Digital humanities

Cvičení: Analýza textů, měření příbuznosti jazyků

Jindřich Marek

1. Analýza textů (frekvence, wordcloud)

- morfologická analýza, lemmatizace
- vizuální znázornění textových dat

Přehled

- toto cvičení se zaměřuje na analýzu textu a zpracování přirozeného jazyka (NLP) v českých textech

- naučíme se pracovat s morfologickou analýzou, lemmatizací, tvorbou wordclouds

Cíle cvičení

- po dokončení tohoto cvičení budete schopni:
  - pochopit základy NLP (zpracování přirozeného jazyka)
  - používat nástroj Voyant Tools pro vizualizaci textu
  - provádět lemmatizaci českých textů
  - vytvářet mračna slov (wordclouds)

Požadované nástroje

- webový prohlížeč pro Voyant Tools

- JupyterLab (jako v předchozích cvičeních)

Dataset

- „velké“ projevy českých prezidentů
  - vánoční poselství prezidenta Miloše Zemana 2022 (poslední)
  - novoroční projev prezidenta Petra Pavla 2024 (první)

- formát: plné texty v češtině

Postup

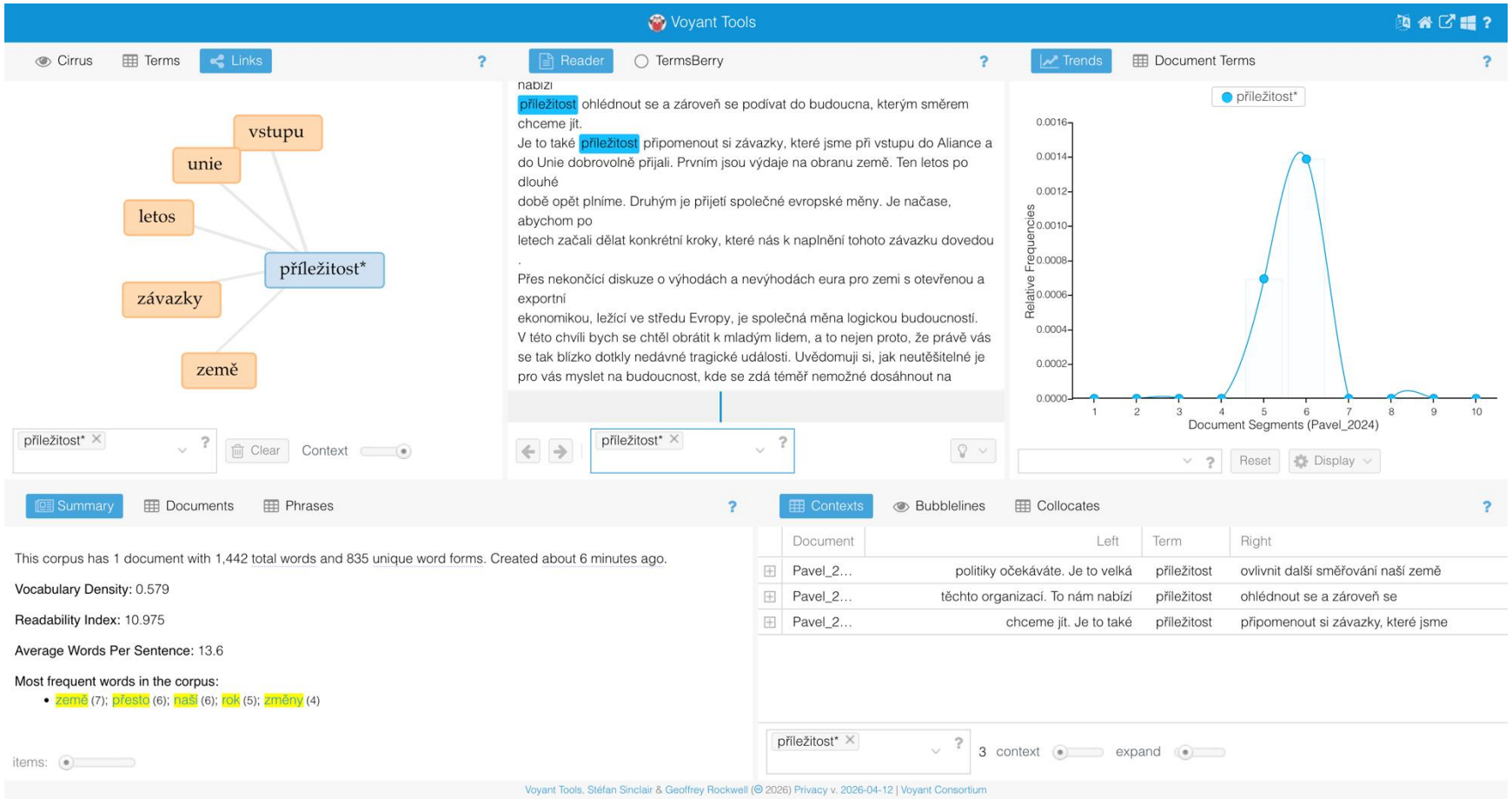
- porovnání slovní zásoby a témat

Rychlá vizualizace ve Voyant Tools

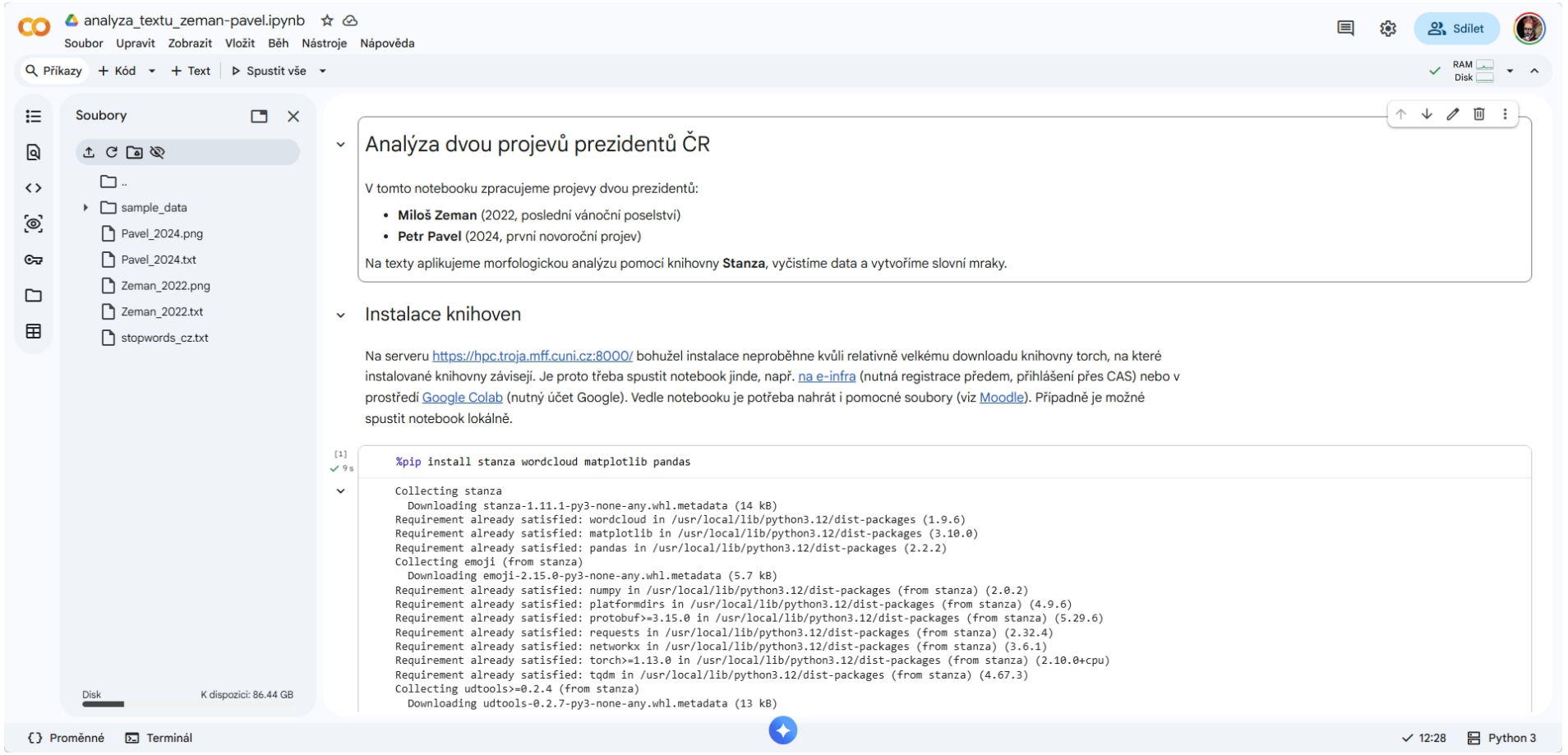
- https://voyant-tools.org
  - načtěte soubor s projevem
  - nahrajte stop-slova (do uživatelem definovaného seznamu)
  - prozkoumejte nástroje
    - Cirrus – slovní mrak
    - Terms – frekvenční seznam
    - Trends – rozložení slov v textu
    - Links – kolokace (která slova se vyskytují spolu)

