Readability of legal texts

Class #11, April 26 2022

Jana Plaňavová Latanowicz j.planavova-latanowicz@uw.edu.pl
Barbora Hladká hladka@ufal.mff.cuni.cz

Data Analytics for Students of Social Studies and Humanities https://ufal.mff.cuni.cz/courses/npfl134

Part I :: Natural Language Processing and Legal Domain

- HW #5 evaluation
- Czech Legal Text Treebank

HW #5 :: assignment

Universal Dependencies and UDPipe
see Lecture #5: video (from 00:58:00)

Subject annotation

Annotation task

Annotate subjects in the sentences of the preamble of the EU regulation 2020/2092 on a general regime of conditionality for the protection of the Union budget. Our motivation to organize this task is twofold:

- evaluate UDPipe system on manually annotated data
- get experience with readability of legal texts

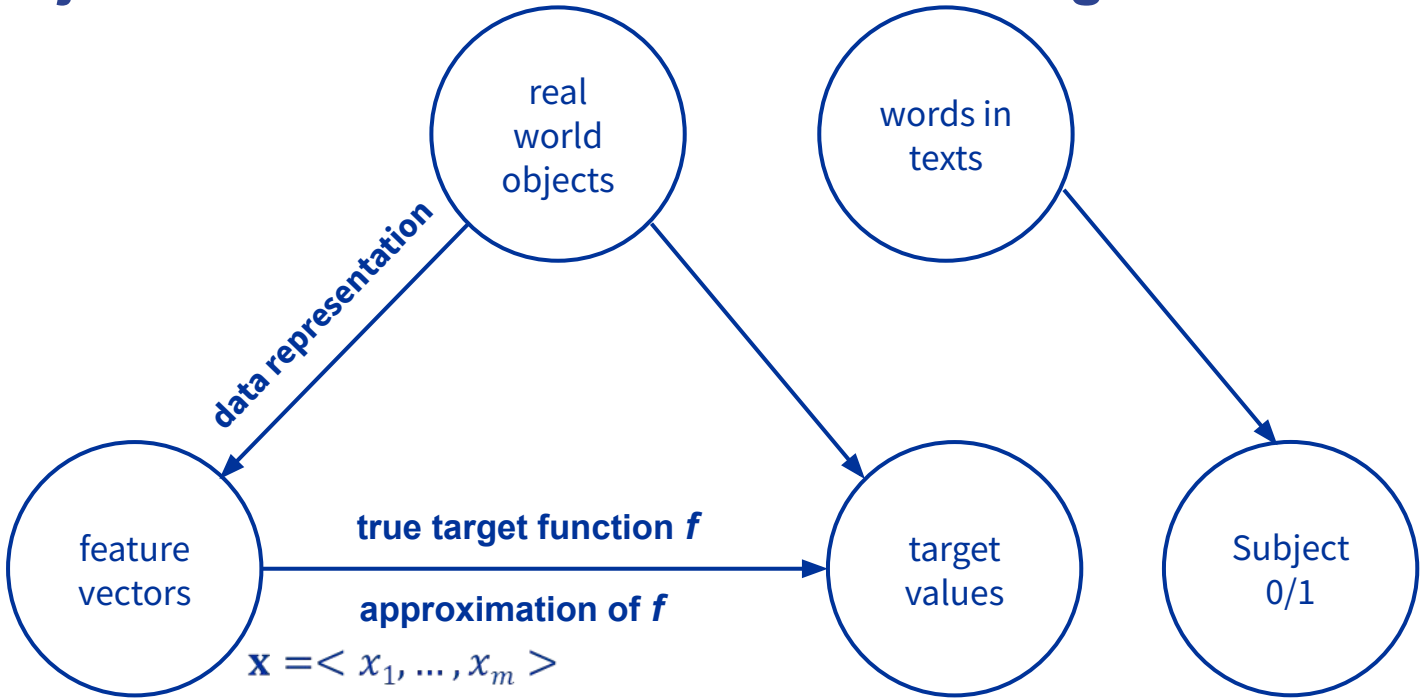
SUB SUB SUB
Once a candidate country becomes a Member State, it joins a legal structure that is based on the fundamental premiss

The homework should be completed by April 19.

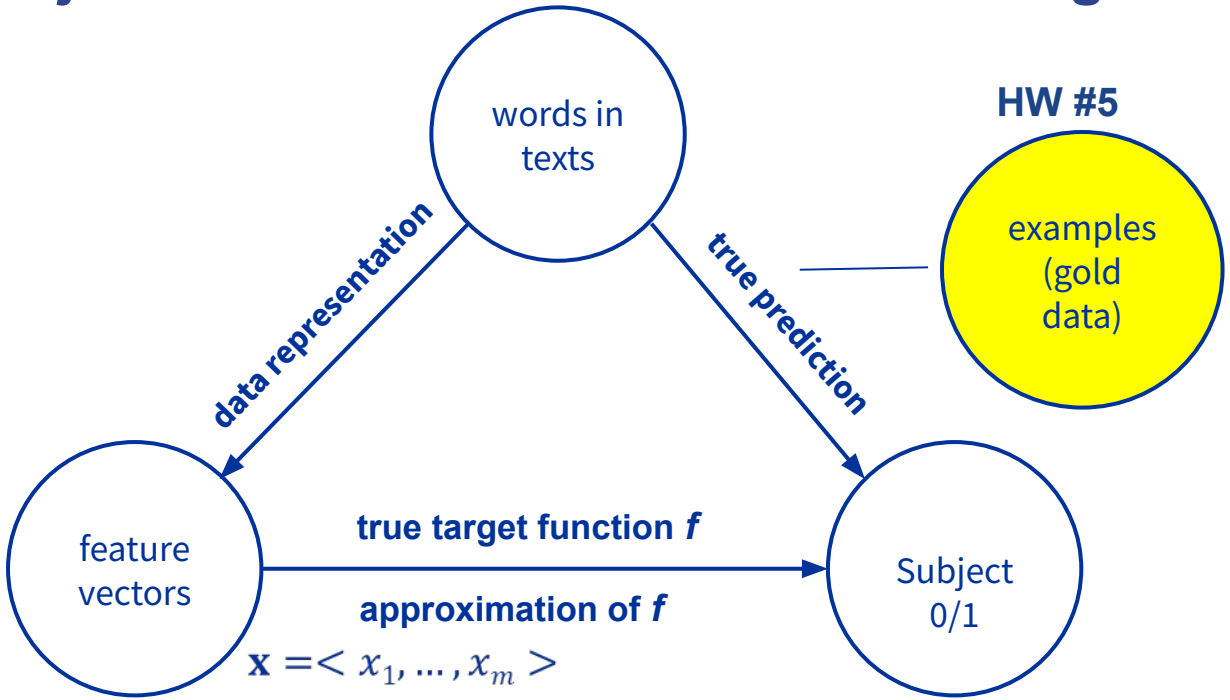
Annotation instructions

- To log in to Brat editor use the credentials sent via e-mail.
- Read carefully the preamble and identify subject(s) in each sentence. We follow the Universal Dependencies annotation guidelines where the basic units of annotation are (syntactic) words, which means that the subject is exactly one word in our annotation task. Typically, it is a noun, pronoun or relative pronoun. Mark all subjects standing in a coordinated construction separately.

