

VŠB TECHNICKÁ  
UNIVERZITA  
OSTRAVA

VSB TECHNICAL  
UNIVERSITY  
OF OSTRAVA



[www.vsb.cz](http://www.vsb.cz)

# Kompresa stromových struktur

## Semestrální projekt

Marek Beran

VŠB – Technická univerzita Ostrava

marek.beran.st@vsb.cz

27. května 2025



- 1 Úvod
- 2 Implementace knihovny
- 3 Převod textu do stromové struktury
- 4 Algoritmy
- 5 Testování
- 6 Literatura



- Je možné efektivně komprimovat textové data převedením do stromové struktury?
- Cíl: Proof of Concept



## Programovací jazyk a platforma:

- C# 9.0
- .NET 5.0 a vyšší
- Visual Studio 2022

## Knihovny:

- UDPipe (rozpoznávání syntaktických stromů)
- MorphoDiTa (morfologická analýza)

## Další nástroje:

- R (datová analýza a vizualizace)
- Mkdocs (dokumentace)
- Bash skripty (podpůrné nástroje)

## Bindings:

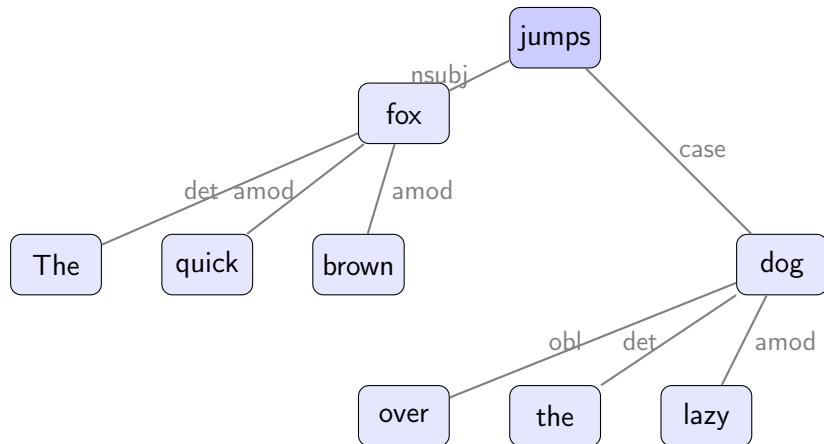
- C# wrapper pro UDPipe (nativní knihovna)
- C# wrapper pro MorphoDiTa (nativní knihovna)



## Převod textu do stromové struktury



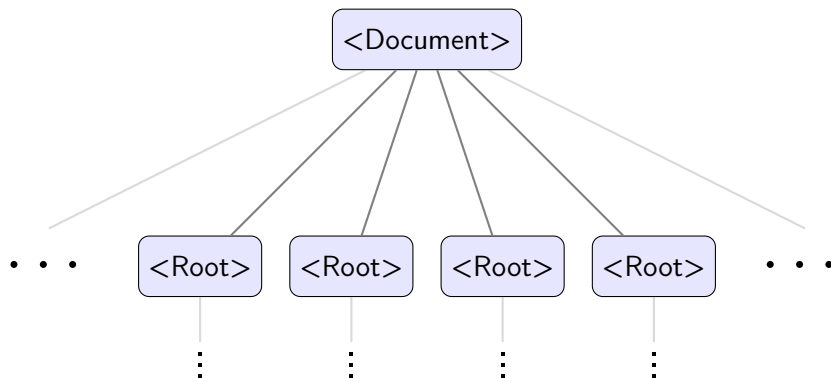
- Dependency parsing - závislosti mezi slovy ve větě.
- Využití knihovny UDPipe pro syntaktickou analýzu textu
- Vytvoření syntaktického stromu pro každou větu



## Umělé rozšíření stromu



- Rozšíření stromu pro podporu více vět bez nutnosti práce s lesem
- Pro zajištění dostatečné velikosti syntaktického stromu pro testování a kompresi







- Zaměření na gramatickou kompresi
- Komprimace pomocí linearizace – převod stromu na posloupnost uzlů a jejich následná komprese pomocí algoritmů pro kompresi textu



- Defakto převod stromu zpět na textovou reprezentaci
- Zvýšená redundance v důsledku zachování stromové struktury
- Řešeno hloubkovým průchodem – dosahoval nejlepších výsledků