- Která slova jsou nejfrekventovanější?
- Jak se liší oba projevy?
- Co Voyant Tools neukazuje?

Možný výsledek (Voyant Tools)



Notebook v prostředí Google Colab



Výsledek (Python, po lemmatizaci)

Miloš Zeman – vánoční poselství 2022

Petr Pavel – novoroční projev 2024

Limity frekvenční analýzy

- frekvence neznamená důležitost
  - častá slova nemusí být klíčová
- chybí kontext
  - nevíme, v jakém významu se slovo používá
- záleží na předzpracování
  - stop-slova, lemmatizace

Možná rozšíření

- kolokační analýza (jaká slova se vyskytují vedle sebe)
- analýza sentimentu (kladný vs. záporný tón)
- tematické modelování (LDA)

- porovnání v čase – jak se jazyk mění

Prompt pro jazykový model

Porovnej hlavní témata v plném textu následujících projevů. Výstupem mají být stručné body na jeden snímek v prezentačním programu (levý a pravý sloupec).

Miloš Zeman (2022, poslední vánoční poselství)

[plný text]

Petr Pavel (2024, první novoroční projev)

[plný text]

Odpověď (Gemma 4 26B A4B)

MILOŠ ZEMAN (2022)

BILANCE A OBHAJOBA TRADICE

• **Aktuální krize**: Agrese Ruska (podpora Ukrajiny vs. ekonomické zájmy), energetická krize a Green Deal.

• **Ekonomická kritika**: Boj proti inflaci, kritika vysokých úrokových sazeb a deficitu státního rozpočtu.

• **Historický regres**: Přehled politické kariéry, obhajoba reform (samospráva, armáda, banky) a úspěchů sociální demokracie.

• **Národní zájem**: Obrana suverenity ČR, ekonomická diplomacie (export) a úzký vztah s Izraelem.

• **Osobní rozloučení**: Reflexe politického života, vyjádření vděčnosti a odchod do důchodu.

PETR PAVEL (2024)

VIZE A ODPOVĚDNOST ZA BUDOUCNOST

• **Bezpečnost a trauma**: Reakce na tragickou událost (střelba), překonání strachu a posílení komunitního bezpečí.

• **Politická kultura**: Kritika vulgarizace diskuse, apel na profesionalitu a vyhýbání se populistickým lákadlům.

• **Státní instituce**: Podpora koalice, řešení dluhu, důležitost opozice a aktivní role prezidenta v obraně práv.

• **Evropský směr**: Posílení role v EU a NATO, plnění závazků (obrana) a podpora přijetí společné měny (euro).

• **Generační odpovědnost**: Podpora mladých lidí, řešení jejich problémů (bydlení, duševní zdraví) a udržitelnost pro děti.