Subject annotation as a machine learning task



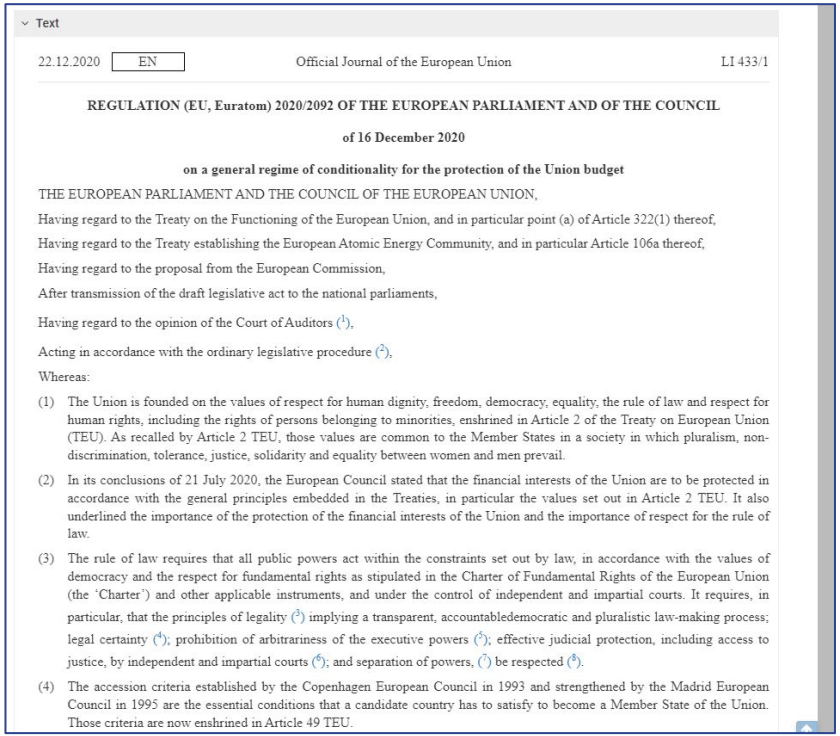
Subject annotation as a machine learning task



HW #5 :: data

Regulation (EU, Euratom) 2020/2092 of the European Parliament and of the Council of 16 December 2020 on a general regime of conditionality for the protection of the Union budget

its preamble (29 items)



Text

22.12.2020 EN Official Journal of the European Union LI 433/1

REGULATION (EU, Euratom) 2020/2092 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL

of 16 December 2020

on a general regime of conditionality for the protection of the Union budget

THE EUROPEAN PARLIAMENT AND THE COUNCIL OF THE EUROPEAN UNION,

Having regard to the Treaty on the Functioning of the European Union, and in particular point (a) of Article 322(1) thereof,

Having regard to the Treaty establishing the European Atomic Energy Community, and in particular Article 106a thereof,

Having regard to the proposal from the European Commission,

After transmission of the draft legislative act to the national parliaments,

Having regard to the opinion of the Court of Auditors (1),

Acting in accordance with the ordinary legislative procedure (2),

Whereas:

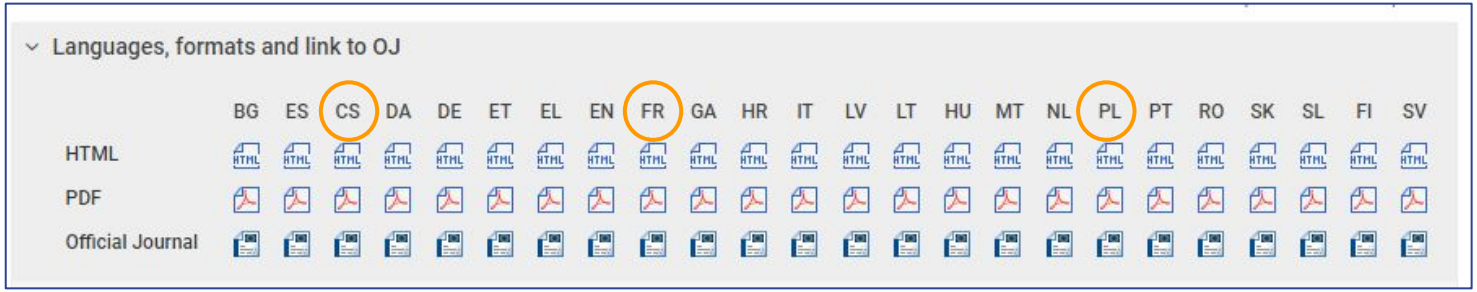
(1) The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities, enshrined in Article 2 of the Treaty on European Union (TEU). As recalled by Article 2 TEU, those values are common to the Member States in a society in which pluralism, non-discrimination, tolerance, justice, solidarity and equality between women and men prevail.

(2) In its conclusions of 21 July 2020, the European Council stated that the financial interests of the Union are to be protected in accordance with the general principles embedded in the Treaties, in particular the values set out in Article 2 TEU. It also underlined the importance of the protection of the financial interests of the Union and the importance of respect for the rule of law.

(3) The rule of law requires that all public powers act within the constraints set out by law, in accordance with the values of democracy and the respect for fundamental rights as stipulated in the Charter of Fundamental Rights of the European Union (the 'Charter') and other applicable instruments, and under the control of independent and impartial courts. It requires, in particular, that the principles of legality (3) implying a transparent, accountabledemocratic and pluralistic law-making process; legal certainty (4); prohibition of arbitrariness of the executive powers (5); effective judicial protection, including access to justice, by independent and impartial courts (6); and separation of powers, (7) be respected (8).

