Kritická práce s daty 1

Radek Čech

Program

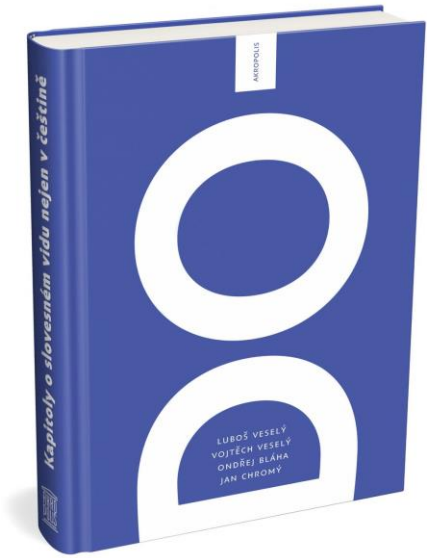
- data v lingvistice
- deskriptivní statistika
- hypotézy a jejich testování
- možnosti textových analýz

Data v lingvistice

- povaha lingvistických dat
  - k jakým datům mámě přístup?
  - na základě jakých metod je můžeme získat?

Data v lingvistice

- data získána metodami (přehledně u Chromý 2020)
  - introspektivními
  - korpusovými (textovými)
  - behaviorálními
  - neurozobrazovacími



Introspektivní metody

- co je introspekce?

Introspektivní metody

- osobní povědomí mluvčího
- analýza vlastních přesvědčení
- „nahlížení“ od vlastního vědomí

Introspektivní metody v lingvistice

- k čemu se používá?
- příklady?

Introspektivní metody v lingvistice

- dokumentace „typických“ příkladů
- gramatičnost

Dokumentace „typických“ příkladů

- stojí na předpokladu (často implicitním), že na základě introspekce se popisují vlastnosti jazyka jako objektivní reality
  - důvody?

Dokumentace „typických“ příkladů

- stojí na předpokladu (často implicitním), že na základě introspekce se popisují vlastnosti jazyka jako objektivní reality

- dichotomický pohled na jazyk
  - langue – parole
  - competence – performance

- normativní pohled na jazyk(?)

Dokumentace „typických“ příkladů

- problémy?

Dokumentace „typických“ příkladů

- problémy?

- interpersonální shoda (či její absence)
- vliv teorie, expertní zkušenosti
- otázka, co se takto vlastně popisuje
  - většinou se opírá o teoretické východisko
    - strukturalismus
    - generativní gramatika

Gramatičnost

- gramatičnost
  - inherentní vlastností jazykové struktury, tzn. věta (či jiný jazykový prostředek) je správný, je-li vytvořen v souladu s gramatickými pravidly daného jazyka

Hrál jsem dva roky na flétnami
Nerad se dívám televizi na
Pavel daroval mamince
Včera Petra česala dvě hodiny před plesem se
Včera Petra dvě hodiny se před plesem česala
Včera Petra česala se dvě hodiny před plesem

Gramatičnost

- gramatičnost
  - inherentní vlastností jazykové struktury, tzn. věta (či jiný jazykový prostředek) je správný, je-li vytvořen v souladu s gramatickými pravidly daného jazyka

**Hrál jsem dva roky na flétnami*

**Nerad se dívám televizi na*

**Pavel daroval mamince*

**Včera Petra česala dvě hodiny před plesem se*

?Včera Petra dvě hodiny se před plesem česala

?Včera Petra česala se dvě hodiny před plesem

Gramatičnost

- kategoriální, nebo škálová proměnná?

Gramatičnost

- kategoriální, nebo škálová proměnná?
- jak vyhodnocovat, pokud škála?

Gramatičnost & akceptabilita věty

- akceptabilita
  - termín nadřazený gramatičnosti, věta je akceptovatelná podle zvolených kritérií (mohou se někdy dokonce vylučovat): gramatičnost, sémantická a pragmatická adekvátnost. Do sféry pragmatické (komunikační) adekvátnosti náleží mnohá mimolingvistická kritéria

    #Petr má rád rajčata a maminku

    #Jaro uvařilo Měsíc a paneláky nachystaly botičky

- více viz heslo **gramatičnost** v NESČ
  - https://www.czechency.org/slovnik/GRAMATI%C4%8CNOST#akceptabilita

Introspektivní metody v lingvistice

(podle Chromý 2020)

