LEXICAL COMPLEXITY AND PROFICIENCY

ROSSINA SOYAN

DECEMBER 7, 2021

DATA SCIENCE FOR LINGUISTS

BACKGROUND

- A summer of teaching intensive Russian at Middlebury college
- A transatlantic flight from New York to Moscow
  - Colleague: Have our students made progress?

MIDDLEBURY CORPUS OF L2 RUSSIAN TEXTS

- Essays written by students as part of a placement exam (pre-test) and final examination (post-test) in the summer of 2019
- 601 essays (103,150 words total) by 133 Russian L2 learners at different levels of proficiency

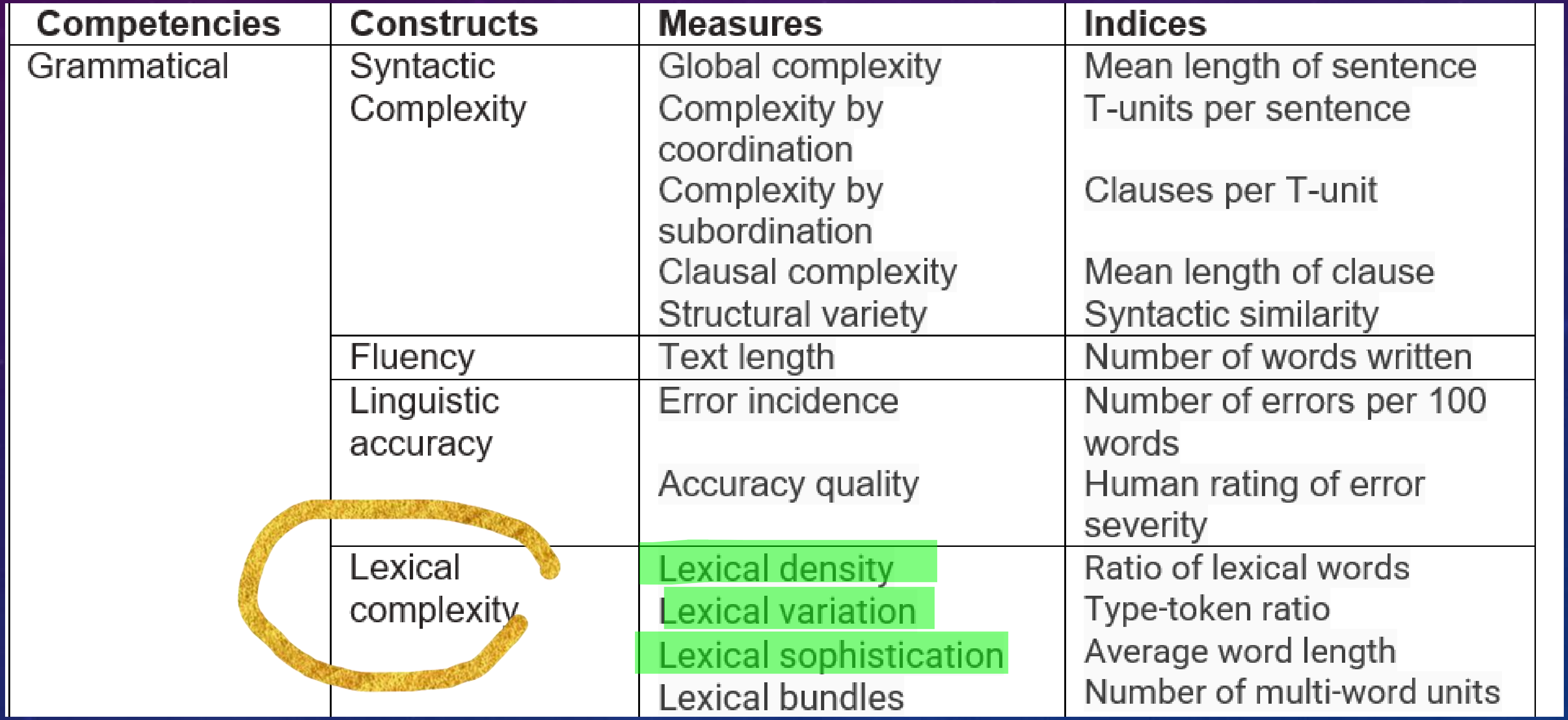
WHAT IS WRITING ABILITY?

- Narrow understanding: organizational knowledge (grammatical knowledge + discourse aspects)
- A wider framework: Writing Competence Model (Connor & Mbaye, 2002; Barkaoui & Hadidi, 2020)

WRITING COMPETENCE MODEL

- Competencies:
  - Grammatical
  - Discourse
  - Sociolinguistic
  - Strategic
  - Content and Source Use