(4) The accession criteria established by the Copenhagen European Council in 1993 and strengthened by the Madrid European Council in 1995 are the essential conditions that a candidate country has to satisfy to become a Member State of the Union. Those criteria are now enshrined in Article 49 TEU.

What we have - manually annotated subjects in CS, PL, FR versions, each version by 2 annotators



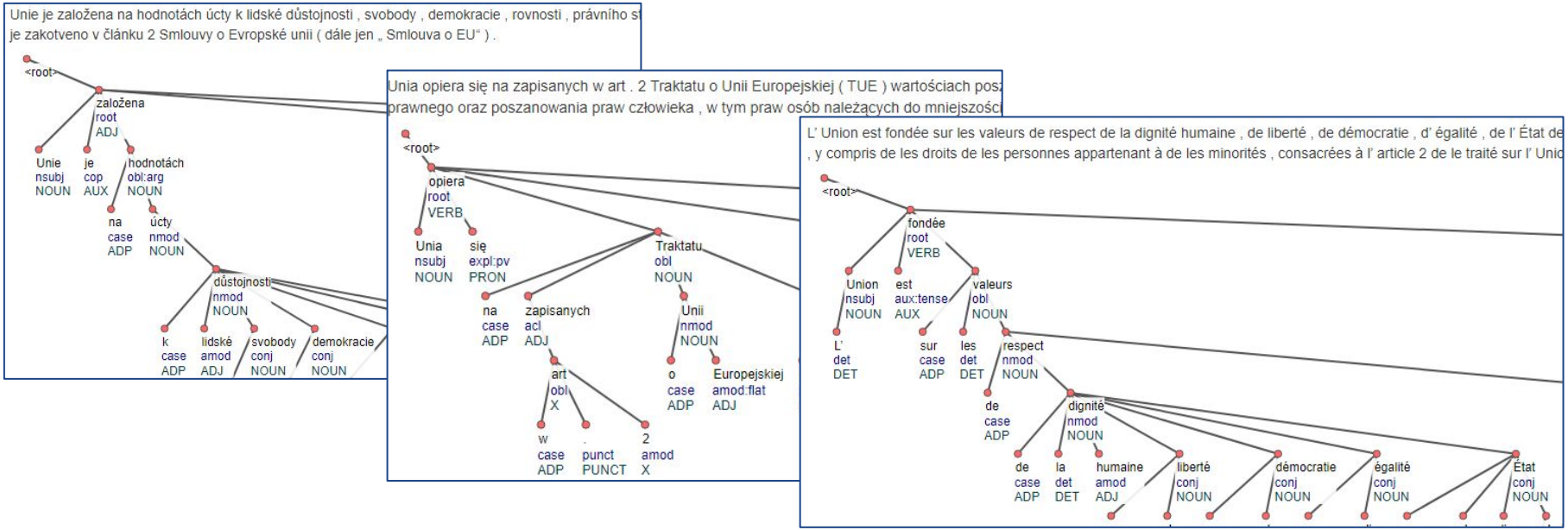
	CS	PL	FR
# of tokens	2,285	2,259	3,160

33 (9) Niezawisłość i bezstronność sądownictwa powinny być zawsze zagwarantowane, a służby dochodzeniowo-śledcze i prokuratura

34 Sądownictwo, służby dochodzeniowo-śledcze i prokuratura powinny dysponować wystarczającymi zasobami finansowymi i ludzkimi c bezstronnego sądu, w tym poszanowanie prawa do obrony.

35 Prawomocne wyroki powinny być skutecznie wykonywane.

What we have - a complex syntactic analysis of each version using UDPipe



Inter-Annotator Agreement :: CS

see Lecture #8

Confusion matrix for 2 annotators treating A1 CS as gold standard and A2 CS as classifier

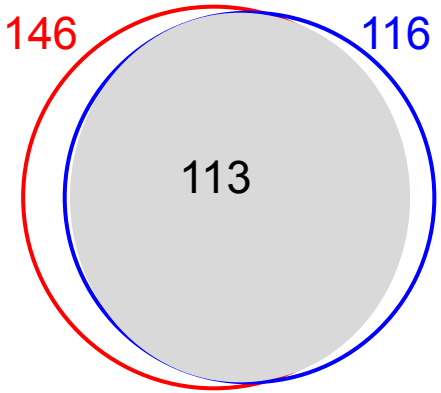
A1 CS \ A2 CS	1	0	∑
1	113	33	146
0	3	N-149	N-146
∑	116	N-116	N

N = total number of words in Czech version

F1 = 2*0.97*0.77/(0.97+0.77) = 0.85

Precision = 113/(113+3) = 0.97

Recall = 113/(111+33) = 0.77



Inter-Annotator Agreement :: CS

Confusion matrix for 2 annotators treating A2 CS as gold standard and A1 CS as classifier

A2 CS

A1 CS	1	0	∑
1	113	33	146
0	3	N-149	N-146
∑	116	N-116	N

F1 = 2*0.97*0.77/(0.97+0.77) = 0.85
Precision = 113/(113+3) = 0.97
Recall = 113/(111+33) = 0.77

