Stylistika VII

ZS 2024

Obsah

- protokoly

Zadání

- vytvořte dva vzorky textů, každý o 7000 - 10000 slovech
- např. dva politici, dva básníci, dva prozaici, "seriozní" noviny vs. bulvár,
- pořádně popište, o jaké texty jde, odkud jste je získali atp.
  - web politiků, městská knihovna v Praze
- porovnejte tyto vzorky
- slovní tvary vs. lemmata
- klíčová slova
- kolokace klíčových slov, tematických slov, slov z trigramů… vyberte sami kriterium
- stylometrické indexy, které jsme probrali na přednáškách a v seminářích

- vytvořte dokument, který má následující strukturu
- cíl analýzy, pokud máte nějaké předpoklady, hypotézy atp., formulujte je
- popis jazykového materiálu - o co jde, zdroje atd.
- stručný popis metod (použijte odkazy na literaturu, jako v odborném textu)
- výsledky
- tabulkově, graficky
- závěry
- přiložit texty

- !! zpracujte úkol tak, aby byl replikovatelný

Vymezení problému

- Cílem analýzy bylo porovnat Babičku od Boženy Němcové a Kříž u potoka od Karoliny Světlé. Při jejich četbě jsem si všimla jistých podobností ve stylu psaní a chtěla jsem si ověřit, jestli je to jen můj dojem, nebo jestli to lze podložit daty. Obě autorky mají širokou slovní zásobu, proto předpokládám, že se tato skutečnost projeví na slovním bohatství jejich textů. Očekávám, že průměrná délka věty bude hodnotově podobná. Jelikož obě autorky tvořily ve stejné době, a obě jejich díla se odehrávají na venkově, nepředpokládám velké rozdíly ve výsledných hodnotách. Předpokládám, že texty budou spíše aktivní než deskriptivní.

Vymezení problému

„Pro stylometrickou analýzu byly vybrány dvě prózy Vladislava Vančury, které reprezentují odlišné žánry autorovy tvorby. Cílem bude porovnat stylometrické vlastnosti těchto textů (např. délka vět, aktivita/deskriptivita, lexikální rozmanitost – dopsat) a ověřit, zda charakteristiky odpovídají rozdílu mezi vybranými žánry. Prvním vybraným textem je pohádková kniha pro děti Kubula a Kuba Kubikula (1931). Vzhledem k rozsahu textu nebylo pracováno s celou knihou, nýbrž jen s její částí. Druhým textem je historická novela Markéta Lazarová (1931), která je charakteristická svým specifickým jazykem. L

Vymezení problému

„Analýza se pokusí vyvrátit hypotézu, kterou ve svém textu Pavoučí ženy a ubohá tlustá moucha definoval Jiří Peňás, jenž psaní rozlišuje na mužské a ženské. Právě ženské psaní je podle této hypotézy řízeno citově a emocionálně, nedosahuje proto takových kvalit jako řemeslné mužské psaní realizováno za účelem předání vyšší myšlenky . Cílem analýzy je stylometricky porovnat podobnost dvou textů, psaných ve stejném žánru na velmi podobné téma a z téže doby. Rozdílem je pohlaví autora.

Na základě této hypotézy bychom čekali u ženou psaného textu vyšší míru expresivních slov, což by se mělo projevit zejména v klíčových slovech. Vzájemné porovnání klíčových slov, ale i dalších indexů určujících slovní bohatství či tematickou koncentraci textu by pak mělo ukázat, zda jsou oba texty opravdu natolik odlišné, jako by dle Peňásovy teorie měly být.“

Vymezení problému - vhodné

„Cílem analýzy je porovnat stylometrické rozdíly mezi projevy Petra Fialy a Andreje Babiše v kontextu jejich vystoupení o vyslovení nedůvěře vládě. V rámci analýzy se zaměřím na strukturu vět, klíčová slova a jejich frekvenci, tematickou koncentraci, argumentační strategii a slovní zásobu.

Hypotéza vychází z předpokladu, že projevy Petra Fialy a Andreje Babiše se budou stylisticky lišit, přičemž každý z politiků využije odlišné jazykové prostředky a argumentační strategie v závislosti na svých politických cílech a zaměření na publikum. Z hlediska stylometrie lze očekávat, že Petr Fiala bude preferovat složitější a formálnější jazyk s vyšší mírou deskriptivity. Naopak projev Andreje Babiše bude pravděpodobně vykazovat vyšší míru aktivity. Jeho projev bude více úderný a bude užívat citově zabarvená slova a apely. Dále přepokládám, že rozdíl v jejich projevech se bude projevovat i v míře, do jaké se každý politik drží hlavního tématu. Očekávám, že Petr Fiala se bude více soustředit na konkrétní téma, zatímco Andrej Babiš bude mít tendenci češtěji odbíhat od hlavní problematiky.“