2. Měření příbuznosti jazyků

- lexikostatistika
  - metoda pro měření příbuznosti jazyků na základě slovní zásoby.

- princip
  - vybereme základní slovní zásobu
    - Swadeshův seznam – slova jako já, ty, hlava, voda, oheň
  - porovnáme podobnost slov mezi jazyky
  - vytvoříme vizualizaci
    - např. dendrogram (strom podobnosti)

Přehled

- toto cvičení se zaměřuje na analýzu podobností mezi jazyky

Cíle cvičení

- po dokončení tohoto cvičení budete schopni:
  - měřit podobnost mezi jazyky pomocí lexikálních dat

Požadované nástroje

- JupyterLab (jako v předchozích cvičeních)

Dataset

- přehled domorodých jazyků jižní Ameriky
  - formát: tabulka se slovy specifickými pro jednotlivé jazyky

Postup

- seznam vybraných slov
- očištění slov na základní tvary
- analýza podobnosti slov
  - hierarchické shlukování
- vizualizace
  - dendrogram
  - matice podobnosti
  - zobrazení ve dvourozměrném prostoru

Tupíjské jazyky

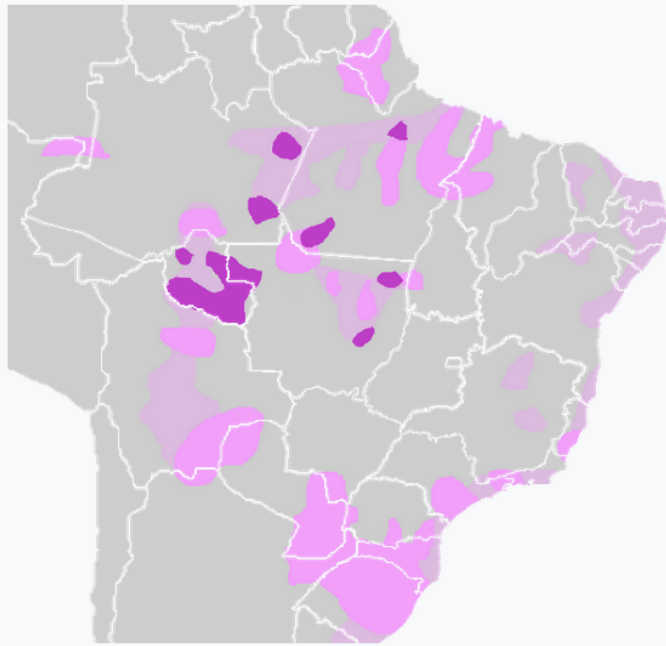
Tupíjské jazyky neboli *jazyková rodina tupí* jsou rodinou jazyků, kterými mluví domorodí obyvatelé Jižní Ameriky, konkrétně v Amazonii, na východním pobřeží Brazílie, v Peru, Bolívii, Paraguayi a Francouzské Guayaně.

Do rodiny se řadí okolo 70 jazyků. Nejvýznamnější skupinou tupíjské rodiny je skupina tupí-guaraní, kam patří guaranština, jíž mluví okolo 7 miliónů lidí a v Paraguayi a Bolívii požívá statusu úředního jazyka.[1] Dalšími tupíjskými jazyky mluví dnes daleko méně lidí: jazyk *cocama* v Peru má 19 000 mluvčích, jazykem *chiriguano* v Bolívii hovoří přibližně 20 000 lidí; ostatní jazyky mají tisíce či stovky mluvčích.[2]

V minulosti se některé tupíjské jazyky (respektive jejich zjednodušené formy) používaly v Brazílii jako lingua franca (všeobecný dorozumívací prostředek) mezi misionáři a domorodým obyvatelstvem.[1]

Jeden z vymřelých jazyků, *tupinambština*, je zdrojem mnoha slov z oblasti pojmenování jihoamerické fauny a flóry v portugalštině, odkud se tato slova šířila i do dalších evropských jazyků.[3] Z těchto slov v češtině používáme třeba slova jaguár či tamarín.[4]

