

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)

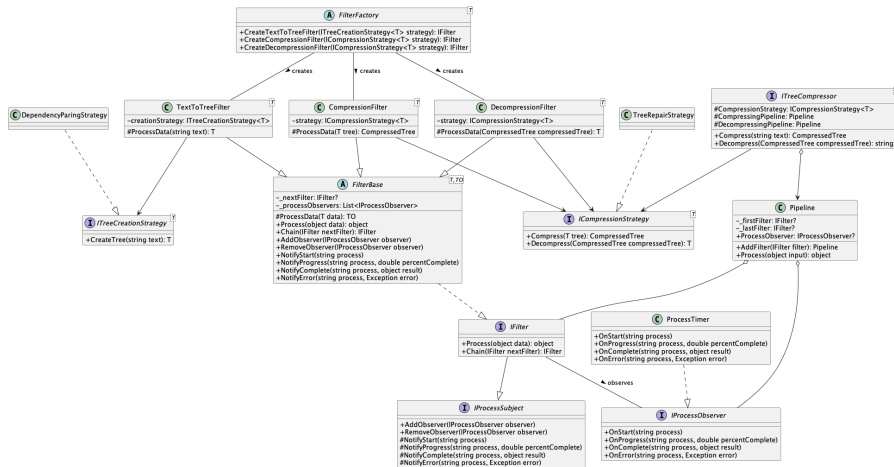
## Další nástroje:

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

## Bindings:

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

## Implementace knihovny

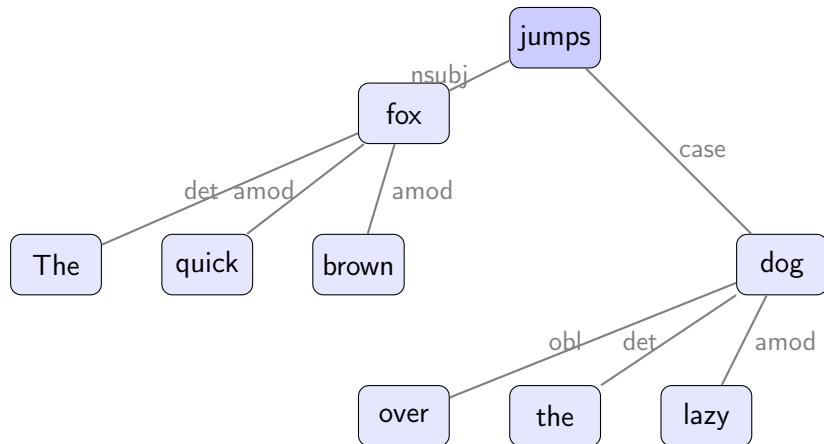


Obrázek: Třídní diagram části implementace zaměřený na řetězení filtrů

# Převod textu do stromové struktury



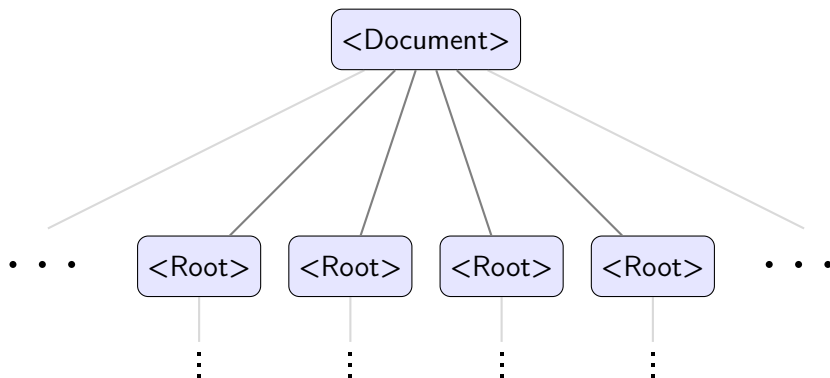
- Dependency parsing - závislosti mezi slovy ve větě.
- Využití knihovny UDPipe pro syntaktickou analýzu textu
- Vytvoření syntaktického stromu pro každou větu



## 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
- TreeRePair – algoritmus pro kompresi stromových struktur inspirovaný RePair algoritmem
- Opět pokus o optimalizaci TreeRePair algoritmu pro kompresi závislostních stromů