Vymezení problému – na co si dát pozor

Prvotním hybatelem tohoto semestrálního úkolu bylo jedno konkrétní heslo ve *Slovníku české literatury po roce 1945*. Heslo se týkalo Martina Friše, vcelku zapomenutého autora jediné knihy *Svědectví o deštivém odpoledni ztráveném v čekání*. Jelikož autorovo dílo obsahuje pouze jedinou experimentální prózu, zapomenutou stejně jako autor, lze nalézt jen málo textů, které by se k autorově osobnosti či jeho textu vyjadřovaly. V tomto malém zlomku lze ale nalézt článek Veroniky Košnárové zabývající se opomenutými prozaickými debuty 60. let. Z tohoto článku vychází i již zmíněné slovníkové heslo o autorovi[1]. A jedno konkrétní tvrzení z této studie nám poslouží i jako hypotéza pro tuto práci: *„Jediná kniha Martina Friše Svědectví o deštivém odpoledni ztráveném v čekání bývá přiřazována k linii experimentálních próz v české literatuře šedesátých let; nejblíže má k textům Věry Linhartové (zvláště k souboru Dům daleko)“*.

- je třeba více referenčních textů, pro možnosti srovnání

Absence popisu metodologie

- nestačí jen odkaz na software

3. Popis metod

Pro výpočty a grafy jazykových a stylistických vlastností textu v této práci byly použity tyto programy: Kwords, Quitaup, LancsBox.

KWords je webová aplikace vyvinutá Českým národním korpusem pro analýzu textů. Díky této aplikaci můžeme identifikovat klíčová slova tím, že aplikace porovnává frekvenci slov v analyzovaném textu s referenčním korpusem[1].

QuitaUp je též webová aplikace vyvinutá Českým národním korpusem ve spolupráci s Ostravskou univerzitou, která slouží ke kvantitativní stylometrické analýze textů. Umožňuje vypočítat různé stylometrické indexy jako například tematickou koncentraci či aktivitu textu[2].

LancsBox je softwar vyvinutý na Lancaster University pro analýzu jazykových dat a korpusů. Umožňuje pracovat s vlastními daty nebo existujícími korpusy, vizualizovat jazyková data a automaticky anotovat texty pro slovní druhy[3].

Absence popisu metodologie

Klíčová slova

Při analýze klíčových slov jsem použila aplikaci v jazykovém korpusu **KWords**. Program vypočítává klíčová prostřednictvím statistického testu, tedy srovnává relativní frekvenci každého slova v textu (úryvku) s relativní frekvenci stejného slova v celém referenčním korpusu.

25 klíčových slov – Zbabělci

lemma – základní tvar slova	textRelFq – relativní frekvence v textu	refRelFq – relativní frekvence v korpusu	statValue – statistická hodnota	effectSize – síla efektu
Harýk	3132,341	0	518,669	100
Catse	559,347	0	82,432	100
wirtemberský	894,955	0	139,066	100
kuci	559,347	0,008	78,935	100
notak	447,477	0,008	60,533	100

- nikde v textu není vysvětleno, co znamená statistická hodnota a co je síla efektu

Metodologie

- ideálně popsat každý způsob měření
- odkazy na literatur
- vzorce

adverbia a nominalizovaná adjektiva. Tato slova slouží k vyjádření vlastností, kvality nebo stavu, čímž text získává větší míru popisnosti a statičnosti. (Čech, 2014, s. 52)

Aplikace Českého národního korpusu QuitaUp, kterou ve své analýze využívám, pracuje pouze s verby (V) a adjektivy (A). To znamená, že se míra dějovosti hodnotí na základě výskytu verb a míra popisnosti na základě adjektiv. Celková aktivita textu (Q) je tedy definována jako poměr verb k součtu verb a adjektiv.

$$Q = \frac{V}{V + A}$$

Tabulka QuitaUp

- prezentovat jen to, co se interpretuje
- někteří mají v metodách popsáno i to, co se nepoužívá (např. h-bod)... proč? protože je to v manuálu...?