Tupíjské jazyky



Území, kde se mluví či mluvilo tupíjskými jazyky

Rozšíření	Jižní Amerika
Počet mluvčích	*není znám*
Počet jazyků	70
ISO 639-2	tup

Čestmír Loukotka

Čestmír Loukotka (12. listopadu 1895 Chrášťany[1] – 13. dubna 1966 Praha) byl český lingvista a etnolog. Je představitel české etnolingvistiky a srovnávací lingvistiky indiánských jazyků. [2]

Život [editovat | editovat zdroj]

Absolvoval obchodní akademii. Před první válkou studoval na Sorbonně americkou etnologii u Paula Riveta. Pracoval jako úředník, teprve od roku 1950 pracoval vědecky v Náprstkově muzeu a poté v Ústavu pro etnografii a folkloristiku ČSAV, účastnil se i expedice v Brazílii.

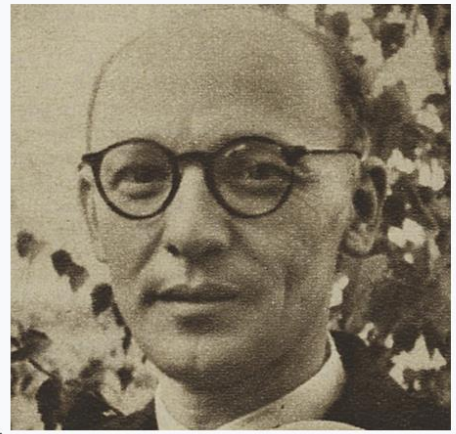
Jeho největší přínos spočívá v dodnes užívané klasifikaci indiánských jazyků Jižní Ameriky na lexikálním základě. Metody klasifikace jazyků uplatnil také pro papuánské jazyky. Vydal knihu o vývoji písma.[3]

Jeho dcera byla prozaička a překladatelka Jarmila Loukotková (1923–2007).

Dílo [editovat | editovat zdroj]

- *Roztřídění jihoamerických jazyků*, Praha, 1935, též španělsky jako *Clasificación de las lenguas sudamericanas*, Praha, 1935 a posmrtně anglicky jako *Classiffication of South American Indian Languages*, Berkeley, 1968.
- *Vývoj písma*, Praha, 1946.
- Classification des langues papoues, *Lingua Posnaniensis* 6 (Poznań), s. 19–83.

Čestmír Loukotka



Čestmír Loukotka (Pražský illustrovaný zpravodaj, 1941)

Narození	12. listopadu 1895 Chrášťany Rakousko-Uhersko
Úmrtí	13. dubna 1966 (ve věku 70 let) Praha Československo
Povolání	jazykovědec, antropolog a etnograf
Děti	• Jarmila Loukotková dcera