- expertní introspekce

- laická introspekce

Expertní introspekce

- domnělá analýza „nezávislé“ entity

- problémy
  - potenciál absence interpersonální shody
  - shoda dána stejnými východisky

- introspekce zaměřená na analýzu užívaní jazyka

Expertní introspekce

- domnělá analýza „nezávislé“ entity
- problémy
  - potenciál absence interpersonální shody
  - shoda dána stejnými východisky
- introspekce zaměřená na analýzu užívaní jazyka
- introspekce jako prostředek formulace hypotéz

Laická introspekce

- většinou se zkoumá větší počet respondentů
- ideálně lidé neovlivněni teorií

Laická introspekce

- většinou se zkoumá větší počet respondentů
- ideálně lidé neovlivněni teorií

- výhody
  - zkoumání řídkých jevů
  - tzv. negativní evidence
  - analýza jazyka jako mentálního systému
    - srov. vliv dichotomií (langue-parol, kompetence-performace)
  - „eliminace“ nežádoucích projevů (přeřeknutí atp.)
    - souvisí s teoretickým rámcem

Laická introspekce

- pokud se zkoumá větší počet respondentů, jedná se de facto o **empirickou** analýzu
  - metodologické důsledky: výběr vzorku, povaha výzkumu, interpretace výsledků, statistika

Laická introspekce – ilustrativní příklad

Označte subjekt

Praha byla hlavní město ČSR

Hlavní město ČSR byla Praha

To byla škola

To byla hlavní města říše

Laická introspekce

- výsledky a důsledky takového přístupu (zpravidla)
  - absence kategorických výsledků → tendence
  - možnost detailnější analýzy výsledků
    - míra korelace různých způsobů vyhodnocení
    - faktory ((věk, pohlaví, vzdělání…)

Introspekce

- v psychologii od druhé pol. 19. stol.
  - vědomí jejích nedostatků

- introspekce jako předmět výzkumu v psychologii
  - nemožnost přesně nahlížet své vnitřní procesy
  - introspekce zřejmě není přímým odrazem mentálních procesů
  - při introspektivním hodnocení vliv implicitních apriorních předpokladů
  - rozdílnost postoje k vlastní a cizí introspekci
    - tendence vnímat vlastní jako spolehlivou, ale u druhých ne
  - vliv vnějšího prostředí, viz následující slide

Introspekce

- Meili et al. (1967, s. 130):

  „Srovnáme-li výpovědi p. o. [pokusných osob] získané při různých výzkumech, vycházejících z **divergentních teoretických** základních názorů, nelze popřít, že p. o. **hovoří řečí teorie**, která udává tón výzkumu, i když se mluví o **tzv. neovlivňujících pokynech**.“

Introspekce

- „With the **decline of structuralism**, introspective methods **lost favor** as a source of psychological data. Then, gone but not forgotten, introspection **re-emerged** in the 1960s with the rise of **cognitive science**, in which verbal protocols were a major source of data and a basis for much theorizing on problem solving.
  But despite their new popularity, introspective methods continued to exhibit **the same weakness** that had aroused critics in the structuralist period – **the lack** of effective means of obtaining **interpersonal agreement** among scientists on the interpretation of introspective data.“
  Estes (2000, s. 21)

Introspektivní metody v lingvistice - shrnutí

- limity
- i když nemůžeme jasně říct, že introspekce interpretuje kognitivní procesy „správně“, nemůžeme tvrdit ani opak
- experimentální výzkum tzv. laické introspekce
- obecně: introspekce často „prvním“ krokem analýz

Korpusové (textové) metody

- analýza pozorovatelného verbálního chování
  - usage-based models/grammars

Korpusové (textové) metody

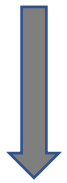
- analýza pozorovatelného verbálního chování
- co nám to vlastně umožňuje zkoumat?