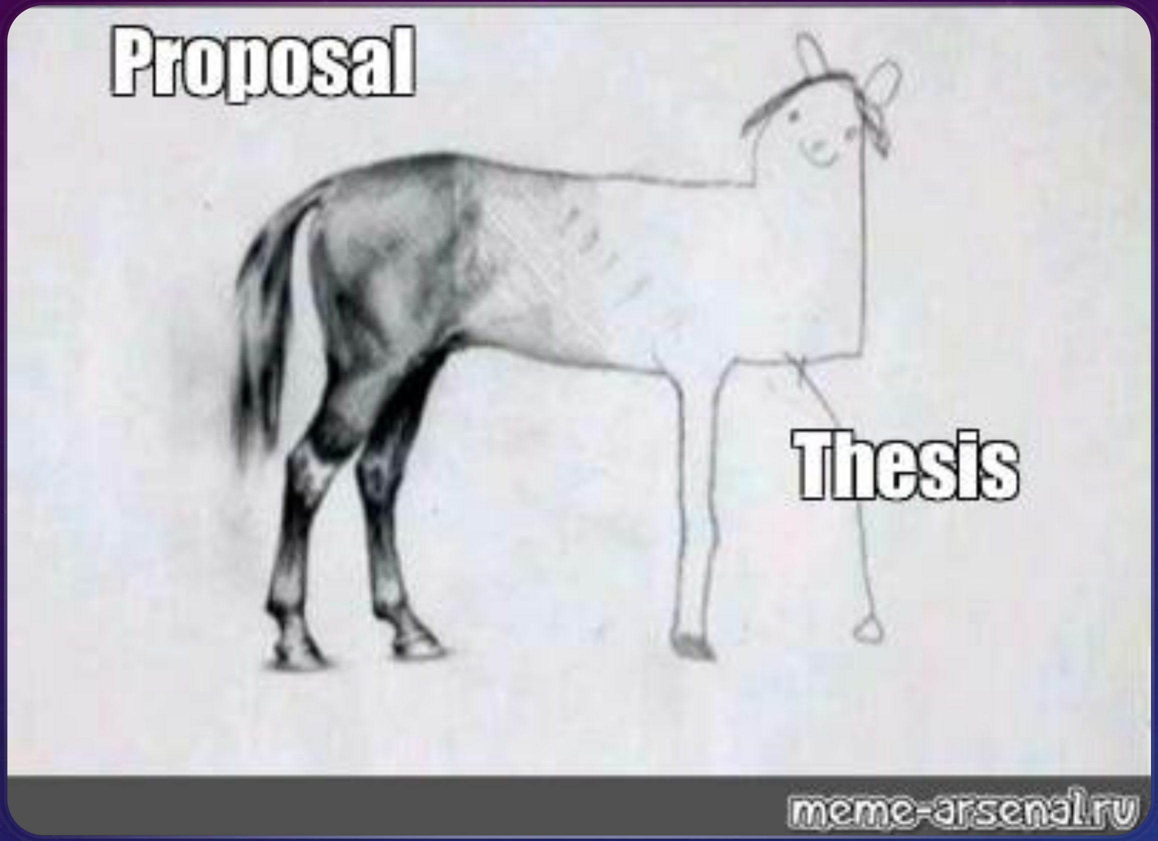
GRAMMATICAL COMPETENCY



Competencies	Constructs	Measures	Indices
Grammatical	Syntactic Complexity	Global complexity	Mean length of sentence
		Complexity by coordination	T-units per sentence
		Complexity by subordination	Clauses per T-unit
		Clausal complexity	Mean length of clause
		Structural variety	Syntactic similarity
	Fluency	Text length	Number of words written
	Linguistic accuracy	Error incidence	Number of errors per 100 words
		Accuracy quality	Human rating of error severity
	Lexical complexity	Lexical density	Ratio of lexical words
		Lexical variation	Type-token ratio
		Lexical sophistication	Average word length
		Lexical bundles	Number of multi-word units

Source: Barkaoui & Hadidi, 2020, p. 10

RESEARCH QUESTION



- What lexical complexity measures correspond to intermediate and advanced proficiency levels in L2 Russian texts?

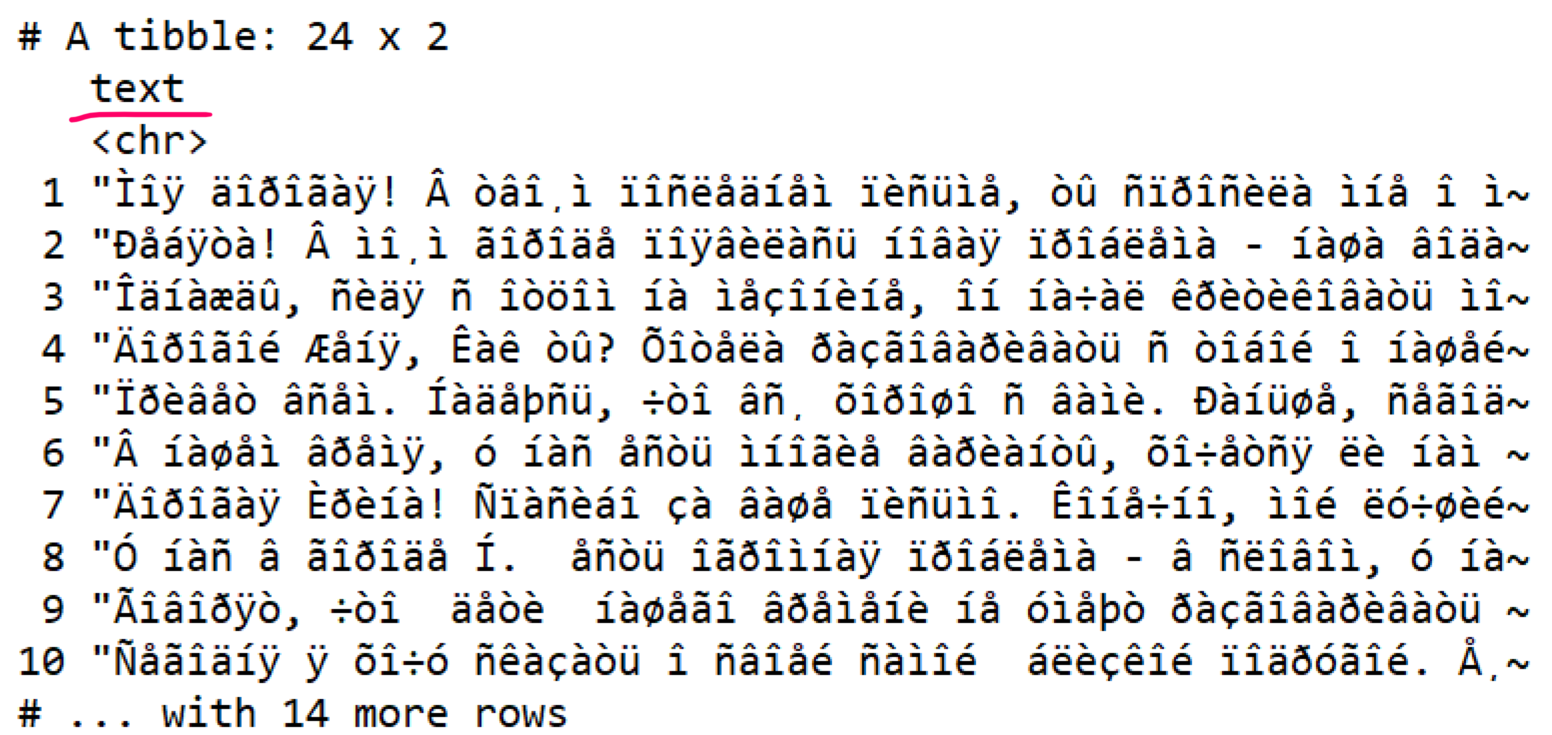
MY ATTEMPT TO ANSWER THE RQ

- Load to R essays written by 8 students (4 intermediate and 4 advanced students). Each student submitted 3 essays. The corpus for analysis: 24 texts
- Calculate lexical diversity, lexical variation, and lexical sophistication for the essays written by each student
- Conduct a hierarchical cluster analysis
- Interpret the results