Slovní bohatství

- objevily se příklady, kdy se interpretovaly výsledky TTR a entropie u různě dlouhých textů…
- je třeba si dávat pozor na vliv délek na jednotlivé hodnoty, viz literatura

Klíčová slova

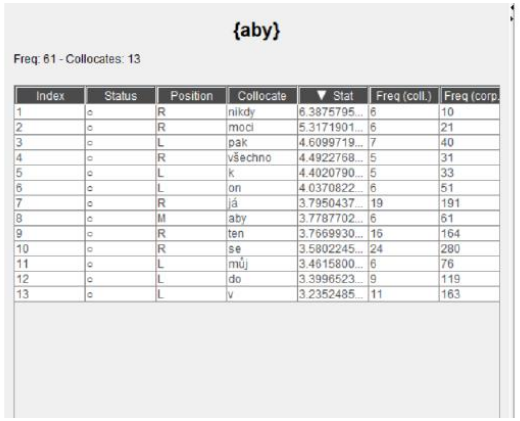
- v textu z literatury uvedeno, jak se měří (např. i statistické testy), ale ve výsledcích uspořádáno podle defaultního nastavení softwaru…
- interpretace vzhledem k vysvětleným způsobům měření

    - ukázka

Kolokace

- nestačí jen dát obrázky a napsat, že se tam objevují ta a ta slova
- pokusit se o interpretaci

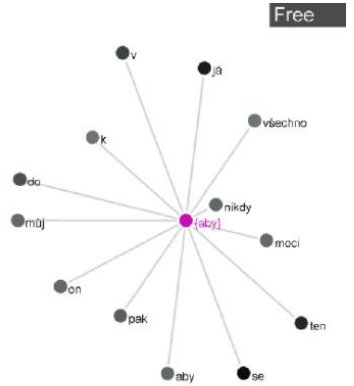
- taky popsat, co obrázky znamenají
  - vzdálenost
  - sytost barev
  - … a to použít pro interpretaci
  - + odkaz na hodnoty v tabulkách



{aby}

Freq: 61 - Collocates: 13

Index	Status	Position	Collocate	▼ Stat	Freq (coll.)	Freq (corp.
1	c	R	nikdy	6.3875795...	6	10
2	c	R	moci	5.3171901...	6	21
3	c	L	pak	4.6099719...	7	40
4	c	R	všechno	4.4922768...	5	31
5	c	L	k	4.4020790...	5	33
6	c	L	on	4.0370822...	6	51
7	c	R	já	3.7950437...	19	191
8	c	M	aby	3.7787702...	6	61
9	c	R	ten	3.7559930...	16	164
10	c	R	se	3.5802245...	24	280
11	c	L	můj	3.4615800...	6	76
12	c	L	do	3.3996523...	9	119
13	c	L	v	3.2352485...	11	163



Desetinná čísla

L věty	poměr hapaxů k tokenům	ATL
18,0328283	0,18792886	4,399
23,0876623	0,18970609	4,344

Slovní tvar vs. lemma

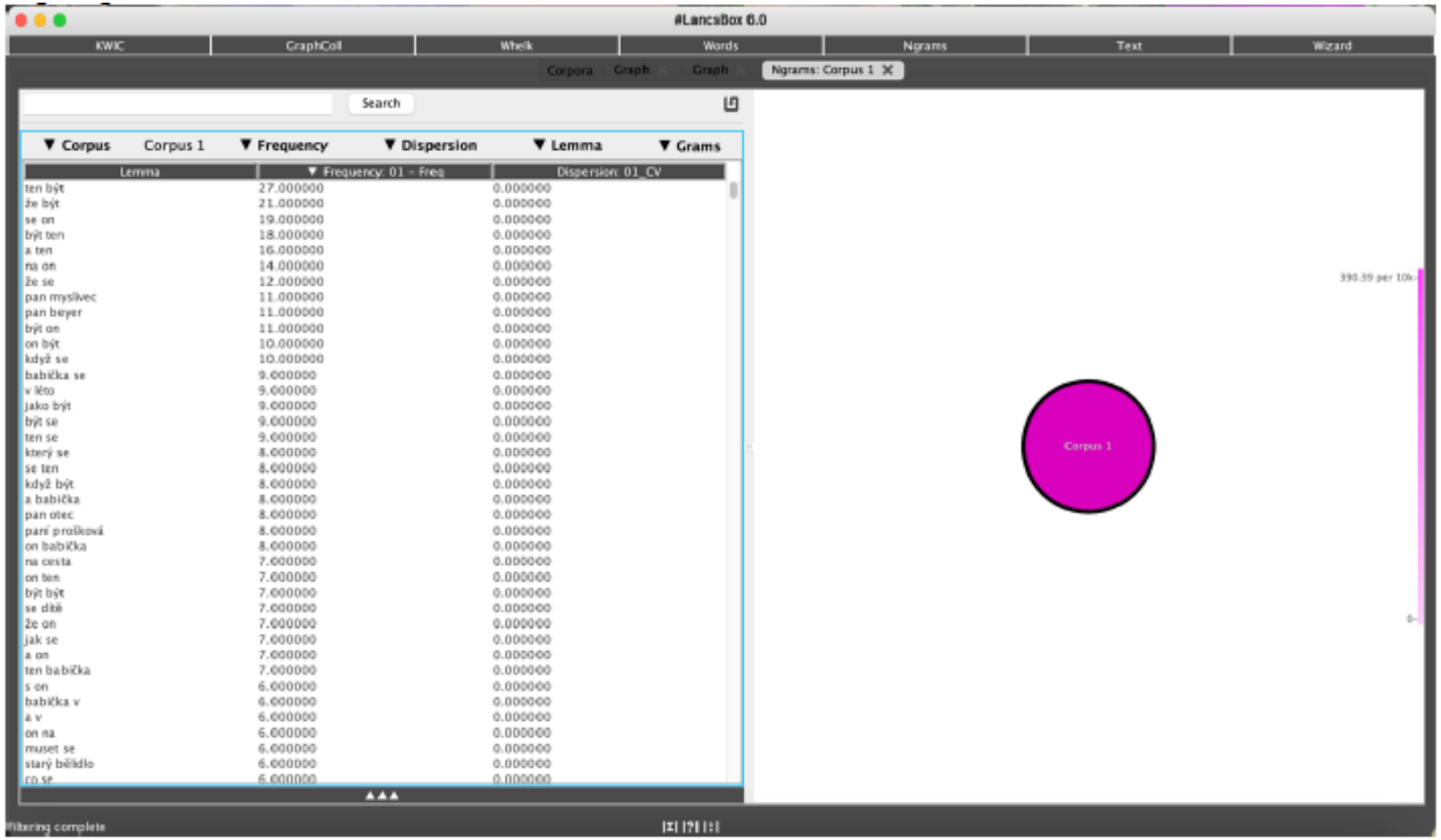
- volba jednotky ovlivňuje výsledky a interpretaci
- jasně uvést, s čím se pracuje a proč