Modely jazykového chování a jejich interpretace

jazykové chování
(texty, promluvy)

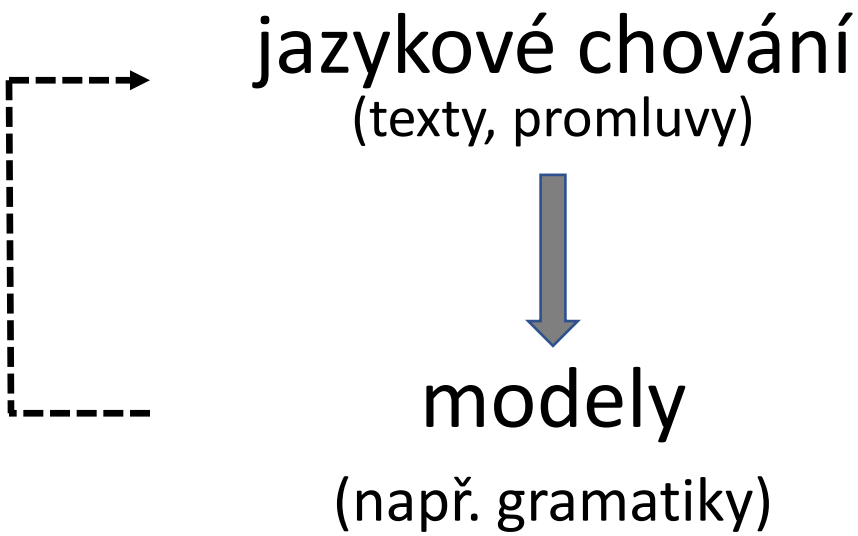
Modely jazykového chování a jejich interpretace

jazykové chování
(texty, promluvy)

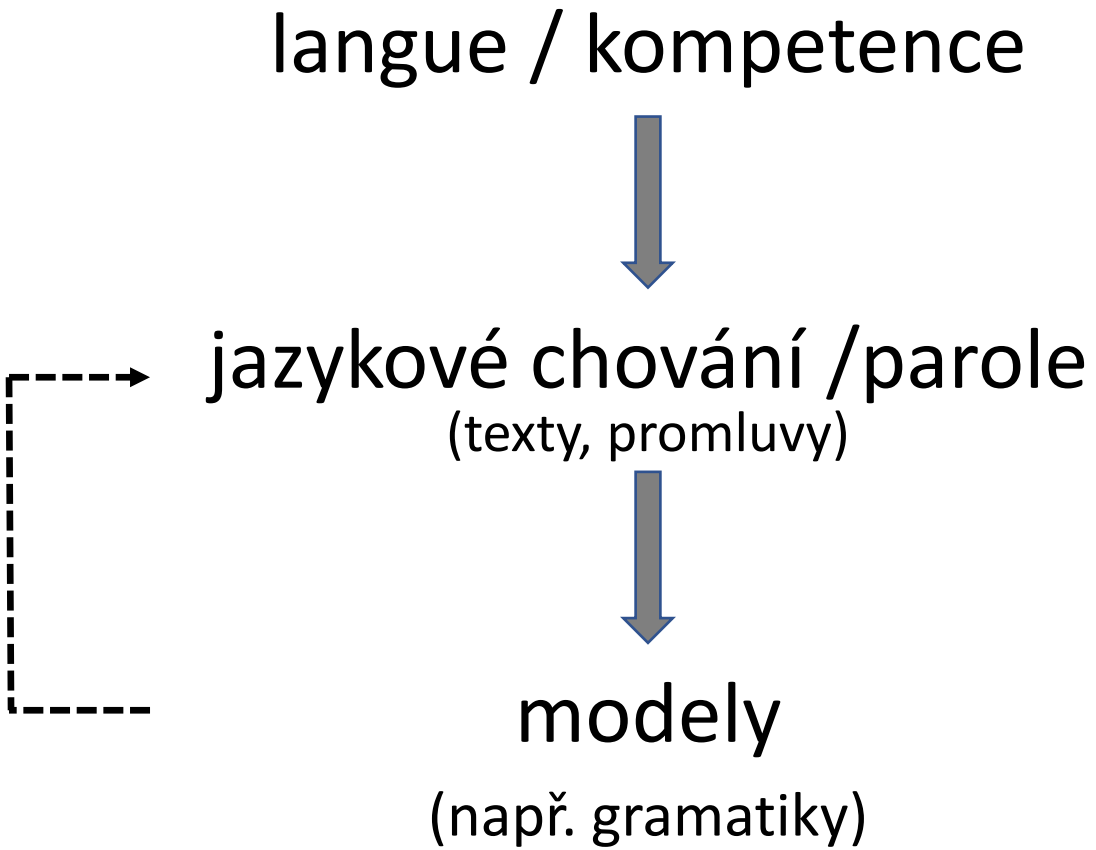


modely
(např. gramatiky)