WHAT I LEARNED

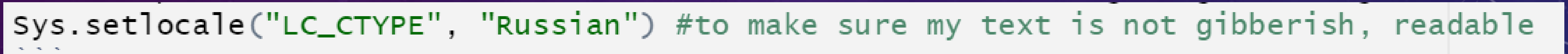
- Loading texts in any language other than English is hard.
- Tokenization of non-English texts may contain serious errors.
- Interpreting your findings after turning words into numbers is the hardest.

DO YOU SPEAK GIBBERISH?

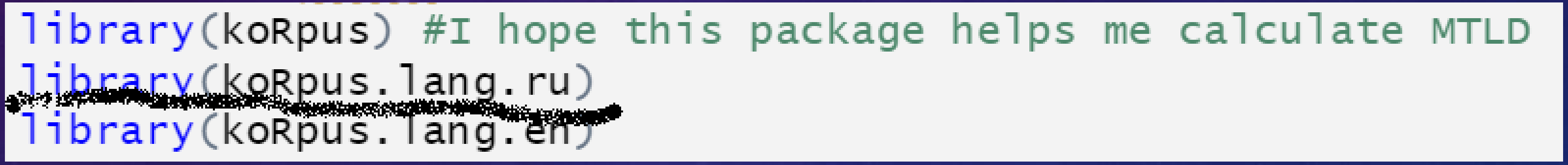


```
# A tibble: 24 x 2
   text
   <chr>
 1 "Ìîÿ äîðîãàÿ! Â òâî¸ì ïîñëåäíåì ïèñüìå, òû ñïðîñèëà ìíå î ì~
 2 "Ðåáÿòà! Â ìî¸ì ãîðîäå ïîÿâèëàñü íîâàÿ ïðîáëåìà - íàøà âîäà~
 3 "Îäíàæäû, ñèäÿ ñ îòöîì íà ìåçîíèíå, îí íà÷àë êðèòèêîâàòü ìî~
 4 "Äîðîãîé Æåíÿ, Êàê òû? Õîòåëà ðàçãîâàðèâàòü ñ òîáîé î íàøåé~
 5 "Ïðèâåò âñåì. Íàäåþñü, ÷òî âñ¸ õîðîøî ñ âàìè. Ðàíüøå, ñåãîä~
 6 "Â íàøåì âðåìÿ, ó íàñ åñòü ìíîãèå âàðèàíòû, õî÷åòñÿ ëè íàì ~
 7 "Äîðîãàÿ Èðèíà! Ñïàñèáî çà âàøå ïèñüìî. Êîíå÷íî, ìîé ëó÷øèé~
 8 "Ó íàñ â ãîðîäå Í.  åñòü îãðîìíàÿ ïðîáëåìà - â ñëîâîì, ó íà~
 9 "Ãîâîðÿò, ÷òî  äåòè  íàøåãî âðåìåíè íå óìåþò ðàçãîâàðèâàòü ~
10 "Ñåãîäíÿ ÿ õî÷ó ñêàçàòü î ñâîåé ñàìîé  áëèçêîé ïîäðóãîé. Å¸~
# ... with 14 more rows
```

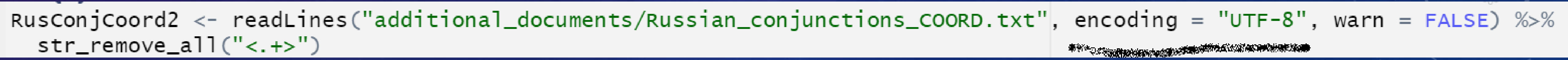
HOW TO SOLVE THE PROBLEM?



```
Sys.setlocale("LC_CTYPE", "Russian") #to make sure my text is not gibberish, readable
```

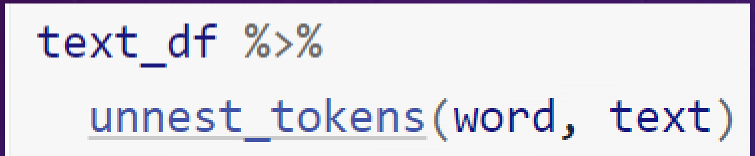


```
library(koRpus) #I hope this package helps me calculate MTLD
library(koRpus.lang.ru)
[redacted]
library(koRpus.lang.en)
```

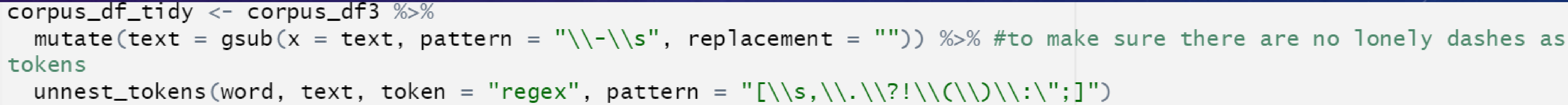


```
RusConjCoord2 <- readLines("additional_documents/Russian_conjunctions_COORD.txt", encoding = "UTF-8", warn = FALSE) %>%
  str_remove_all("<.+>")                                                                  [redacted]
```

UNNEST_TOKEN()