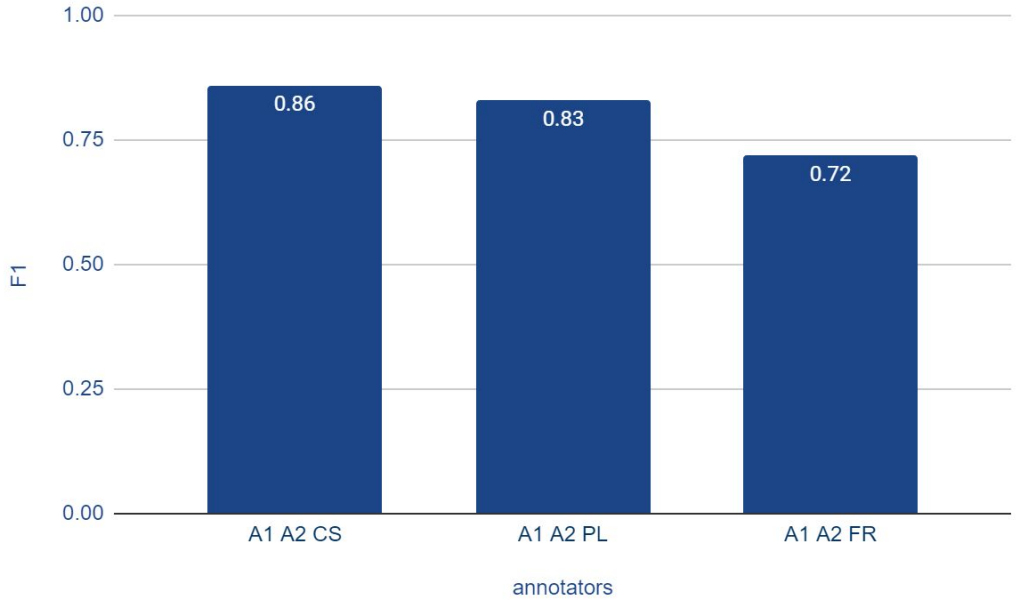
A1 CS

A2 CS	1	0	∑
1	113	3	116
0	33	N-149	N-116
∑	146	N-146	N

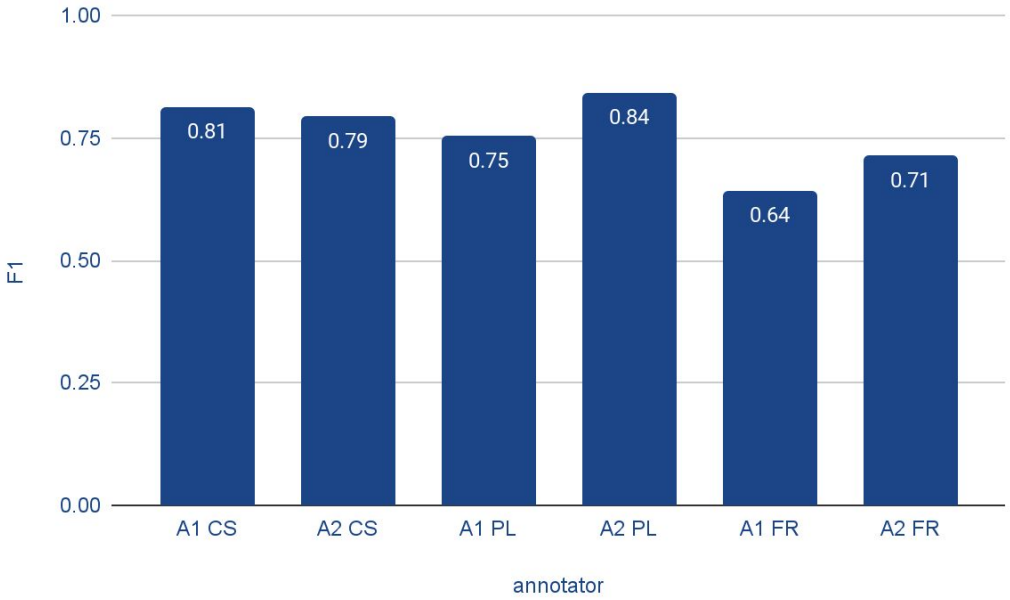
F1 = 2*0.97*0.77/(0.97+0.77) = 0.85
Precision = 113/(113+33) = 0.77
Recall = 113/(111+3) = 0.97

N = total number of words in Czech version

Inter-Annotator Agreement



UDPipe evaluation



Accounting units shall take inventory of their assets and liabilities pursuant to section 29 and 30. The relation of obligation is present: what accounting units have to do – take inventory.

Czech Legal Text Treebank http://hdl.handle.net/11234/1-2498

is a collection of 1,133 manually annotated sentences from two Czech legal documents: The Accounting Act (563/1991 Coll., as amended) and Decree on Double-entry Accounting for undertakers (500/2002 Coll., as amended):

- morphologically
- syntactically
- entities = terms from the accounting domain
- relations of definition, right, obligation