- objevily se případy, kdy KWords prostřednictvím slovních tvarů, tematická slova jako lemmata a pak se to porovnávalo…

- nastavení KWords & QuitaUp

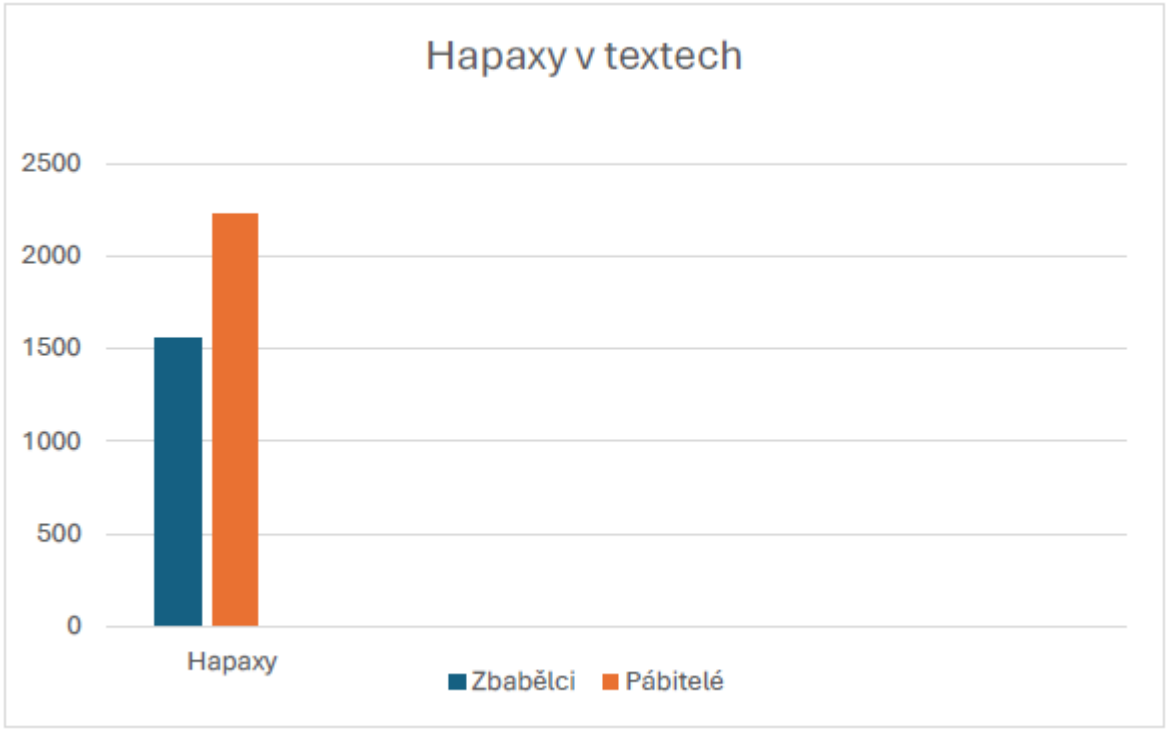
Nevhodné printscreeny

- např. bigramy



Formální úpravy

- nespoléhat na nevhodné standradní nastavení Excelu



Formální úpravy

- seznam literatury
  - formát uspořádání
    - abecedně, číslovaně…
    - jednotně!