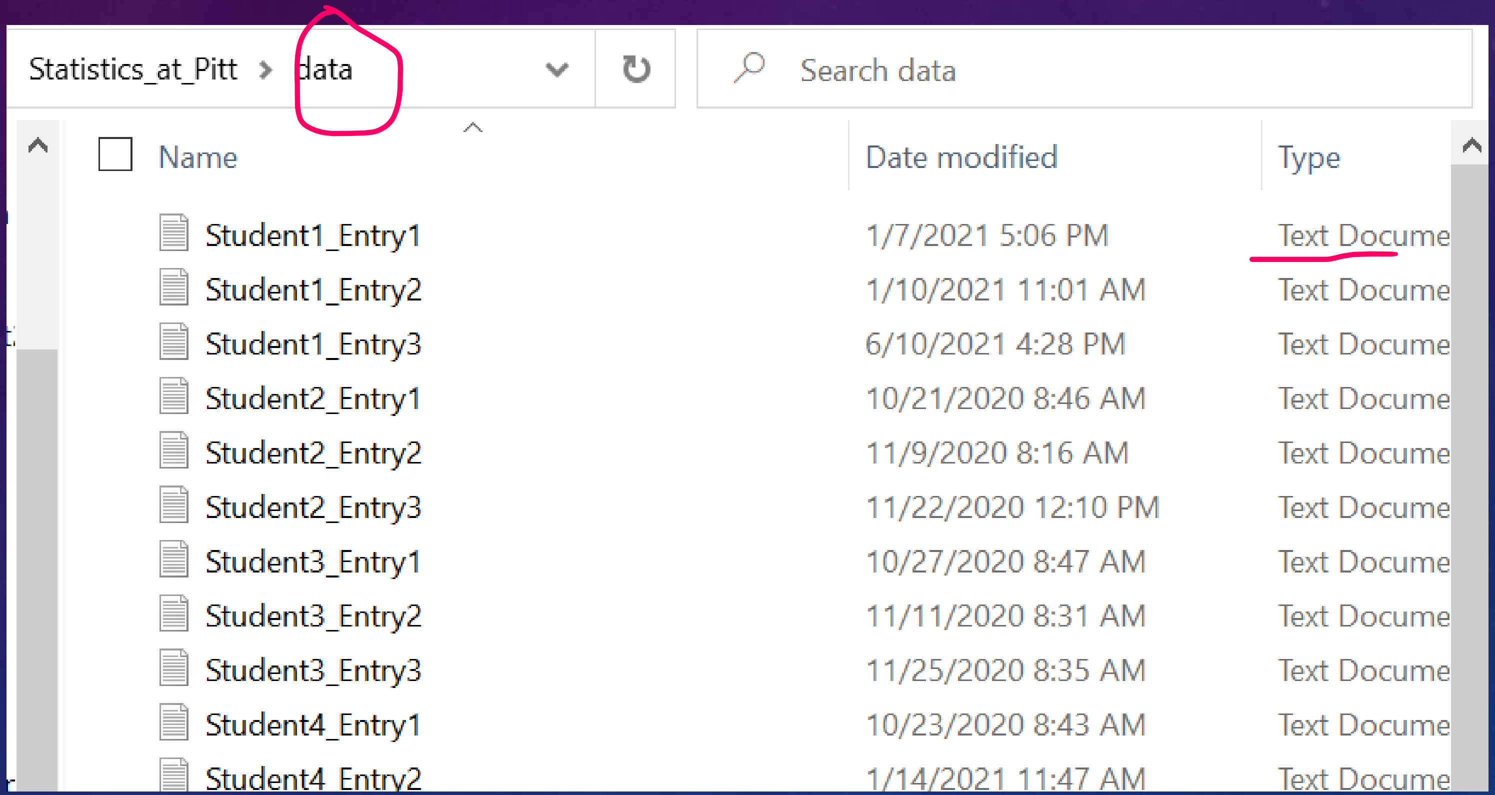


```
text_df %>%
  unnest_tokens(word, text)
```



```
corpus_df_tidy <- corpus_df3 %>%
  mutate(text = gsub(x = text, pattern = "\\-\\s", replacement = "")) %>% #to make sure there are no lonely dashes as
tokens
  unnest_tokens(word, text, token = "regex", pattern = "[\\s,\\.\\?!\\(\\)\\:\";]")
```

MY DATA

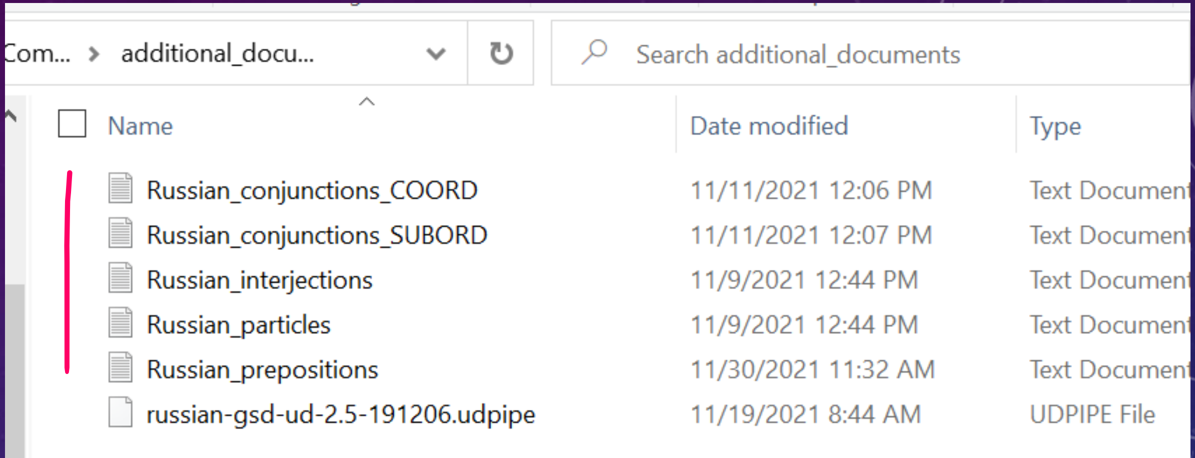


Name	Date modified	Type
Student1_Entry1	1/7/2021 5:06 PM	Text Docume
Student1_Entry2	1/10/2021 11:01 AM	Text Docume
Student1_Entry3	6/10/2021 4:28 PM	Text Docume
Student2_Entry1	10/21/2020 8:46 AM	Text Docume
Student2_Entry2	11/9/2020 8:16 AM	Text Docume
Student2_Entry3	11/22/2020 12:10 PM	Text Docume
Student3_Entry1	10/27/2020 8:47 AM	Text Docume
Student3_Entry2	11/11/2020 8:31 AM	Text Docume
Student3_Entry3	11/25/2020 8:35 AM	Text Docume
Student4_Entry1	10/23/2020 8:43 AM	Text Docume
Student4_Entry2	1/14/2021 11:47 AM	Text Docume