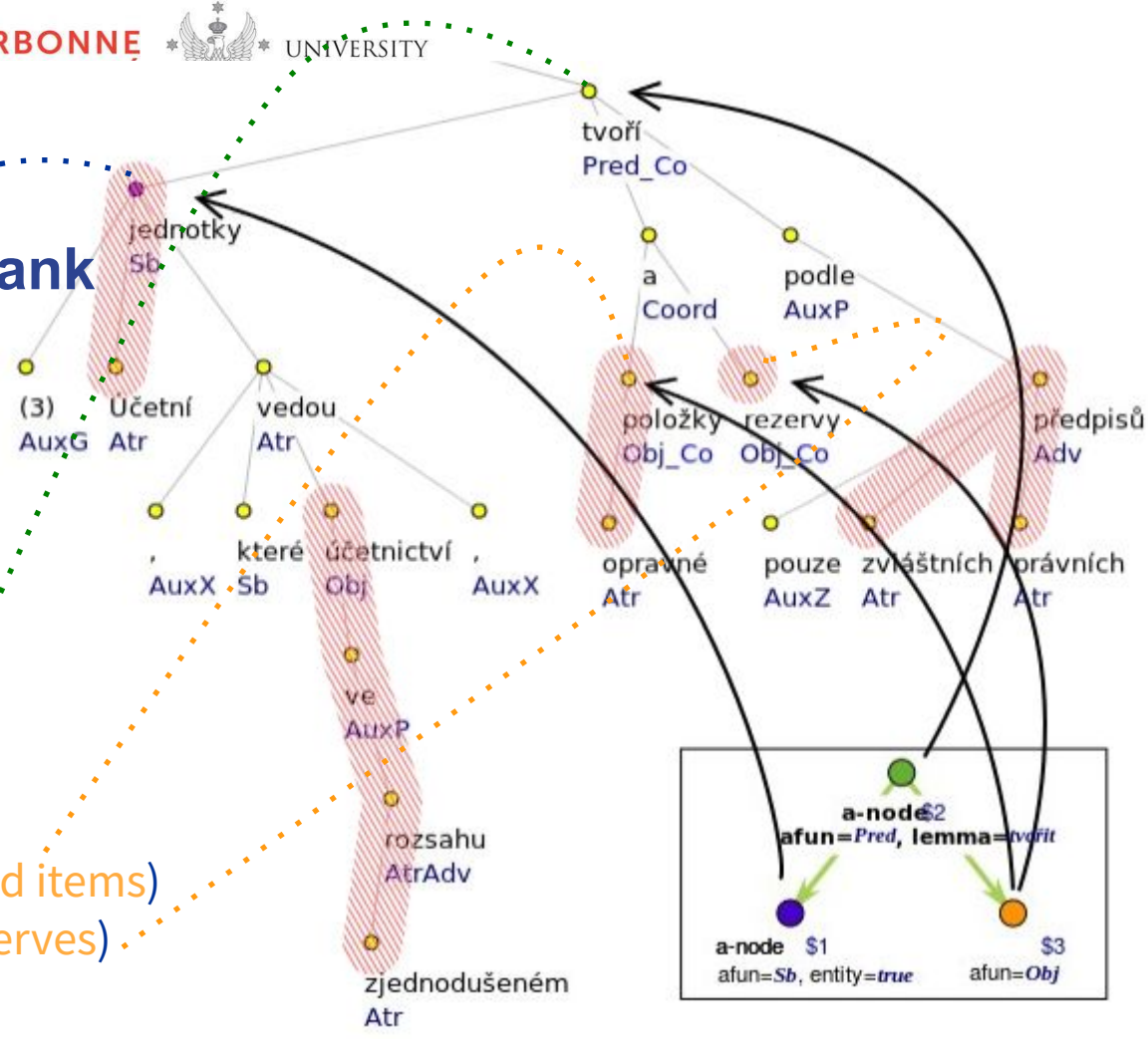
Czech Legal Text Treebank

Example

(3) Účetní jednotky, …, tvoří
opravné položky a rezervy …

(3) Accounting units, …,
create
fixed items and reserves …

(accounting units, create, fixed items)
(accounting units, create, reserves)



A survey on Subject annotation

1. How much time did you spend on the annotation task?
2. Do you consider reading and understanding the text difficult?

Part II :: Readability of legal texts

Why readability of legal texts and policy documents is important?

- rule of law element (https://worldjusticeproject.org/)
  - accountability
  - law is clear, publicized, stable and is applied evenly
  - accessibility of law and justice (incl. enforcement)
  - part of good administration
- economic impact (legal counselling is not attainable)
- need to understand and translate law into technical specifications (e.g. EU regulations on the technological sectors like digital single market legislation)

Why measuring readability of legal texts and policy documents is important?

- decline in literature skills (incl. understanding, evaluating and using written texts)
- intentional obfuscation (e.g. lawyers gobbledygook)
- complexity of topics covered
- multilinguality of EU legislation is taken into account during the law-making processes

How to measure readability of EU law? (State of the art)

Jukka Ruohonen research (2021) https://arxiv.org/pdf/2102.11625.pdf

sample of 201 EU documents on Single Digital Market

5 readability indices:

- mostly based on the grade level of education (eg. score 9,3 - ninth grader would be able to read the text, score 22 for a university graduate)
- https://pypi.org/project/textstat/

Conclusion: hypothetical grade level around 30

Criticism: does not take into account different types of EU acts and documents

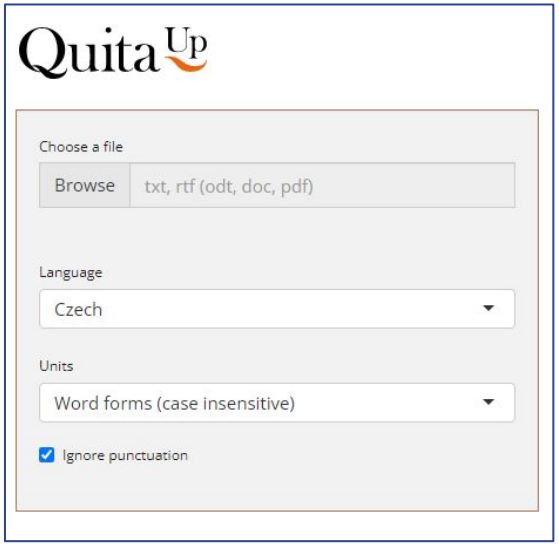
QuitUp comparison and Ruohonen research

Caveats:

- stylometry (QuitUp) vs readability (Ruohonen)
- multilingual vs monolingual approach

QuitaUp

- https://korpus.cz/quitaup/
- pre-processes input texts using UDPipe
  - i.e., tokenization, lemmatization, POS tagging, syntactic parsing



QuitaUp :: Terminology

You cannot end a sentence with because because because is a conjunction.

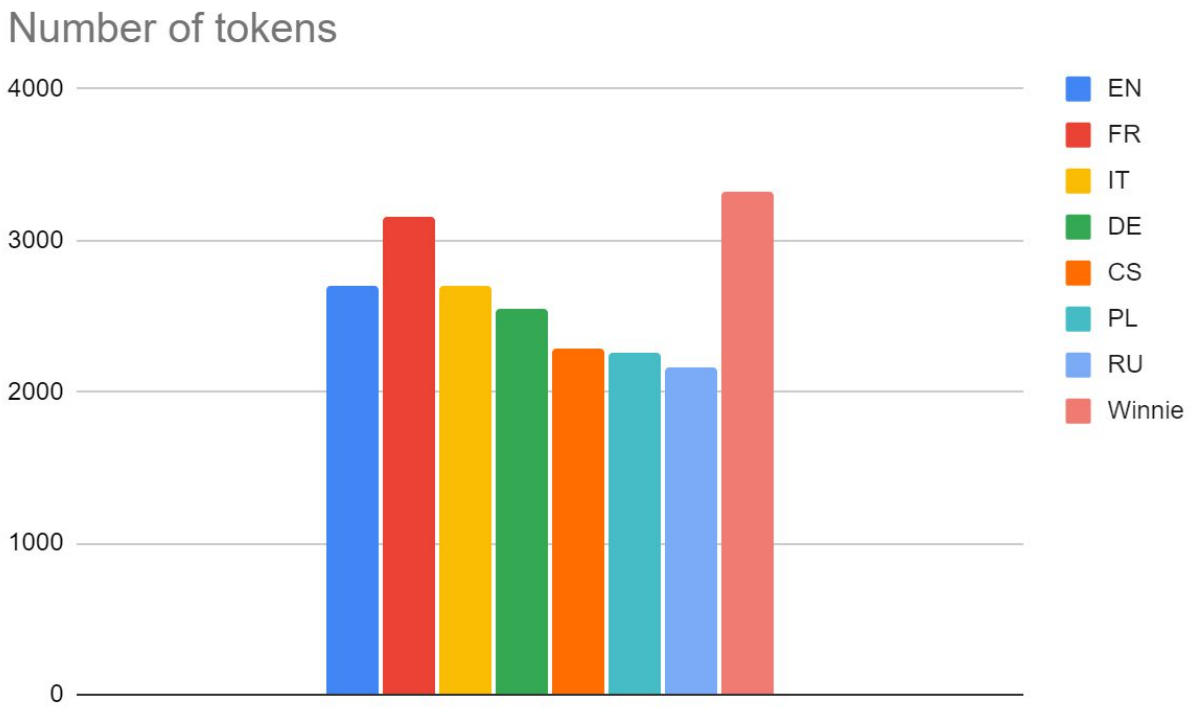
- tokens = smaller units in a text
- token length = number of characters (because 7)
- word types = different tokens in a text (9)
- function words (synsemantic words) have little lexical meaning (prepositions, conjunctions, pronouns, …) you, a, with, is, cannot
- content words (autosemantic words) possess semantic content (nouns, adjectives, most verbs, most adverbs) end, sentence, conjunction