- Ústav Českého národního korpusu. *KWords*. *Příručka ČNK*. [online]. [cit. 2024-11-21] Dostupné z: https://wiki.korpus.cz/doku.php/manualy:kwords.
- CVRČEK, Václav, ČECH, Radek a KUBÁT, Miroslav. *QuitaUp – nástroj pro kvantitativní stylometrickou analýzu*. Praha: Český národní korpus a Ostravská univerzita, 2020. Dostupné z: https://korpus.cz/quitaup/.
- BREZINA, Václav a PLATT, William. #LancsBox X [software]. Lancaster: Lancaster University, 2024. Dostupné z: http://lancsbox.lancs.ac.uk.

Formální úpravy

- seznam literatury

11. Seznam literatury

1) Křen, M. Kolokační míry a čeština: srovnání na datech Českého národního korpusu. In Čermák, F. & M. Šulc (eds.), *Kolokace*, 2006, 223–248.
2) *Klíčové slovo:* Jan Štráfelda. 2020. Dostupné z: https://www.strafelda.cz/klicove-slovo
3) Radek Čech, Miroslav Kubát (2017): TEMATICKÁ KONCENTRACE TEXTU. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.
   URL: https://www.czechency.org/slovnik/TEMATICKÁ KONCENTRACE TEXTU(poslední přístup: 21. 11. 2024)
4) MCINTYRE, Dan a Brian WALKER. *CORPUS STYLISTICS THEORY AND PRACTICE*. Edinburgh University Press, 2019. ISBN 978 1 4744 1322 0.

Formální úpravy

- seznam literatury

Použitá literatura

ČECH, Radek. *Tematická koncentrace textu v češtině*. Praha: Institute of Formal and Applied Linguistics, 2016. ISBN 978-80-88132-00-4.

ČECH, Radek, Ioan-Iovitz POPESCU a Gabriel ALTMANN. *Metody kvantitativní analýzy (nejen) básnických textů*. Olomouc: Univerzita Palackého, 2014. ISBN 978-80-244-4044-6.

František Čermák (1,2), Václav Cvrček (3) (2017): *Kolokace.* In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.
URL: https://www.czechency.org/slovnik/KOLOKACE (poslední přístup: 18. 11. 2024)

Asociační (kolokační) míry. *Český národní korpus* [online]. [cit. 2024-11-18]. Dostupné z: https://wiki.korpus.cz/doku.php/pojmy:asociacni_miry.

KWords. *Český národní korpus* [online]. [cit. 2024-11-18]. Dostupné z: https://wiki.korpus.cz/doku.php/manualy:kwords?s%5b%5d=din.

Michal Křen (2017): *Asociační míra*. In: Petr Karlík, Marek Nekula, Jana Pleskalová (eds.), CzechEncy - Nový encyklopedický slovník češtiny.
URL: https://www.czechency.org/slovnik/ASOCIAČNÍ MÍRA (poslední přístup: 18. 11. 2024)

KUBÁT, Miroslav. *Kvantitativní analýza žánrů*. 2016.

MCINTYRE, Dan a Brian WALKER. *Corpus Stylistics*. Edinburgh University Press, 2019. ISBN 978 1 4744 1323 7.

Formální úpravy

- seznam literatury

- zkuste ChatGPT :-)

Způsoby vyjádření – na co si dát pozor

- Máj je signifikantně kratší, což je třeba brát v potaz při zhodnocování výsledků, ale přesto jsem ho chtěla dát do porovnání dvou vrcholných děl.

Způsoby vyjádření – na co si dát pozor

- Máj je **signifikantně kratší**, což je třeba brát v potaz při zhodnocování výsledků, ale přesto jsem ho chtěla dát do porovnání dvou vrcholných děl.

Závěry

- je třeba si být vědom/vědoma limit kvantitativních analýz
- s ohledem na tyto limity pak přistupovat i k závěrům

- „Práce dokázala úspěšně vyvrátit argumentačně překonané tvrzení i z hlediska kvantitativní analýzy. A přinesla tak další důkaz o nepravdivosti a neúčinnosti dělení literatury na mužskou a ženskou.“
  - je to ale jen na 2 textech…

Závěry

- „Mé hypotézy jsem tedy potvrdila.“

- ve statistice se nikdy hypotézy nepotvrzují…

Závěry

- „Na základě provedené analýzu mohu konstatovat, že se potvrdila část mé hypotézy. Petr Fiala skutečně používá složitější a formálnější jazyk s vyšší mírou deskriptivity, zatímco Andrej Babiš se vyjadřuje jednodušším jazykem a častěji využívá citově zabarvená slova. Hypotéza se však nepotvrdila v části, která předpokládala, že se Andrej Babiš bude méně držet tématu než Petr Fiala. Hodnoty tematické koncentrace u obou projevů vyšly velmi podobné, což svědčí o tom, že se oba politici při svých projevech věnovali hlavním tématům srovnatelně.“