LEXICAL DENSITY

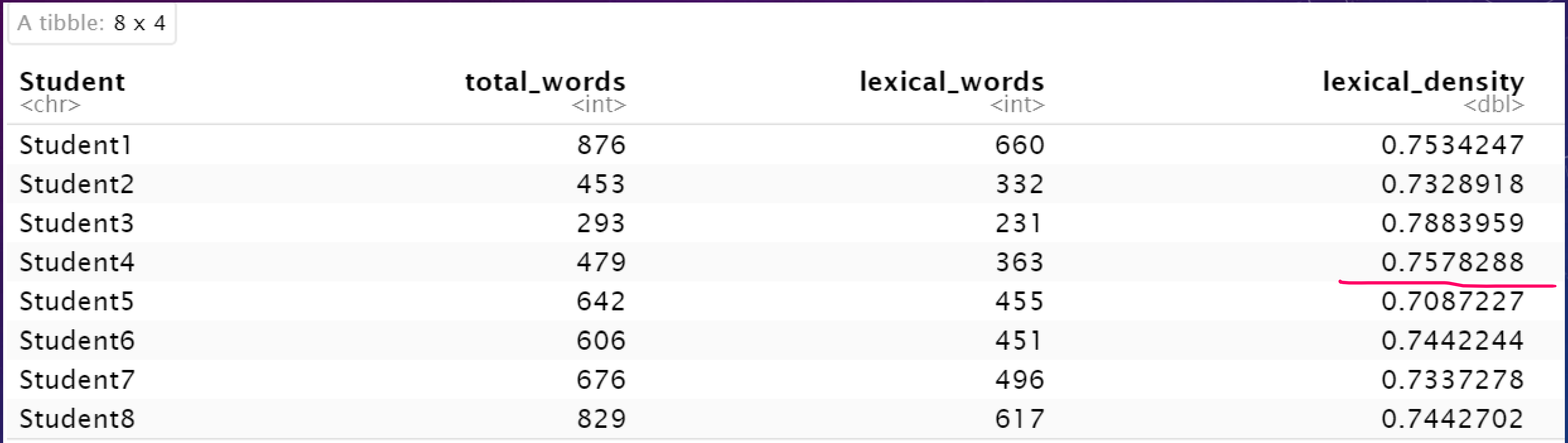
- The ratio of lexical words to the total number of words per essay (Bakaoui & Hadidi, 2020)
- Hypothesis: The higher the density, the higher the proficiency

LEXICAL DENSITY



Name	Date modified	Type
Russian_conjunctions_COORD	11/11/2021 12:06 PM	Text Document
Russian_conjunctions_SUBORD	11/11/2021 12:07 PM	Text Documen
Russian_interjections	11/9/2021 12:44 PM	Text Document
Russian_particles	11/9/2021 12:44 PM	Text Document
Russian_prepositions	11/30/2021 11:32 AM	Text Document
russian-gsd-ud-2.5-191206.udpipe	11/19/2021 8:44 AM	UDPIPE File

- I created .txt files with a possible list of non-lexical words in Russian
- I anti-joined non-lexical words and lexical words in each essay
- I divided the number of lexical words to the total number of words



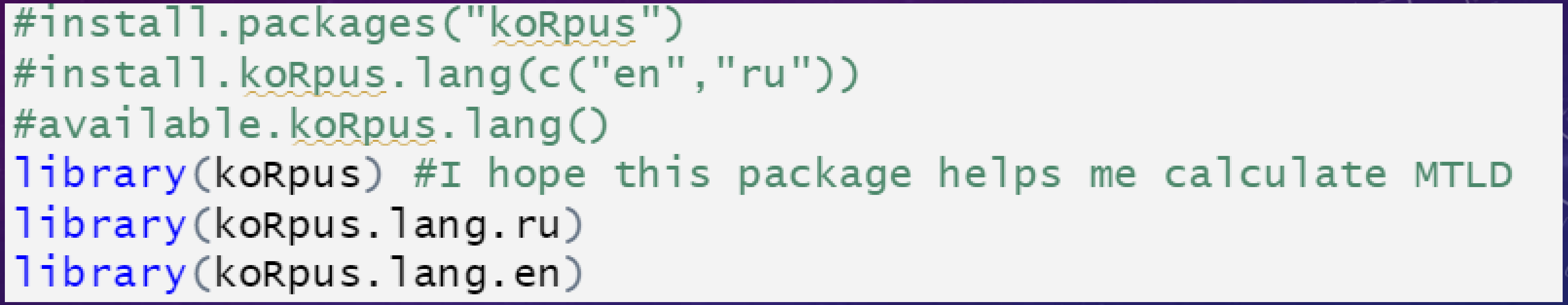
A tibble: 8 x 4

Student	total_words	lexical_words	lexical_density
<chr>	<int>	<int>	<dbl>
Student1	876	660	0.7534247
Student2	453	332	0.7328918
Student3	293	231	0.7883959
Student4	479	363	0.7578288
Student5	642	455	0.7087227
Student6	606	451	0.7442244
Student7	676	496	0.7337278
Student8	829	617	0.7442702

LEXICAL VARIATION (DIVERSITY)