multimediální obsah na Commons

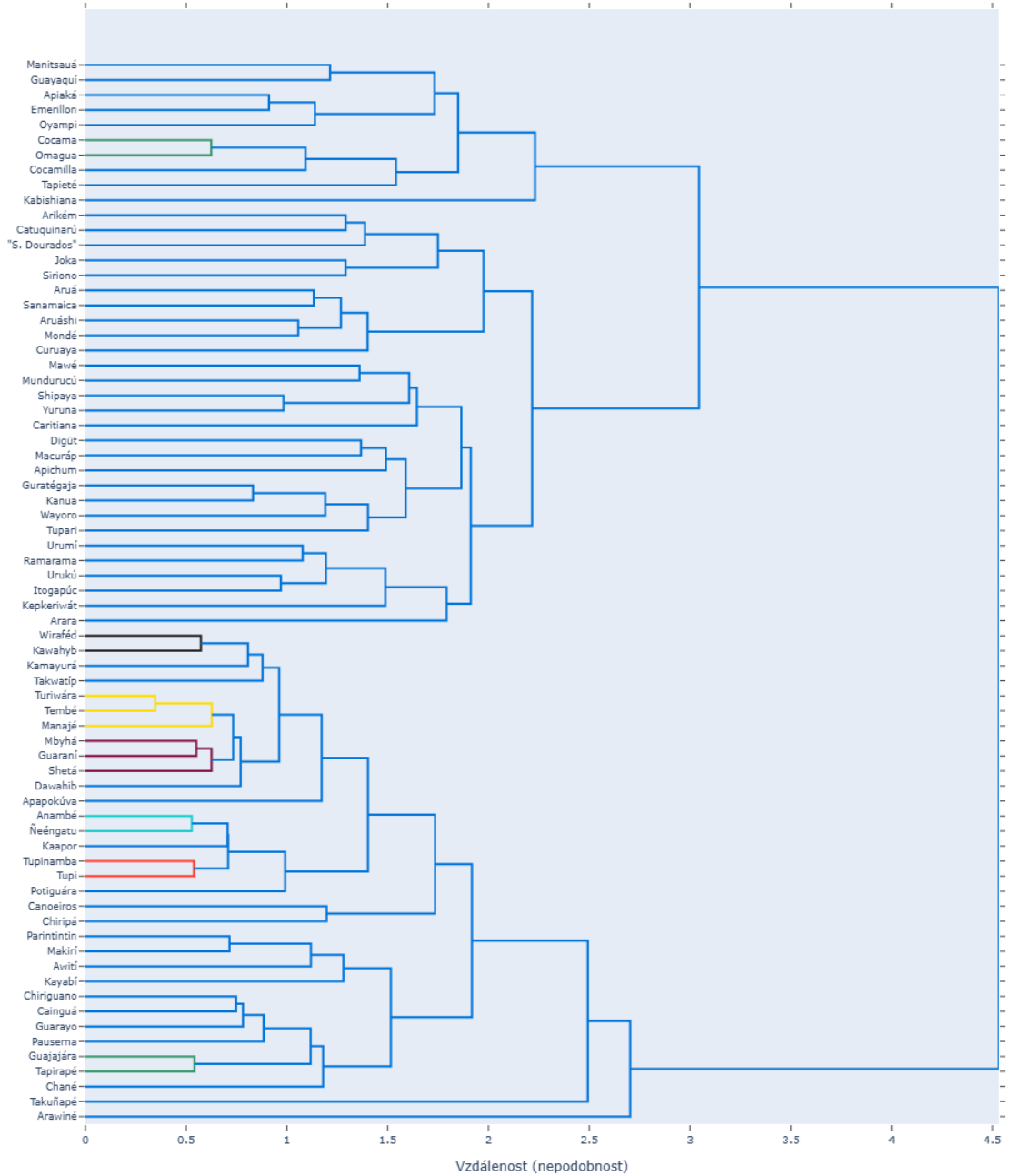
Některá data mohou pocházet z datové položky.