## Typy průchodů stromem:

- Preorder
- Inorder
- Postorder
- Průchod do šířky

## Preorder linearizace:

- Nejlepší výsledky pro TreeRePair
- Asymptotická složitost:  $O(n)$



---

**Algorithm 1** TreeRePair – zjednodušený pseudokód

---

**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
-



- Celková časová složitost:
  - Dominantní částí je kompresní cyklus:  $O(k \cdot n)$
  - Vzhledem k tomu, že  $k$  je v implementaci omezeno konstantou `MAX_ITERATIONS(100)`, lze říci, že asymptotická složitost je  $O(n)$
  - Bez omezení na počet iterací je složitost  $O(n^2)$
- Prostorová složitost:
  - Slovník digramů:  $O(n)$  – v nejhorším případě máme  $O(n)$  unikátních digramů
  - Seznam pravidel gramatiky:  $O(k) = O(1)$  vzhledem ke konstantnímu omezení
  - Celková prostorová složitost:  $O(n)$



## 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
- Hloubka a vyváženost struktury

## Optimalizační techniky:

- Efektivní hašování podstromů
- Inkrementální aktualizace metrik – pouze pro změněné podstromy



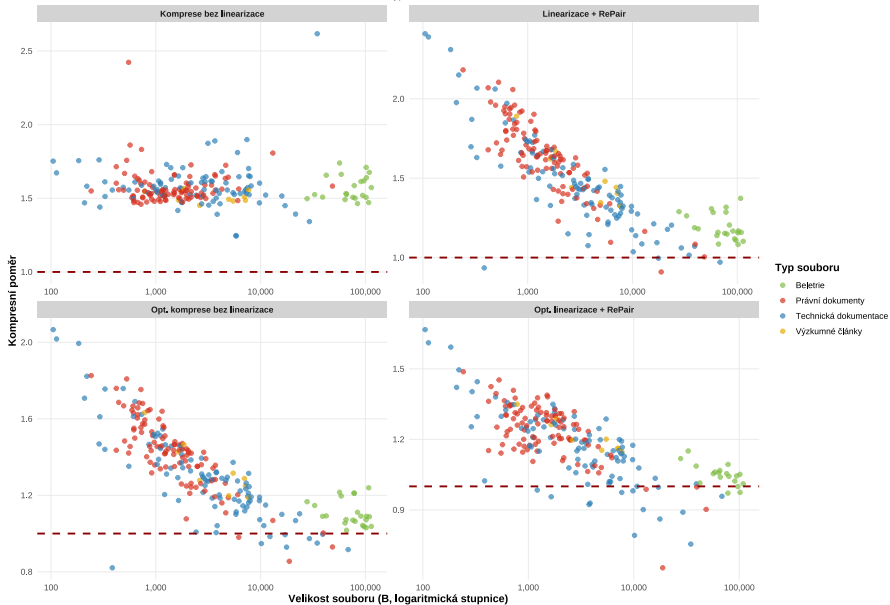
- 4 různé typy textu
- Celkový počet souborů: 242
- Celkový objem 10 MB
- Všechny texty jsou v angličtině

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

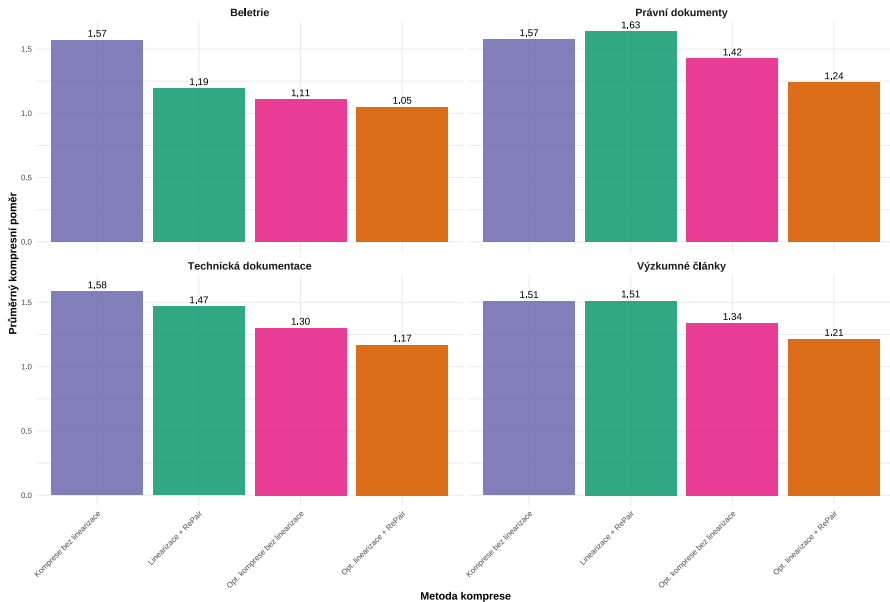
## Kompresní poměr podle velikosti souboru pro každou metodu zvlášť

Barva = typ souboru







## Srovnání kompresních poměrů stromových algoritmů

Průměrný kompresní poměr podle 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