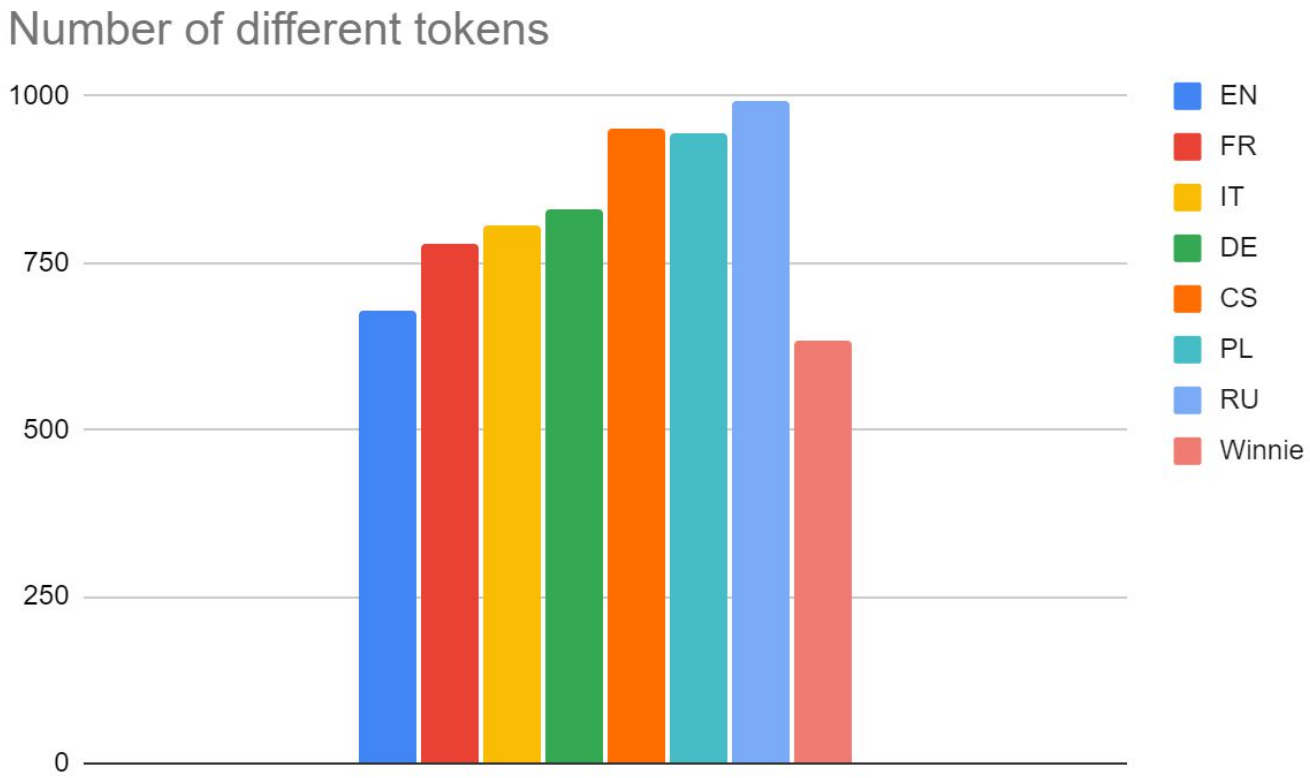
QuitaUp :: Comparison

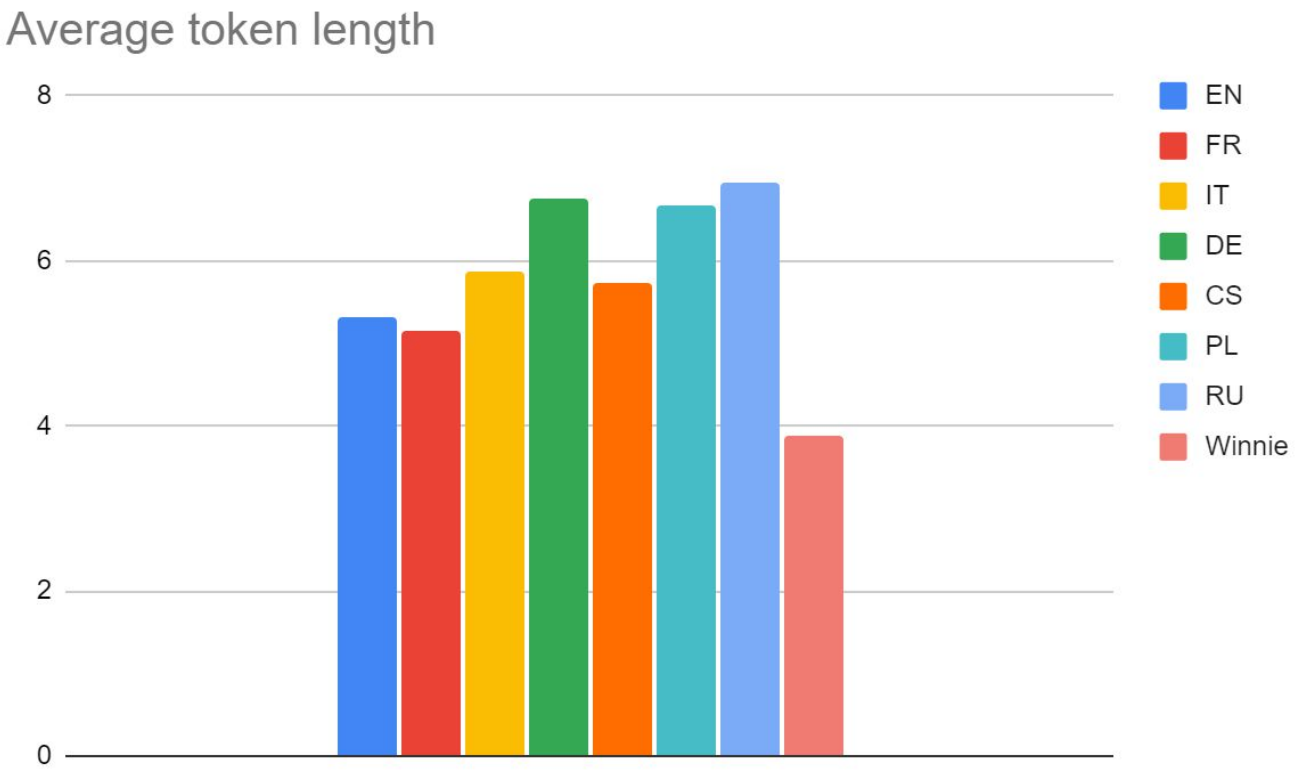
What we have compared:

- 6 official translations of a preamble of the Regulation 2020/2092 of 16 December 2020 on a general regime of conditionality for the protection of the Union budget
- mechanical translation of the Regulation from English into Russian by using https://lindat.mff.cuni.cz/services/translation
- A.A. Milne “Winnie-the-Pooh” (in original language, about 1,5 chapter, number of characters roughly similar to the one of Regulation’s preamble in English)

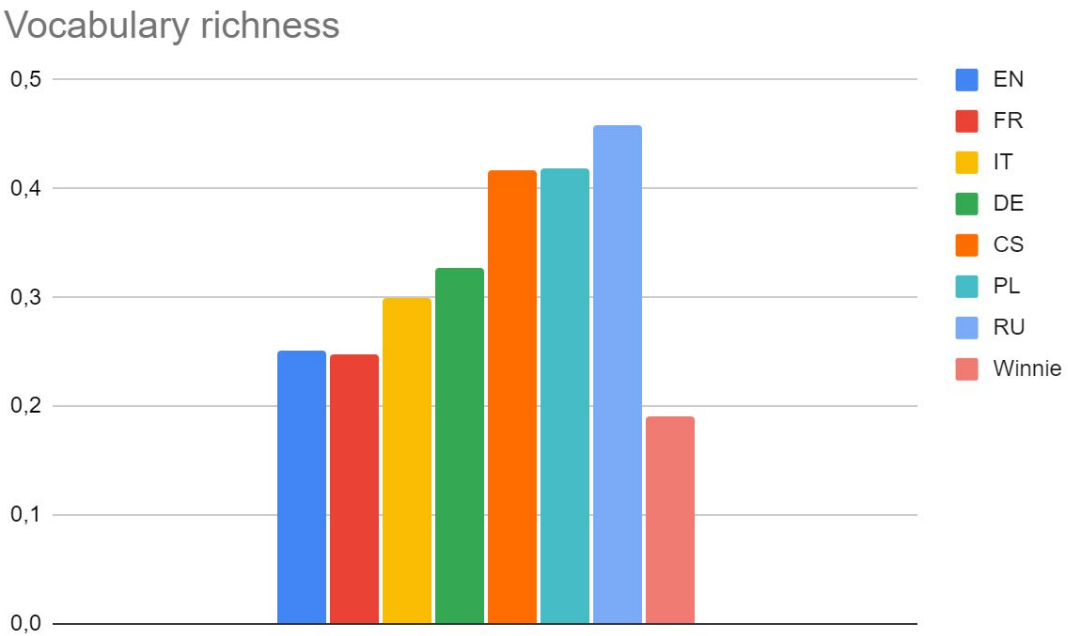
Token-related measures

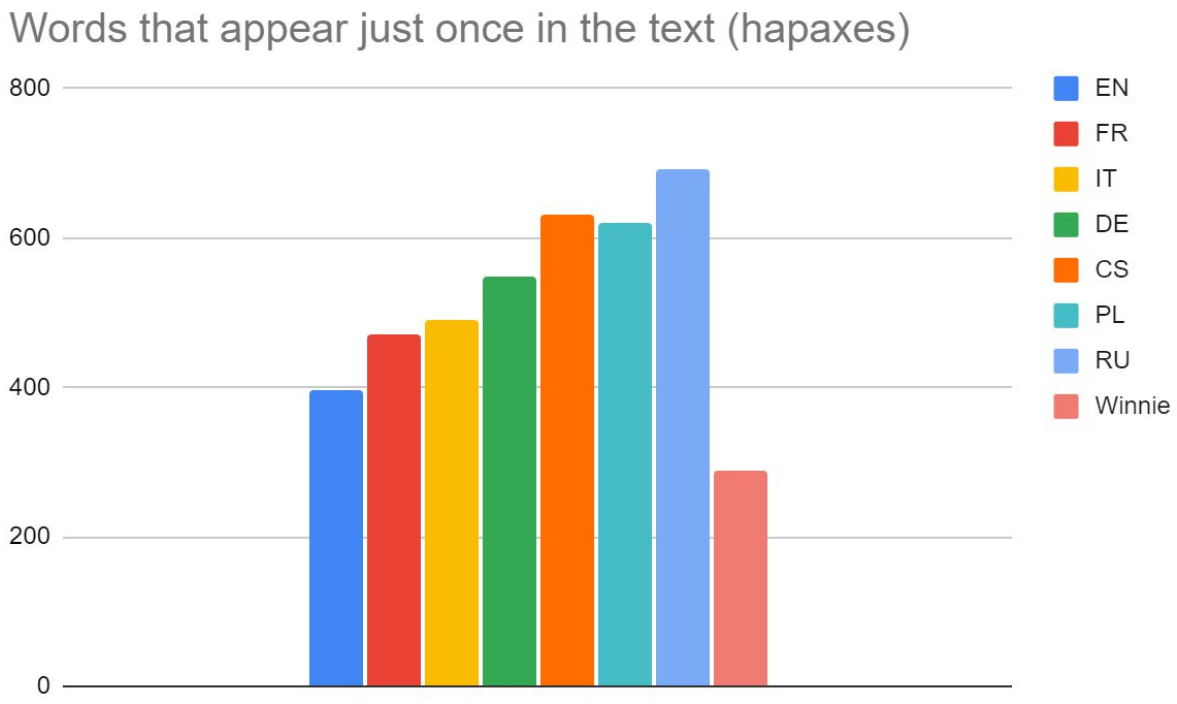



Number of different tokens
1000
750
500
250
0
EN
FR
IT
DE
CS
PL
RU
Winnie

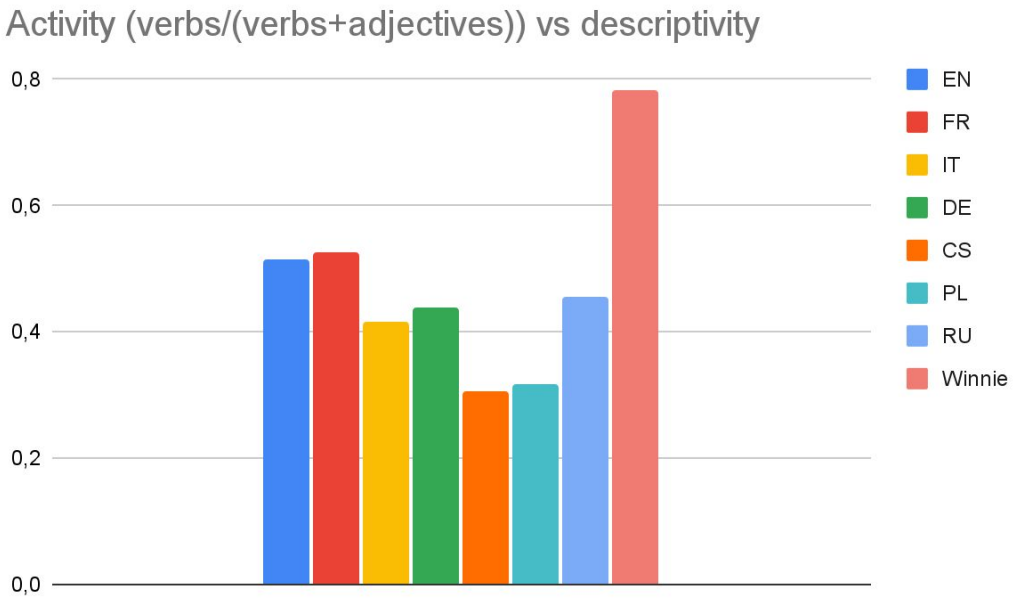

Average token length
8
6
4
2
0
EN
FR
IT
DE
CS
PL
RU
Winnie

Word types

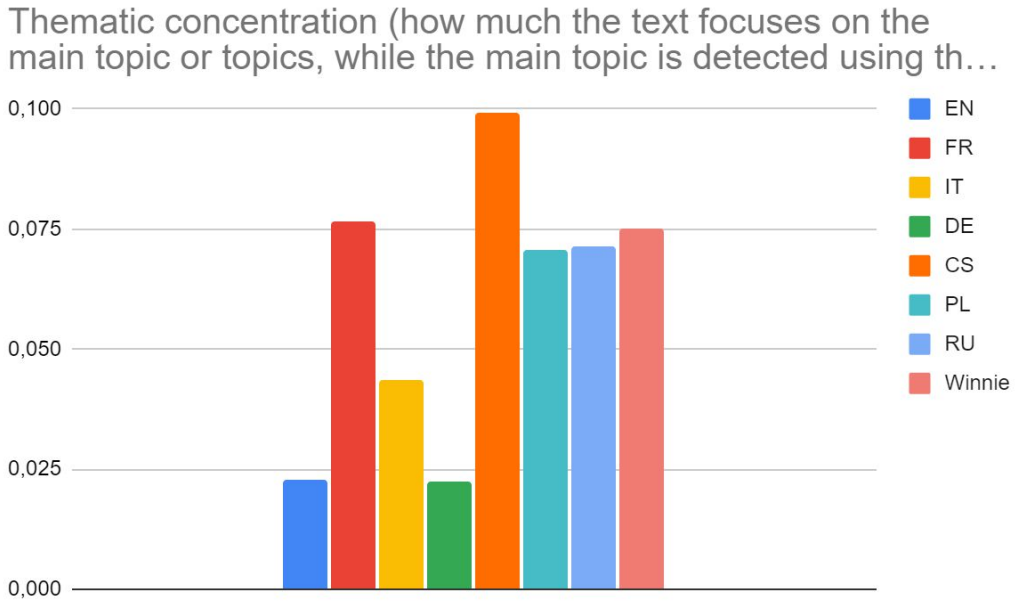

Vocabulary richness
0,5
0,4
0,3
0,2
0,1
0,0
EN
FR
IT
DE
CS
PL
RU
Winnie


Words that appear just once in the text (hapaxes)
800
600
400
200
0
EN
FR
IT
DE
CS
PL
RU
Winnie

Content words



Thematic concentration



	rank	word type	frequency
function branch	1.	w_1	f_1
	2.	w_2	f_2
	3.	w_3	f_3
	…	…	…
content branch	V.	w_v	f_V

$f_1 >= f_2 >= f_3 >= \ldots >= f_V$

Thematic words are content words occurring in the function branch. This occurrence is considered a certain anomaly caused by the importance of the word type in a text.