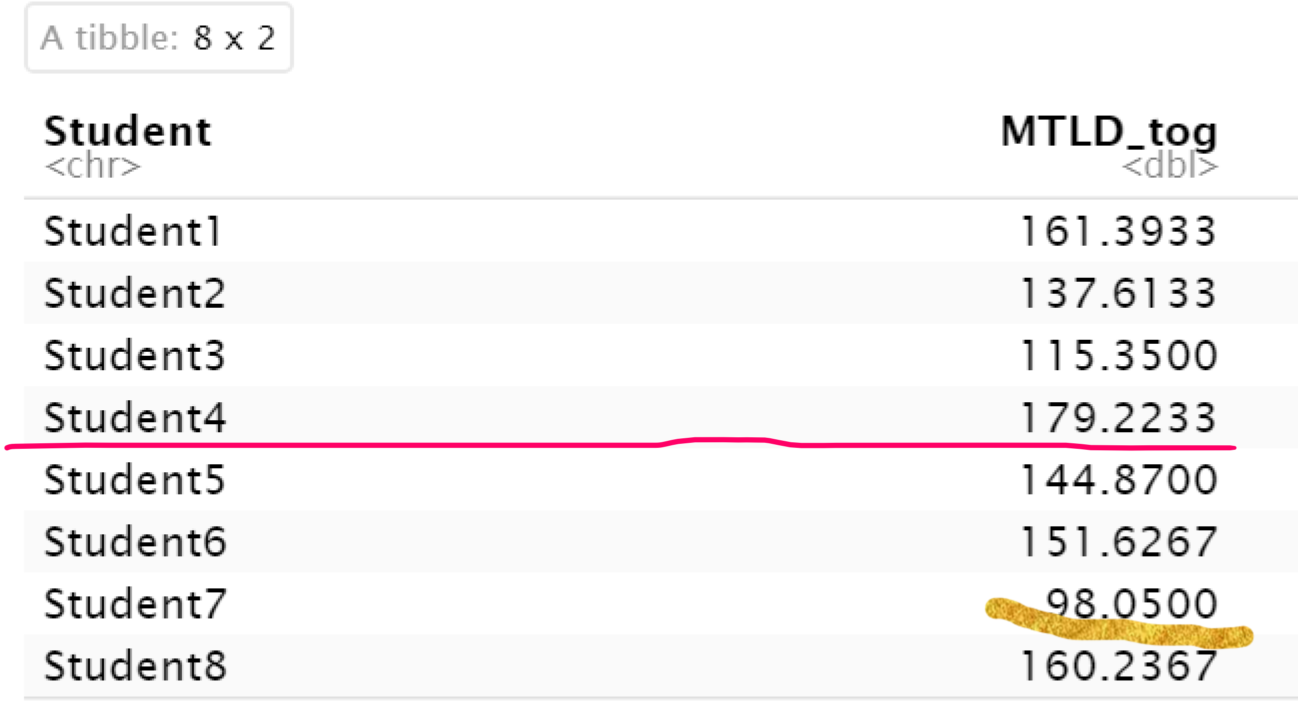
- The ratio of the types (the number of different types of words used) to the tokens (the total number of words used) (Barkaoui & Hadidi, 2020)
- A version of TTR less dependent on text length is Measure of Textual Lexical Diversity (MTLD)
- Hypothesis: The higher the variation, the higher the proficiency

LEXICAL VARIATION (DIVERSITY)



```
#install.packages("koRpus")
#install.koRpus.lang(c("en","ru"))
#available.koRpus.lang()
library(koRpus) #I hope this package helps me calculate MTLD
library(koRpus.lang.ru)
library(koRpus.lang.en)
```

- I installed a package for calculating MTLD
- I calculated MTLD for all students
- I had to create a vector for MTLD manually



A tibble: 8 x 2

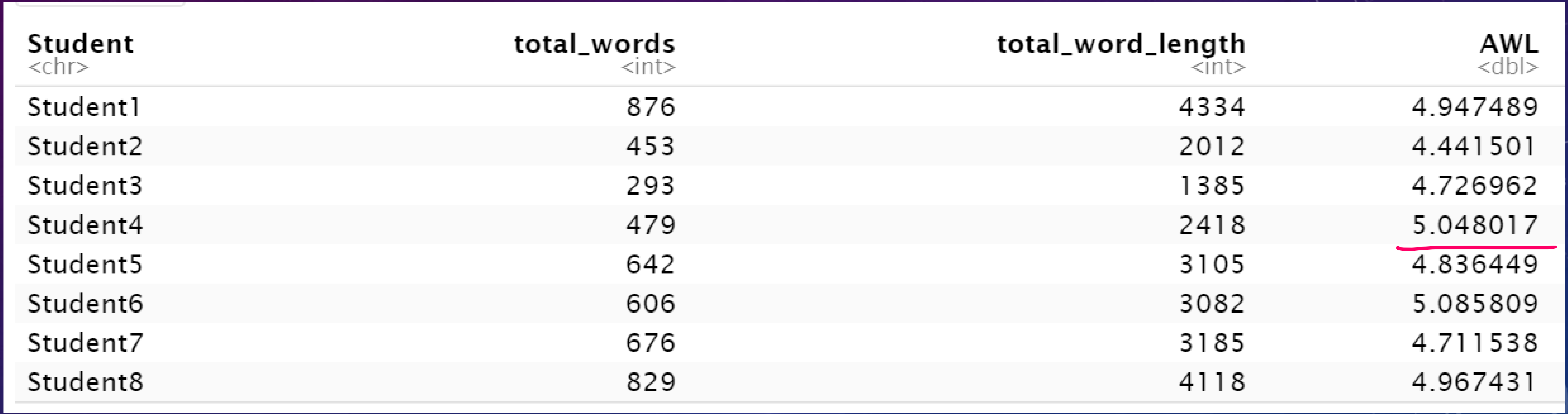
Student	MTLD_tog
<chr>	<dbl>
Student1	161.3933
Student2	137.6133
Student3	115.3500
Student4	179.2233
Student5	144.8700
Student6	151.6267
Student7	98.0500
Student8	160.2367

LEXICAL SOPHISTICATION

- The proportion of relatively unusual, advanced, or low-frequency words to frequent words used in a text
- Can be calculated through average word length (AWL) by dividing the total number of letters by the total number of words (Bakaoui & Hadidi, 2020)
- Hypothesis: the larger the AWL, the higher the proficiency

LEXICAL SOPHISTICATION

- I calculated the length of each word
- I added everything up
- I divided the total length by the total number of words