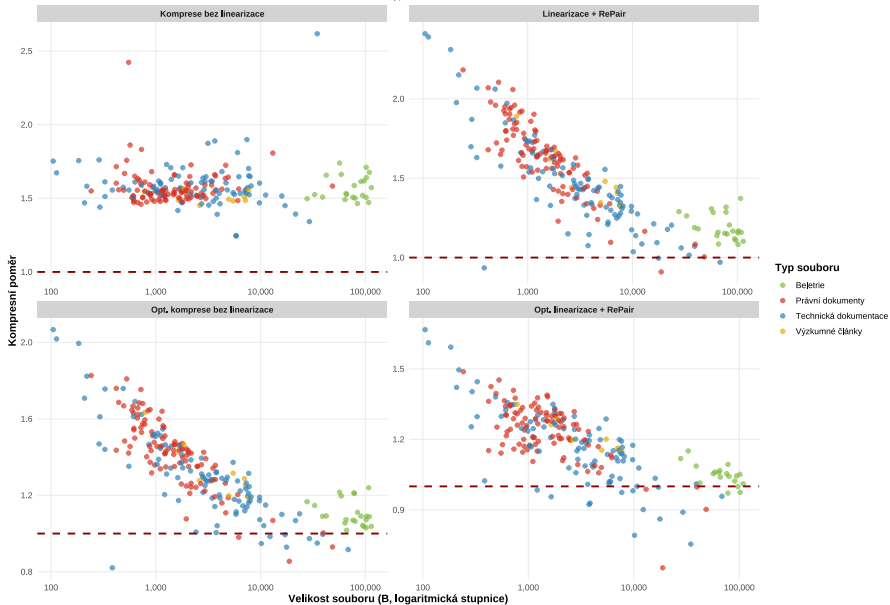
- 4 různé typy textu
- Celkový počet 200 textů
- Celkový objem 10 MB
- Všechny texty jsou v angličtině
- Texty jsou uloženy v textovém formátu

### Tabulka datasetu

Typ textu	Počet souborů	Rozsah velikostí
Beletrie	23	28 – 120 KB
Právní dokumenty	103	1 KB – 800 KB
Technická dokumentace	96	<1 KB – 100 KB
Vědecké články	20	1 KB – 100 KB

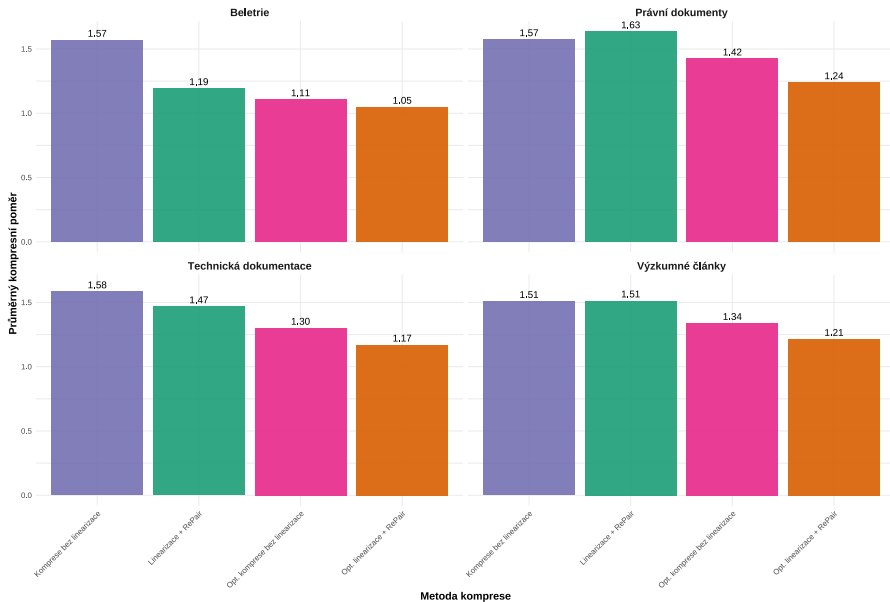
## Kompresní poměr podle velikosti souboru pro každou metodu zvlášť

Barva = typ souboru







## Srovnání kompresních poměrů stromových algoritmů

Průměrný kompresní poměr podle typu souboru a metody





-  Katja Filippova a Michael Strube. “Dependency tree based sentence compression”. In: *Proceedings of the Fifth International Natural Language Generation Conference*. 2008, s. 25–32.
-  Daniel Jurafsky a James H Martin. “Speech and Language Processing: An introduction to Natural Language Processing”. In: *Computational Linguistics, and Speech Recognition with Language Models. Third Edition draft* (2024).
-  Sandra Kübler, Ryan McDonald a Joakim Nivre. “Dependency parsing”. In: *Dependency parsing*. Springer, 2009, s. 11–20.
-  Colt McAnlis a Aleks Haecky. *Understanding compression. data compression for modern developers*. 1st Edition. Sebastopol, CA: O'Reilly, 2016. ISBN: 978-1-491-96153-7.



F. Oquendo, J. Leite a T. Batista. “Pipe-filter architectural style”. In: *Undergraduate Topics in Computer Science* (2016), s. 171–177. DOI: 10.1007/978-3-319-44339-3\_13.



Milan Straka a Jana Straková. “Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe”. In: *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada: Association for Computational Linguistics, srp. 2017, s. 88–99. URL: <http://www.aclweb.org/anthology/K/K17/K17-3009.pdf>.





Jana Straková, Milan Straka a Jan Hajič. “Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition”. In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore, Maryland: Association for Computational Linguistics, čvn. 2014, s. 13–18. URL: <http://www.aclweb.org/anthology/P/P14/P14-5003.pdf>.

Děkuji za pozornost

Marek Beran

VŠB – Technická univerzita Ostrava

marek.beran.st@vsb.cz

27. května 2025