Struktura textu a jeho srozumitelnost

- „První nástroj najdeme na internetových stránkách Národního digitálního korpus (NDK), jedná se o aplikaci QuitaUp (QU). První měření ukázala, jak jsou si texty až překvapivě podobné. Oba jsou přibližně stejně dlouhé, výbor z Povídek malostranských (dále jen povídky) shromažďuje 10 109 tokenů. Token je grafické slovo, přesněji jeho realizace, která je oddělená mezerami v textu, většinou značena N. Výbor z dopisů (dále jen dopisy) má 9475 tokenů. Velmi blízký je také počet typů (type – typizovaný, abstraktní token; token vyjádřený jediným tvarem). Pro povídky je to 3766, pro dopisy 3693. Počet typů označuje množství různých jednotek v textu; type je značen V = vocabulary. Kdybychom zohlednili také pozdravy a loučení v dopisech, toto číslo by ještě narostlo. Našim cílem však bylo porovnat texty samy o sobě, bez formálních vycpávek. K tomu mimo jiné posloužilo srovnání průměrné délky vět. K této proměnné dospějeme jednoduše proveditelnou rovnicí: počet slov vydělíme počtem vět. U dopisů je průměrná délka 13,17 slov (719 vět), u povídek 13,96 (počet vět 724). Čísla jsou si opět blízká, ale z výše uvedených dat šlo těsný výsledek předpokládat. „

Struktura textu a jeho srozumitelnost

- „Aplikace QuitaUp spočítala, že jde o poměrně stejně dlouhé texty, ale ještě těsnější se ukázaly být průměrné délky vět. „

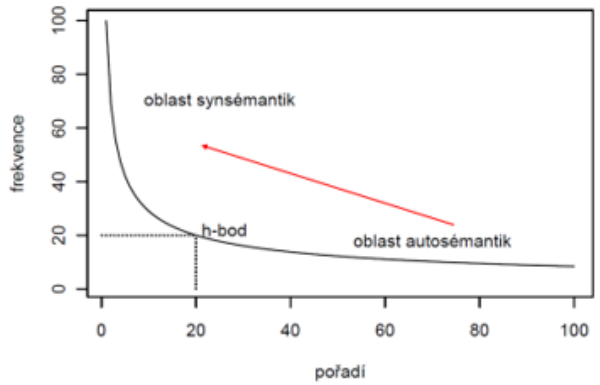
Obrázky, tabulky

- vždy popisky
- vždy na ně odkazovat v textu

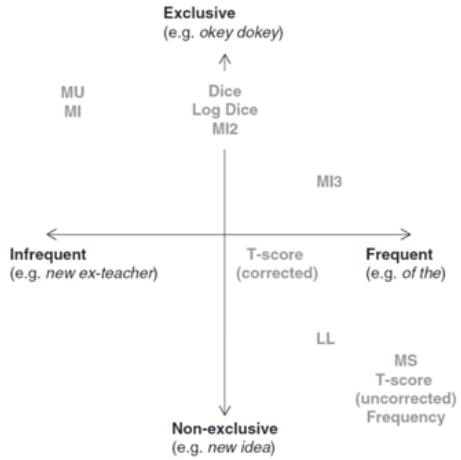
Obrázky – kopie

- pokud si „vypůjčíte“ graf, obrázek, vždy citujte odkud

Pro hledání tematických slov je tak nutné využít h-bod, „pro nějž platí, že rank = frekvence (např. 32. nejfrekventovanější slovo má frekvenci 32 výskytů). Všechna plnovýznamová slova nad tímto bodem (tj. v našem případě s frekvencí vyšší než 32)“[9] mají tematickou váhu závislou na vzdálenosti od h-bodu. Tematická koncentrace odpovídá pak součtu tematických vah jednotlivých slov. Sekundární tematickou koncentraci textu získáme, pokud h-bod vynásobíme 2.