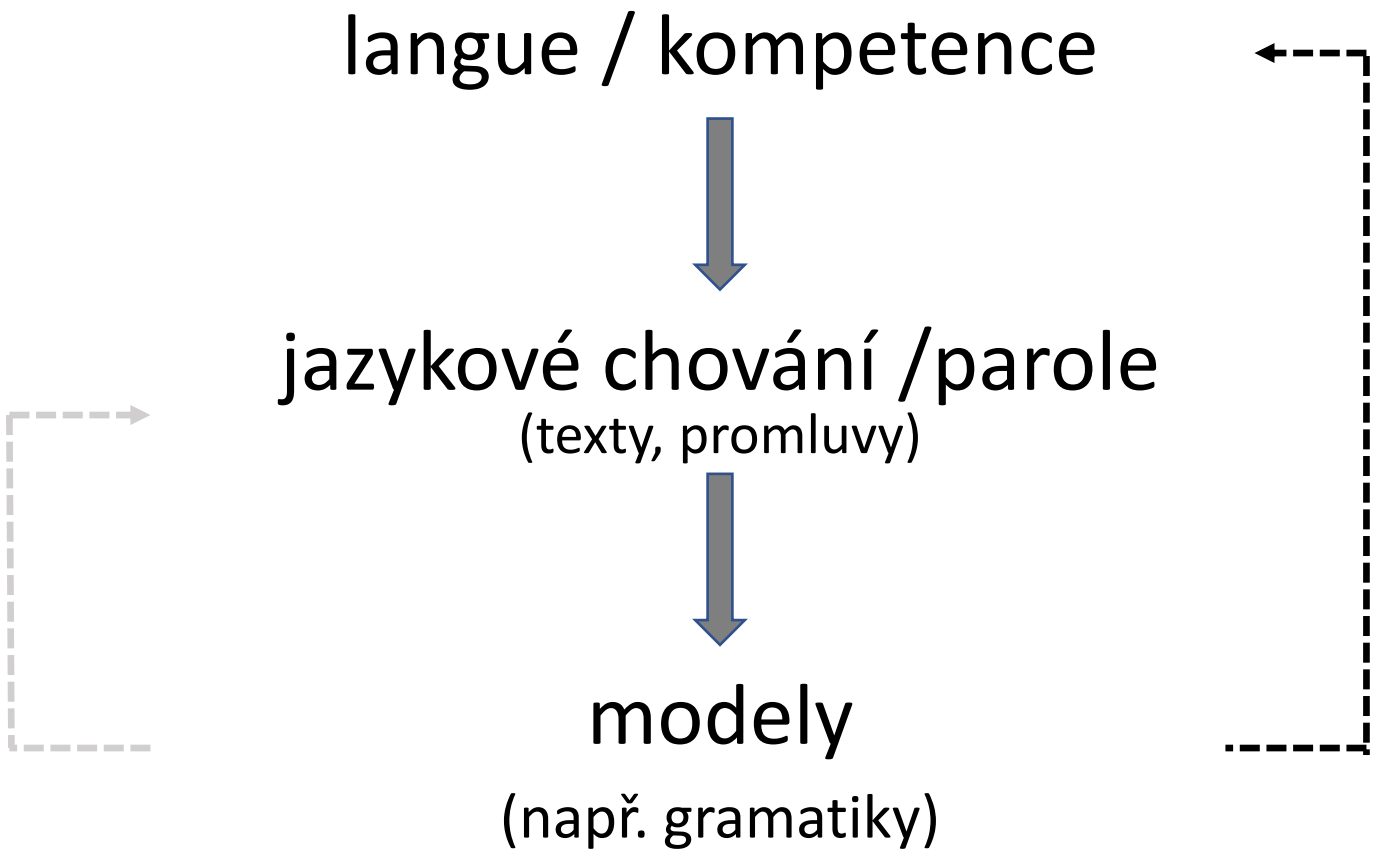
Modely jazykového chování a jejich interpretace



