

Have you ever wondered what kind of adjectives Trump likes to use? Let us find out:

**trump\_ud$lemma[trump\_ud$upos == "ADJ"]**

…

#Save it as a list

**trump\_adj <- trump\_ud$lemma[trump\_ud$upos == "ADJ"]**

#Compute the frequencies:

**trump\_adj\_freq <- table(trump\_adj)**

#sort and take top 20

**sort(trump\_adj\_freq, decreasing = TRUE)[1:20]**

trump\_adj

great many big good more

692 256 240 180 163

american strong other bad new

130 125 116 109 104

total better illegal last first

99 90 83 83 70

much wonderful Rt special fake

69 61 60 60 57

**3. How to search for sequences**

Let us search for "fake news" by lemmas. First, we will extract all indices of the words that are ‘news’.

**trump\_news\_id <- which(trump\_ud$lemma == "news")**

**trump\_news\_id**

…

For example:

**trump\_ud$sentence[115072]**

[1] "Great and we should boycott Fake News CNN."

Next, we need to extract the preceding words:

**trump\_news\_prev <- trump\_news\_id - 1**

**before\_news <- trump\_ud[trump\_news\_prev,]**

**fake\_news <- before\_news[before\_news$lemma == "fake",]**

**dim(fake\_news)**

[1] 147 17

**Exercise**

Check the noun *immigrant*: which are the preceding adjectives? How many times are they preceded by the word *illegal*? What is the proportion with regard to the total number of their occurrences?