„Podstatný rozdíl je např. mezi MI-score a T-score: zatímco MI-score nachází silné kolokace s velkou relativní frekvencí, a tedy spíše výjimečné až náhodné, T-score naopak kolokace nenáhodné, pravidelné a ustálené, ale nepříliš výrazné.“[13]



Ukázky prací

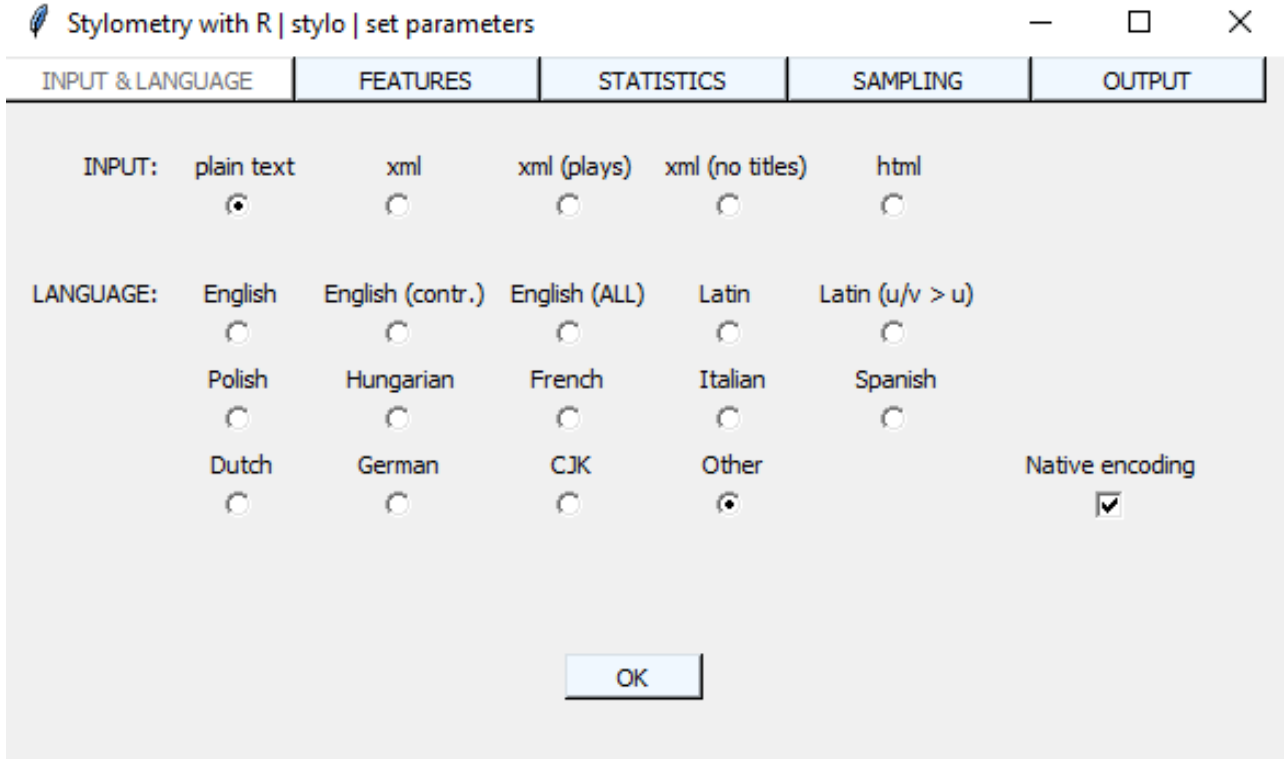
Stylo - postup

- R
  - https://www.r-project.org/
- RStudio
  - https://posit.co/download/rstudio-desktop/
- package stylo
  - Tools -> Install packages -> stylo

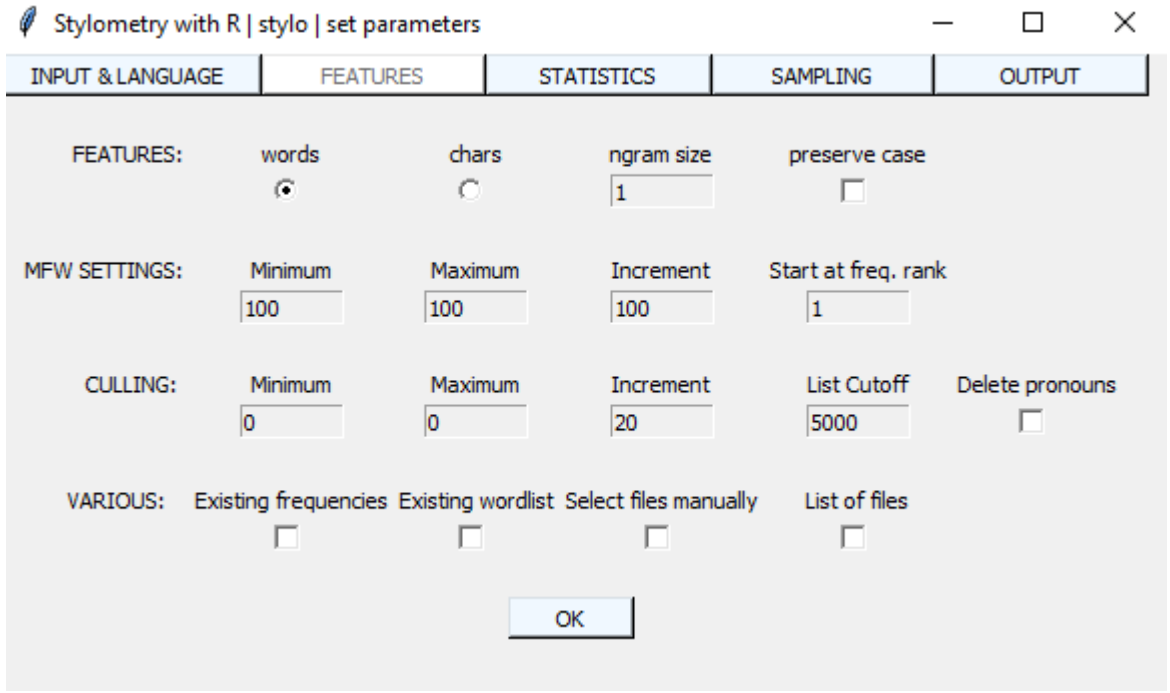
- library(stylo)
- stylo()