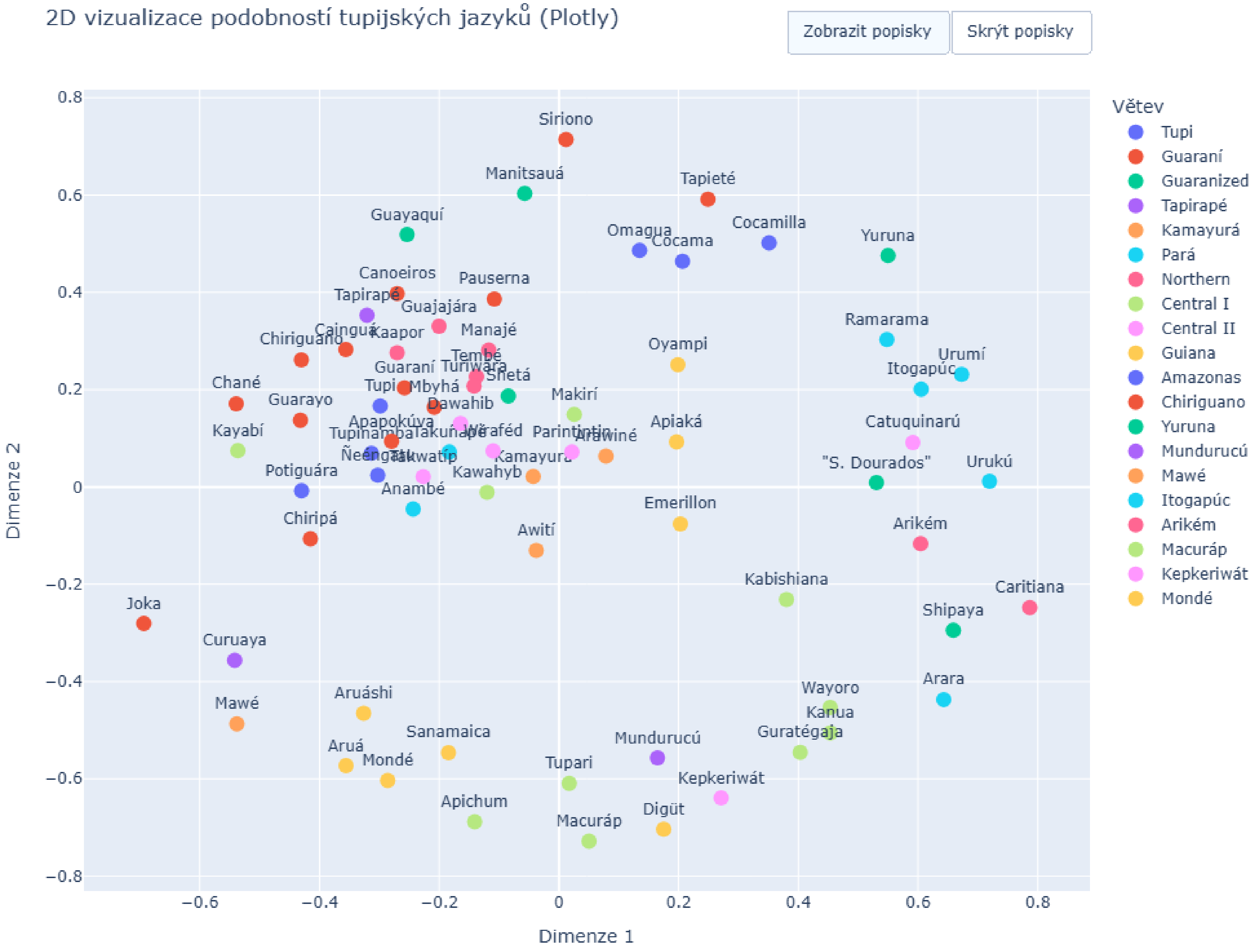
Vocabulary [edit]

Loukotka (1968) lists the following basic vocabulary items.[6]

