

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 Převod textu do stromové struktury
- 3 Algoritmy
- 4 Testování
- 5 Literatura



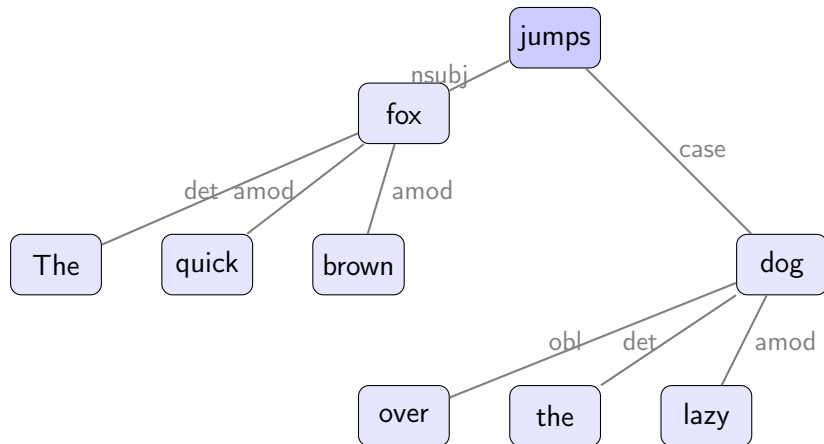
- Je možné efektivně komprimovat přirozený jazyk převedením do stromové struktury?
- Cíl: Proof of Concept



# Převod textu do stromové struktury



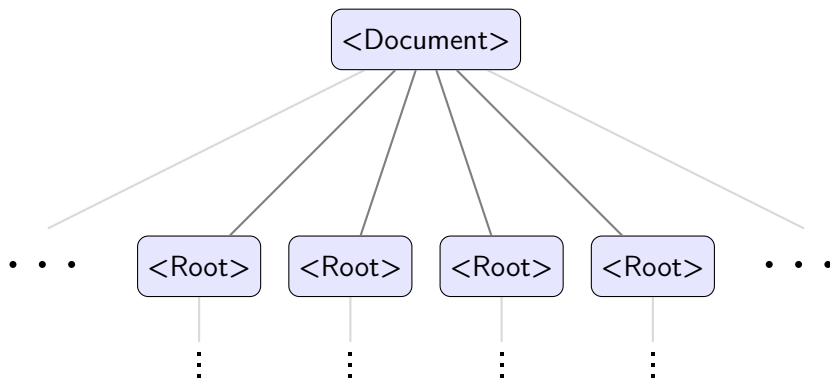
- Dependency parsing - závislosti mezi slovy ve větě.
- Využití knihovny UDPipe pro syntaktickou analýzu textu
- Vytvoření syntaktického stromu pro každou větu



## 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
- Zpočátku exaktní metody (generování hashů pro všechny podstromy)
- Komprimace pomocí linearizace – převod stromu na posloupnost uzlů a jejich následná komprese pomocí algoritmů pro kompresi textu
- Pokus o optimalizaci linearizovaného RePair (Recursive Pairing) algoritmu pro kompresi linearizovaných stromových struktur (maximální komprimace redundantních struktur) – hledání n-tic
- Komprese bez linearizace – algoritmus pro kompresi stromových struktur inspirovaný RePair algoritmem
- Opět pokus o optimalizaci





---

**Algorithm 1** Algoritmus pro kompresi bez linearizace

---

**Require:** Závislostní strom  $T$

- 1: Inicializuj čítač pravidel a slovník digramů
  - 2: Projdi strom a vytvoř index všech digramů (rodič, dítě, pozice)
  - 3: **while** existuje digram  $D$  s četností  $\geq 2$  **do**
  - 4:   Najdi digram  $D$  s nejvyšší četností
  - 5:   Vytvoř nový neterminál  $N_i$  a pravidlo  $N_i \rightarrow D$
  - 6:   Nahraď všechny výskyty digramu  $D$  neterminálem  $N_i$
  - 7:   Aktualizuj index digramů
  - 8:   Pokud komprese není efektivní, ukonči
  - 9: **end while**
  - 10: Odstraň nepoužitá pravidla
  - 11: **return** Komprimovaný strom  $T$  a pravidla gramatiky
-