Student	total_words	total_word_length	AWL
<chr>	<int>	<int>	<dbl>
Student1	876	4334	4.947489
Student2	453	2012	4.441501
Student3	293	1385	4.726962
Student4	479	2418	5.048017
Student5	642	3105	4.836449
Student6	606	3082	5.085809
Student7	676	3185	4.711538
Student8	829	4118	4.967431

CLUSTER ANALYSIS

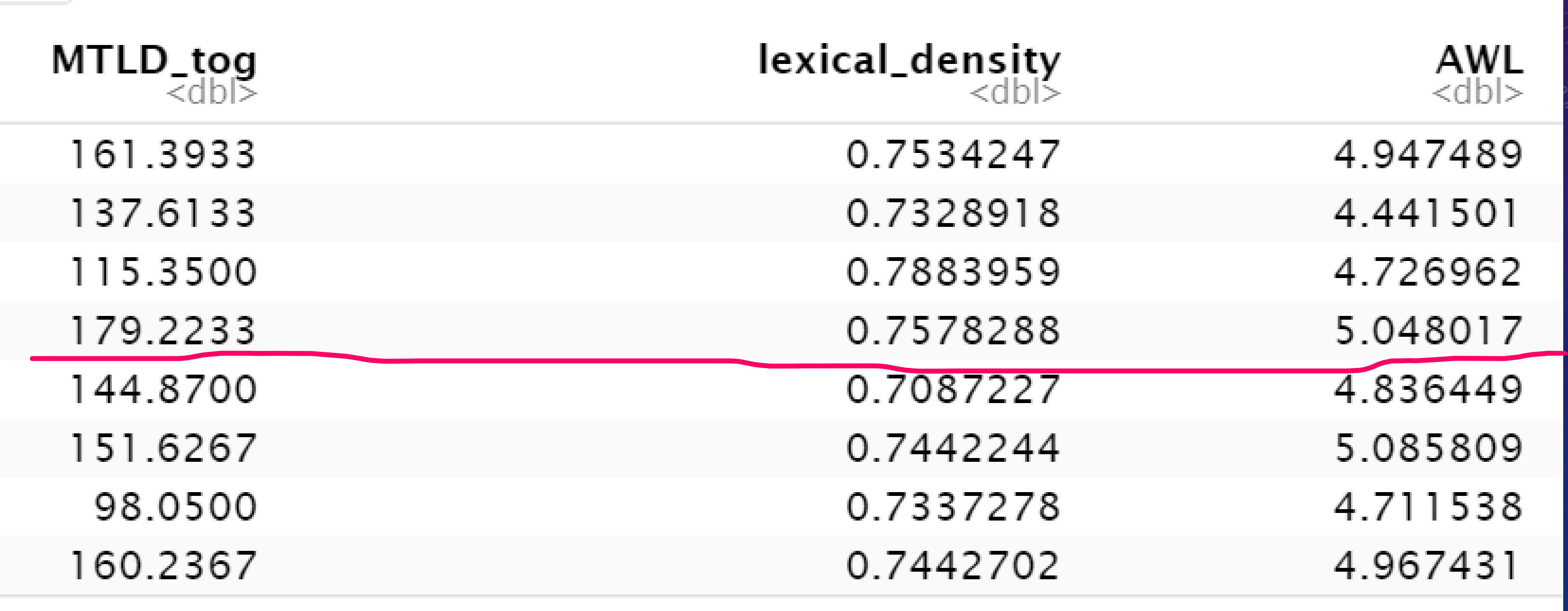
- “Agglomerative hierarchical cluster analysis is a mathematical procedure for classifying cases (e.g., texts) into groups based on their shared similarities across a number of measures (e.g., linguistic features)” (Jarvis et al., 2003, p. 384)

CLUSTER ANALYSIS

- I calculated the three lexical complexity measures
- I scaled my data points
- I performed hierarchical cluster analysis
- I measured the goodness of clusters

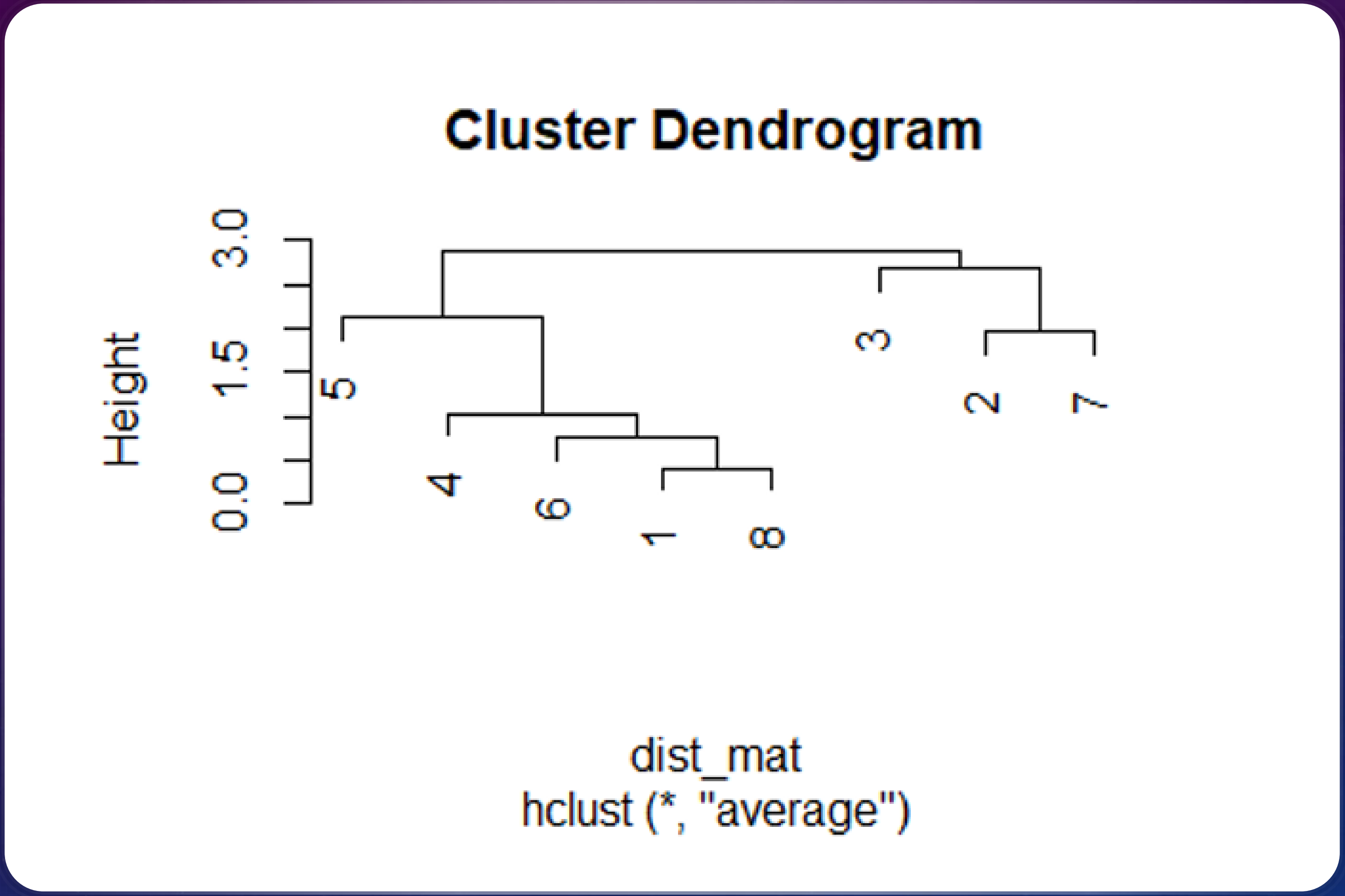
Inspired by the datacamp tutorial:

https://www.datacamp.com/community/tutorials/hierarchical-clustering-R

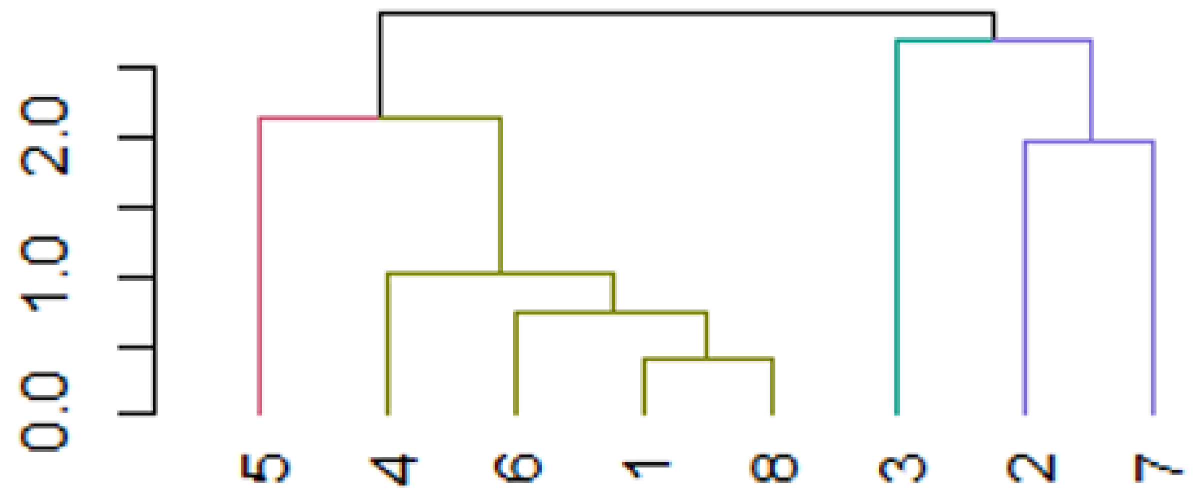


MTLD_tog	lexical_density	AWL
<dbl>	<dbl>	<dbl>
161.3933	0.7534247	4.947489
137.6133	0.7328918	4.441501
115.3500	0.7883959	4.726962
179.2233	0.7578288	5.048017
144.8700	0.7087227	4.836449
151.6267	0.7442244	5.085809
98.0500	0.7337278	4.711538
160.2367	0.7442702	4.967431

CLUSTER ANALYSIS

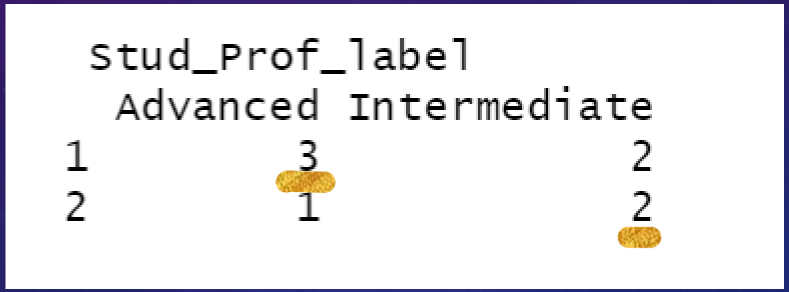


CLUSTER ANALYSIS



- The same data, but with colored branches
- Students 3, 2, 7 have the lowest MTLD and AWL

GOODNESS OF CLUSTERS



```
   Stud_Prof_label
    Advanced Intermediate
  1        3            2
  2        1            2
```

INTERPRETATION OF THE FINDINGS

- Despite high MTLD and AWL, two students were rated as Intermediate , although they were advanced
- Despite low MTLD and AWL, one student was rated as advanced, although they were intermediate
- Lexical complexity does not influence proficiency ratings at intermediate and advanced levels
- Lexical complexity measures relevant for English texts may not be relevant for Russian texts

LIMITATIONS OF THE STUDY

- The corpus size is too small
- The tokenization rules should be checked once more
- The length of essays should be controlled
- The division into lexical and non-lexical items may be revised
- MTLD should be compared with other TTR measures

AN OPTIMISTIC ENDING

- I now can easily calculate three lexical complexity measures in my students' essays and tell my colleagues and my students whether lexical density, lexical variation, and lexical sophistication of their texts has increased or not.