**Topic 2**

**First Steps with R and Universal Dependencies**

April 16 2019

Goals:

1. To learn the basic types of objects and commands in R, using RStudio.

2. To learn to annotate sentences using the Universal Dependencies package udpipe.

See the file *Intro\_R*.

**Annotating your data in R with udpipe**

**1. How to install and load the package udpipe**

Step 1. In RStudio, go to *Packages* and press *Install.* Type in *udpipe*. Make sure the Install dependencies box is checked and press the Install button. The package will be installed. **You need to do this only once!** (unless you change the version of R)

Step 2. Every time you start a new session in R, you need to load the packages you want to work with. To load *udpipe*, you need to type in the following:

**library(udpipe)**

**2. How to annotate a sentence in R**

To annotate a sentence in R, you need two things:

a) the sentence:

**text <- "This is my sentence."**

b) the language model:

For use it for the first time, you need to download the model:

**ud\_eng <- udpipe\_download\_model(language = "english")**

**ud\_eng <- udpipe\_load\_model(ud\_eng)**

If you have already downloaded it, there is a more efficient way:

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

**#change the name of the model if you work with another language!**

Now you are ready to do the parsing:

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

To see the output, you can simply type the name of the new object:

**text\_ud**

# doc\_id paragraph\_id sentence\_id sentence start end te#rm\_id token\_id token lemma upos xpos

#1 doc1 1 1 This is my sentence. 1 4 # 1 1 This this PRON DT

#2 doc1 1 1 This is my sentence. 6 7 # 2 2 is be AUX VBZ

#3 doc1 1 1 This is my sentence. 9 10 # 3 3 my my PRON PRP$

#4 doc1 1 1 This is my sentence. 12 19 # 4 4 sentence sentence NOUN NN

#5 doc1 1 1 This is my sentence. 20 20 # 5 5 . . PUNCT .

# feats head\_token\_#id dep\_rel deps misc

#1 Number=Sing|PronType=Dem # 4 nsubj <NA> <NA>

#2 Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin # 4 cop <NA> <NA>

#3 Number=Sing|Person=1|Poss=Yes|PronType=Prs # 4 nmod:poss <NA> <NA>

#4 Number=Sing # 0 root <NA> SpaceAfter=No

#5 <NA> # 4 punct <NA> SpacesAfter=\\n

This looks messy. We can look at the structure of this object:

**str(text\_ud)**

#'data.frame': 5 obs. of 17 variables:

# $ doc\_id : chr "doc1" "doc1" "doc1" "doc1" ...

# $ paragraph\_id : int 1 1 1 1 1

# $ sentence\_id : int 1 1 1 1 1

# $ sentence : chr "This is my sentence." "This is my sentence.#" "This is my sentence." "This is my sentence." ...

# $ start : int 1 6 9 12 20

# $ end : int 4 7 10 19 20

# $ term\_id : int 1 2 3 4 5

# $ token\_id : chr "1" "2" "3" "4" ...

# $ token : chr "This" "is" "my" "sentence" ...

# $ lemma : chr "this" "be" "my" "sentence" ...

# $ upos : chr "PRON" "AUX" "PRON" "NOUN" ...

# $ xpos : chr "DT" "VBZ" "PRP$" "NN" ...

# $ feats : chr "Number=Sing|PronType=Dem" "Mood=Ind|Number=#Sing|Person=3|Tense=Pres|VerbForm=Fin" "Number=Sing|Person=1|Poss=Y#es|PronType=Prs" "Number=Sing" ...

# $ head\_token\_id: chr "4" "4" "4" "0" ...

# $ dep\_rel : chr "nsubj" "cop" "nmod:poss" "root" ...

# $ deps : chr NA NA NA NA ...

# $ misc : chr NA NA NA "SpaceAfter=No" ...

This is a data frame. The most important linguistic information is in the following columns:

* token\_id – the ID of the word (from 1 to …)
* token – the wordform
* lemma
* upos – the part of speech from the Universal Dependencies tagset
* feats – morphological features
* head\_token\_id – the ID of the syntactic head
* dep\_rel – the syntactic function (dependency)

The full list of POS tags can be found here:

<http://universaldependencies.org/u/pos/all.html>

The full list of syntactic dependencies can be found here:

<http://universaldependencies.org/u/dep/all.html>

It may be helpful to select only the relevant columns:

**text\_ud[, 8:15]**

# token\_id token lemma upos xpos feats

#1 1 This this PRON DT Number=Sing|PronType=Dem

#2 2 is be AUX VBZ Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin

#3 3 my my PRON PRP$ Number=Sing|Person=1|Poss=Yes|PronType=Prs

#4 4 sentence sentence NOUN NN Number=Sing

#5 5 . . PUNCT . <NA>

# head\_token\_id dep\_rel

#1 4 nsubj

#2 4 cop

#3 4 nmod:poss

#4 0 root

#5 4 punct

**Exercise**

Annotate the following sentences and answer the following questions:

1) May the force be with you.

What is the POS and syntactic dependency of *may*?

2) I'm going to make him an offer he can't refuse.

What is the syntactic dependency and head of *refuse*? Can you explain why?

3) There's no place like home.

What is the POS of *like* and *home*?

4) Nobody puts Baby in a corner.

What is the syntactic dependency of *corner*?

5) I want to break free.

What is the syntactic dependency of *break*?