Thematic Concentration :: CS and EN

Word	POS	TW (primary TC)
unie	NOUN	0.0314
státu	NOUN	0.0189
komise	NOUN	0.0145
opatření	NOUN	0.0145
měla	VERB	0.0102
právního	ADJ	0.0069
zásad	NOUN	0.0028

Word	POS	TW (primary TC)
law	NOUN	0.0063
union	NOUN	0.0056
member	PROPN	0.0033
measures	NOUN	0.0025
rule	NOUN	0.0025
commission	NOUN	0.0016
financial	ADJ	0.0009

Conclusions

- Readability of legal acts is language-specific
  (it should be measured for each language)
- Measuring readability of legal documents should reflect the type of the legal document in question
  (e.g. EU regulation, national law, contracts, policy documents)
- Readability of legal documents should be defined against “average user” of the given type of document
  (generally applied rules vs individually addressed rules)

Further reading

- Jukka Ruohonen: Assessing the Readability of Policy Documents on the Digital Single Market of the European Union (2021) https://arxiv.org/pdf/2102.11625.pdf
- Hilary Frooman: Lawyers and Readability (1981) https://journals.sagepub.com/doi/10.1177/002194368101800405
- In Czech: Jak umělá inteligence chápe občanský zákoník (Noc vědců, 2020) https://www.youtube.com/watch?v=3EtQIzQ7CsE (from 3:00:00)