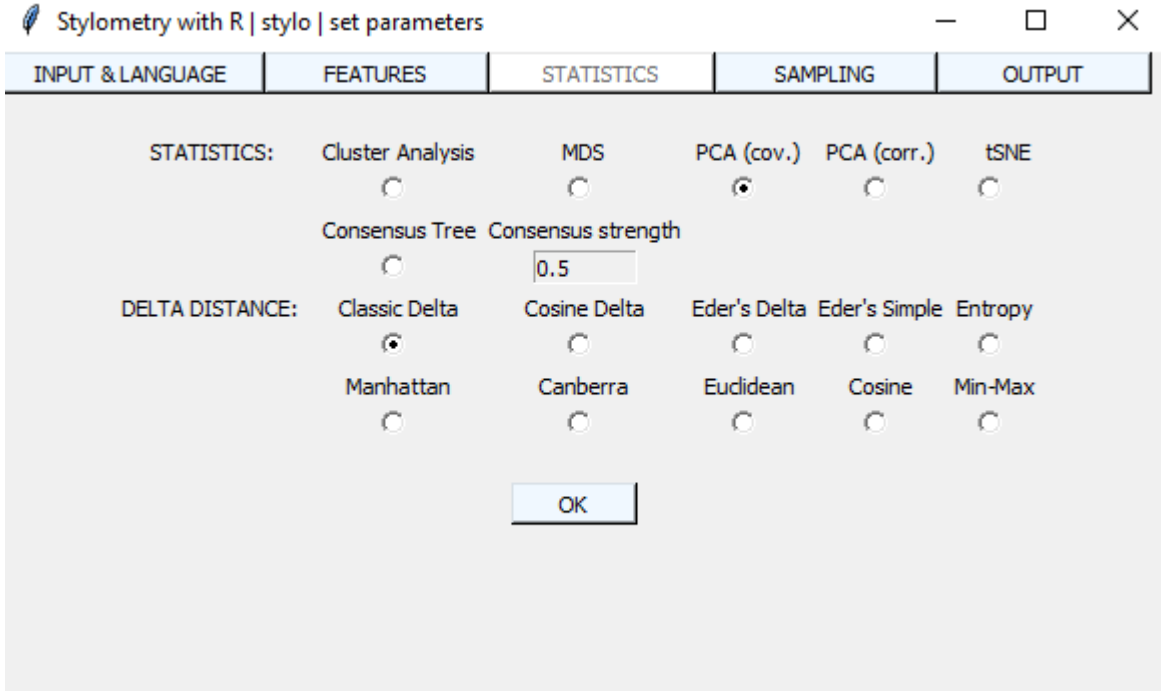
Stylo

- Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: a package for computational text analysis. *The R Journal*, *8*(1).



Stylo


Stylometry with R | stylo | set parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
FEATURES:
words
chars
ngram size
preserve case
1
MFW SETTINGS:
Minimum
Maximum
Increment
Start at freq. rank
100
100
100
1
CULLING:
Minimum
Maximum
Increment
List Cutoff
Delete pronouns
0
0
20
5000
VARIOUS:
Existing frequencies
Existing wordlist
Select files manually
List of files
OK


Stylometry with R | stylo | set parameters
INPUT & LANGUAGE
FEATURES
STATISTICS
SAMPLING
OUTPUT
STATISTICS:
Cluster Analysis
MDS
PCA (cov.)
PCA (corr.)
tSNE
Consensus Tree
Consensus strength
0.5
DELTA DISTANCE:
Classic Delta
Cosine Delta
Eder's Delta
Eder's Simple
Entropy
Manhattan
Canberra
Euclidean
Cosine
Min-Max
OK