

Exploring readability in Czech legal writing

Ivan Kraus

ÚČJTK FF UK

May 12, 2025

Readable legal writing?

Rooted in the Plain English movement
cf. Mellinkoff (1963)

Still new to Czech...

... but there are handbooks!
cf. Šamánková and Kubíková (2022), Šváb (2023)

PONK

‘Web application for the automatic assessment of the accessibility (comprehensibility or clarity) of Czech legal texts.’

<https://ufal.mff.cuni.cz/grants/ponk>

The PONK App measures...

Stylometric indicators (Cvrček et al., 2020)

Traditional **readability formulas** (Bendová & Cinková, 2021; Coleman & Liau, 1975; Flesch, 1948; Gunning, 1952; Kincaid et al., 1975; McLaughlin, 1969; Senter & Smith, 1967)

Select **grammar** and **lexicon** (Chromý & Ceháková, 2023; Šamánková & Kubíková, 2022; Sgall & Panevová, 2014; Šváb, 2023)

Adding up to 61 variables in this paper

Corpora

KUK 1.0 (Hladká et al., 2024): Czech administrative and legal texts
FrBo, OmbuFlyers, ESO

KUKY 1.0 (Cinková et al., 2024): Czech administrative and legal texts

CzCDC (Novotná & Harašta, 2019): Czech court decisions

LiFR–Law (Cinková et al., 2023): Czech legal and administrative texts

Data set composition

KUKY and FrBo documents
annotated by experts on readable
legal writing.

55:45 bad:good split.

section	bad	good
CzCDC	211	0
FrBo	78	229
KUKY	84	110
LiFRLaw	3	0
OmbuFlyers	38	0
<i>total</i>	415	339

Variables with strong effects on readability

Kruskal-Wallis test
with epsilon squared
effect size

Variables related to
verbs, nouns,
adjectives, and
sentence length

Readability formulas
(some)

variable	p-value	effect size	sign
activity	< .0001	0.275	+
verbdist	< .0001	0.273	-
sentlen.m	< .0001	0.272	-
ari	< .0001	0.271	-
gf	< .0001	0.259	-
smog	< .0001	0.252	-
NOUNcount.m	< .0001	0.246	-
VERBfrac.m	< .0001	0.238	+
...			

Variables with weak / nonsignificant effects on readability

Grammatical ambiguities

Weak semantics

Pragmatics (hedges, extreme case formulations)

Text length

(and others; 12 nonsignificant in total)

Exploratory factor analysis

To see the main tendencies in readability, let us...

- 1 Take variables with significant effects on readability*
... because we know they are relevant
- 2 Group similar ones together into **factors**
... meaning we need to leave stand-alones** out

* Except for the readability formulas. They already *are* several variables grouped together.

** With $|r| < 0.35$.

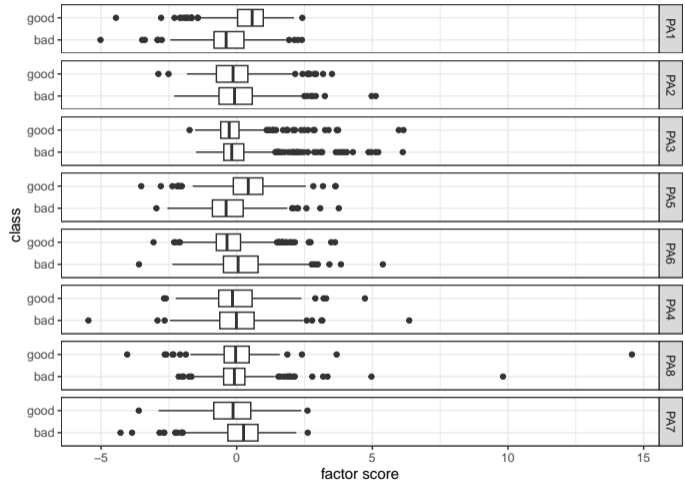
See Watkins (2021) or Fabrigar and Wegener (2012) for more on EFA.