### Metriky pro hodnocení podstromů:

- Četnost výskytu: počet identických instancí
- Velikost: počet uzlů v podstromu
- Kompresní zisk:  $(\text{velikost} \times \text{četnost}) - (\text{velikost} + \text{četnost})$  - vyjadřuje zisk z komprese – tzn. rozdíl mezi velikostí podstromu a velikostí gramatického pravidla

### Optimalizační techniky:

- Využití heuristiky pro výběr podstromů



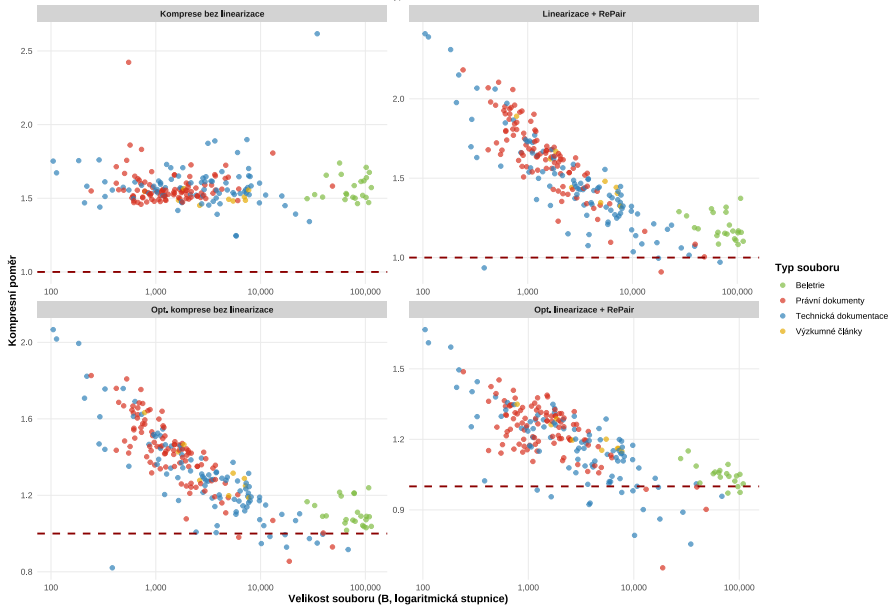
- 4 různé typy textu
- Celkový počet souborů: 242
- Celkový objem 10 MB
- Všechny texty jsou v angličtině

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

**Tabulka:** Typy textu a jejich velikosti

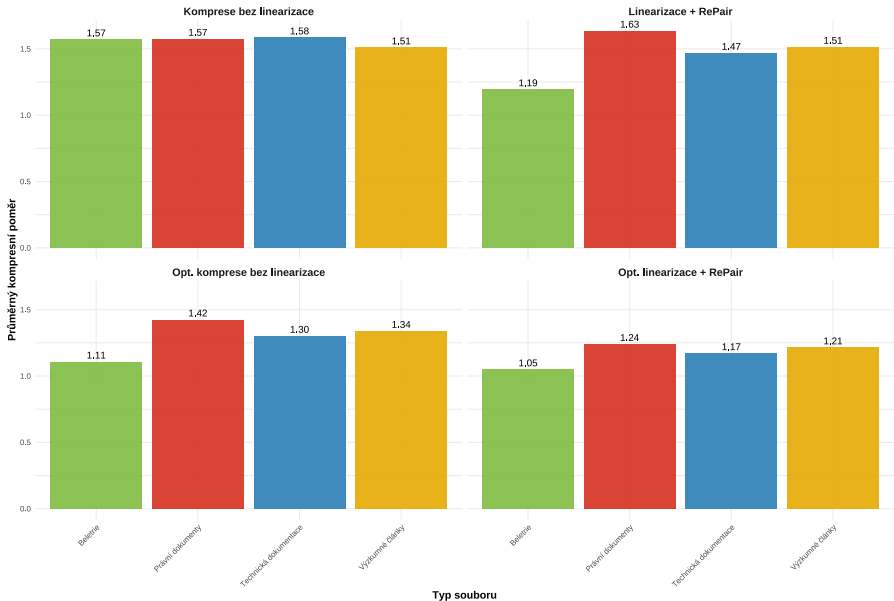
## Kompresní poměr podle velikosti souboru pro každou metodu zvlášť

Barva = typ souboru







## Srovnání kompresních poměrů stromových algoritmů

Průměrný kompresní poměr podle metody a typu souboru





-  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.

Děkuji za pozornost

Marek Beran

VŠB – Technická univerzita Ostrava

marek.beran.st@vsb.cz

27. května 2025