Language	Branch	head	ear	tooth	hand	one	two	three	woman	water	fire	stone	maize	tapir
Tupi	Tupi	a-kang	nambi	táña	pó	peteĩ	mokoĩ	mbohapüi	kuñá	ü	tatá	itá	abai	tapüíra
Tupinamba	Tupi	a-kán	nambü	ráña	pó	angepé	mokoin	musaput	kuñá	ü	tatá	itá	auvati	tapirusu
Potiguára	Tupi	a-kanga	nambi	tanha	in-bó	oyepe	mokoy	mosapür	kuña	üü	tata:	ita:		
Ñeéngatu	Tupi	a-kanga	namü	taña	pó	yepé	mokoin	musapeire	kuñan	üüg	tatá	itá	auati	tapira
Guaraní	Guaraní	ãkan	nambi	apen-kun	pó	peteĩ	mokõi	mbohapy	kuñá	ü	tatá	itá	avatí	tapií
Apapokúva	Guaraní				pó	aépi	mokõi	moapi	kuña	ü	tatá			
Chiripá	Guaraní	rakã	nambi			aépi				ü	tata		avati	mborevi
Cainguá	Guaraní	aká	nambi	kú	pó	petein	mókoin	mbohapi	koñá	ü	tatá	itá	avachi	mborevi
Mbyá	Guaraní	che-ahká	chen-nambüh	che-rain	cheh-pó	peteí	mokoi	mboapü	kuña	ü	tatá	itá	avachi	tapií
Canoeiros	Guaraní	eaushmã			de-pó				uainvi	üg		itá	avashi	
Shetá	Guaranized	sh-aka	che-nambi	tienai	che-pó	matinkam	mokoi	ñiiru	kuñá	ü	tată	itá	avachi	tapi
"S. Dourados"	Guaranized	ñ-ãka	ela:me	nénai	e:-po	ua:i	mo:gai	mágatei	ko:ña	ho:ñe	agel'á	i:tá	nutya	tela:goi
Guayaquí	Guaranized	ni-aka	nambi	ã	i-pá	eteyã	meno	tanã	kuña	ü	dadá	itá	waté	mberevi
Tapirapé	Tapirapé	dzyane-akánga	dzyane-inamí	dzyane-roi	dzyane-pó	anchepé	mukúi	mãpít	kudzá	ü	tatá	itá	awachí	tapiíra
Kamayurá	Kamayurá	ye-akang	ye-nami	ye-nai	ye-po	yepete	mokoi	moapit	kuña	ü	tata	ita	avatsi	tapiít
Awití	Kamayurá	apot	inte-yambe	inte-ngu	i-po	mayepete	monkói	munitaruka	kuñá	ü	tara	ita	avachi	tapií
Arawiné	Kamayurá		ne-nami		ye-po									
Anambé	Pará	a-kánga	hä-nambi	se-raña	pó	yanäpo	mukuẽ	muhapi	kuña	ü	tata	ita	awat	tapiri

Hierarchické shlukování tupijských jazyků na základě lexikálních podobností (Plotly)




2D vizualizace podobností tupijských jazyků (Plotly)
Zobrazit popisky
Skrýt popisky
Dimenze 1
Dimenze 2
Větev
Tupi
Guaraní
Guaranized
Tapirapé
Kamayurá
Pará
Northern
Central I
Central II
Guiana
Amazonas
Chiriguano
Yuruna
Mundurucú
Mawé
Itogapúc
Arikém
Macuráp
Kepkeriwát
Mondé
Siriono
Manitsauá
Tapieté
Guayaquí
Omagua
Cocama
Cocamilla
Yuruna
Canoeiros
Pauserna
Tapirapé
Guajajára
Manajé
Chiriguano
Cainguá
Kaapor
Tembé
Guaraní
Turiwára
Shetá
Oyampi
Ramarama
Urumí
Itogapúc
Chané
Tupi
Mbyhá
Makirí
Guarayo
Dawahib
Apiaká
Catuquinarú
Kayabí
Apapokúva
Tupinambá
Takunapé
Wiraféd
Parintintin
Arawiné
Ñeengatú
Takwatip
Kamayurá
"S. Dourados"
Urukú
Potiguára
Kawahyb
Anambé
Emerillon
Chiripá
Awití
Arikém
Kabishiana
Caritiana
Joka
Shipaya
Curuaya
Arara
Wayoro
Aruáshi
Kanua
Mawé
Guratégaja
Sanamaica
Mundurucú
Aruá
Mondé
Tupari
Kepkeriwát
Apichum
Digüt
Macuráp
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−0.6
−0.4
−0.2
0
0.2
0.4
0.6
0.8

Limity měření podobnosti

- měříme podobnost znaků – co když se slova liší jen v jednom znaku, ale mají stejný původ?
- ignorujeme sémantiku – co když se význam slova v čase změnil?
- malý vzorek slov – stačí 13 slov pro spolehlivou klasifikaci?

Klíčové poznatky a otázky k diskusi

- klíčové poznatky
  - tupijské jazyky tvoří jasně oddělitelné větve
  - podobnost slovní zásoby koreluje s klasifikací lingvistů, ale metoda má své limity
- co ovlivňuje podobnost slovní zásoby kromě příbuznosti?
  - geografická blízkost
  - jazykový kontakt
  - doba oddělení
- možná rozšíření
  - použít větší Swadeshův seznam (100 nebo 200 slov)
  - zahrnout fonologickou podobnost
  - porovnat s výsledky komparativní metody