8 factors

factor	denotation
PA1	technical / spoken-like
PA2	short / long
PA3	uniform syntax / diverse syntax
PA5	prestigious / neutral
PA6	less negation / more negation
PA4	fewer objects / more objects
PA8	early predicate / predicate further in the clause
PA7	lexically uniform / lexically diverse

8 factors

factor	denotation
PA1	technical / spoken-like
PA2	short / long
PA3	uniform syntax / diverse syntax
PA5	prestigious / neutral
PA6	less negation / more negation
PA4	fewer objects / more objects more nominalizations / fewer nominalizations
PA8	early predicate / predicate further in the clause
PA7	lexically uniform / lexically diverse



Factors and readability

Spokenness (PA1) and **neutral style** (PA5) are strong predictors of readability

Negation (PA6), **lexical diversity** (PA7) and **syntactic diversity** (PA3) weaker, but still significant

Objects (nominalizations) (PA4), predicate position (PA8) and text length (PA2) are not

factor	p-value	effect size
PA1	< 0.0001	0.178
PA2	0.21	0.002
PA3	< 0.01	0.011
PA5	< 0.0001	0.149
PA6	< 0.0001	0.047
PA4	0.16	0.003
PA8	0.19	0.002
PA7	< 0.0001	0.034

What about the stand-alones?

They were not very strong to begin with

variable	p-value	effect size	sign
NOUNfrac.v	< 0.0001	0.036	+
relativisticexprs	< 0.0001	0.024	-
caserepcount.v	< 0.001	0.015	-
redundexprs	< 0.01	0.014	-
extrcaseexprs	< 0.01	0.014	-
abstractNOUNs	< 0.01	0.010	+
anaphoricrefs	< 0.01	0.009	+
VERBcompdist.m	< 0.05	0.008	-

What to take away

What to take away

Readable Czech legal writing **does** exist.

What to take away

Readable Czech legal writing **does** exist.

Use verbs and keep the sentences short when writing!

Bendová, K., & Cinková, S. (2021). Adaptation of Classic Readability Metrics to Czech. *24th International Conference on Text, Speech and Dialogue*, 159–171.

https://doi.org/10.1007/978-3-030-83527-9_14

Chromý, J., & Ceháková, M. (2023). Diversity of garden-path structures - SPR.

<https://doi.org/10.17605/OSF.IO/KSTPE>

Cinková, S., Chromý, J., Šamánková, J., Hořeňovská, K., Kettnerová, V., Kolářová, V., Kubištová, H., & Panevová, J. (2023). LiFR-Law. Corpus of Paraphrased Czech Administrative Texts with Reading Comprehension for Readability Studies.

<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5225>

Cinková, S., Kuk, M., Šamánková, J., Kubíková, B., Pospíšil, P., Mírovský, J., Hladká, B., & Novotná, T. (2024). KUKY1.0. <https://ufal.mff.cuni.cz/grants/ponk/kuky>.

<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5812>

- Coleman, M., & Liau, T. L. (1975). A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2), 283–284. <https://doi.org/10.1037/h0076540>
- Cvrček, V., Čech, R., & Kubát, M. (2020). QuitaUp – a tool for quantitative stylometric analysis. <https://korpus.cz/quitaup/>
- Fabrigar, L. R., & Wegener, D. T. (2012). *Exploratory Factor Analysis*. Oxford University Press.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221–233. <https://doi.org/10.1037/h0057532>
- Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill.
- Hladká, B., Cinková, S., Kuk, M., Mírovský, J., Novotná, T., & Zahálková, K. N. (2024). KUK 1.0. <https://ufal.mff.cuni.cz/grants/ponk/kuk1.0>.
<https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5821>

- Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel* (tech. rep.). Institute for Simulation and Training, University of Central Florida.
- McLaughlin, G. H. (1969). SMOG grading – a new readability formula. *Journal of reading*, 12(8), 639–646.
- Mellinkoff, D. (1963). *The Language of the Law*. Resource Publications.
- Novotná, T., & Harašta, J. (2019). Czech Court Decisions Corpus (CzCDC 1.0).
<https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-3052>
- Šamánková, J., & Kubíková, B. (Eds.). (2022). *Jak psát srozumitelné úřední texty: Příručka srozumitelného psaní pro úředníky*. Kancelář veřejného ochránce práv.
https://www.ochrance.cz/uploads-import/ESO/p%C5%99%C3%ADru%C4%8Dka/Prirucka_srozumitelneho_psani_tisk.pdf

- Senter, R., & Smith, E. A. (1967). *Automated readability index* (tech. rep.). AMRL-TR. Aerospace Medical Research Laboratories (U.S.)
- Sgall, P., & Panevová, J. (2014). *Jak psát a jak nepsat česky* (2nd ed.). Nakladatelství Karolinum.
- Šváb, J. (2023). *Jak psát, aby se to dalo číst: Příručka přístupného psaní* (2nd ed.). Leges.
- Watkins, M. (2021, January). *A Step-by-Step Guide to Exploratory Factor Analysis with R and RStudio*. Routledge.