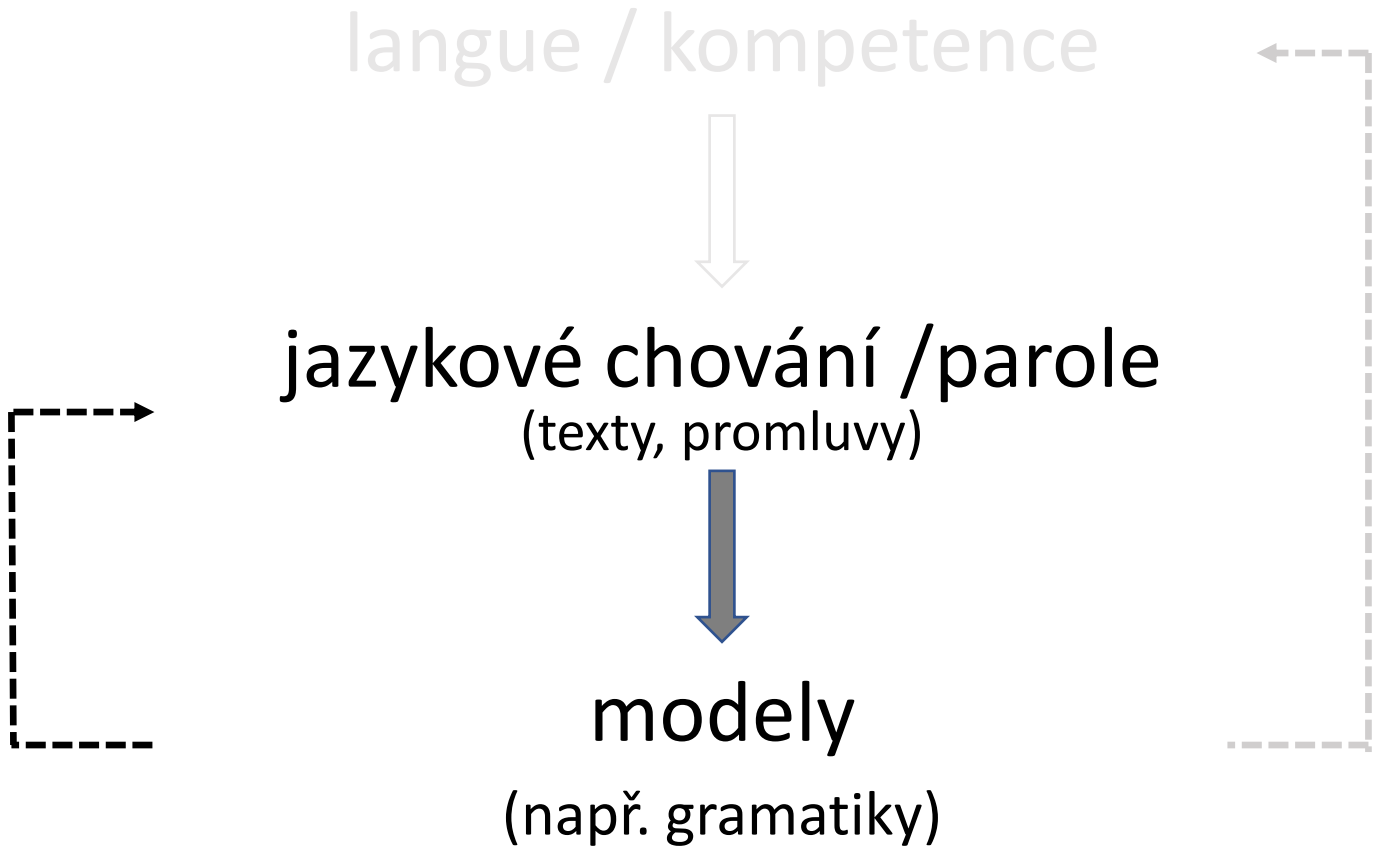
Modely jazykového chování a jejich interpretace



Modely jazykového chování a jejich interpretace



Modely jazykového chování a jejich interpretace



Modely jazykového chování a jejich interpretace

- jazykové chování
  - dynamika, „nestabilita“
  - náhodné fluktuace

Modely jazykového chování a jejich interpretace

- jazykové chování
  - dynamika, „nestabilita“
  - náhodné fluktuace
  - počínající tendence (srov. jazyková změna a její evoluce)

Korpusové (textové) metody

- výhody?

Korpusové (textové) metody

- výhody
  - nezávislost na postoji badatele
  - množství dat
  - rychlé zpracování
  - replikovatelnost

Korpusové (textové) metody

- výhody
  - nezávislost na postoji badatele
  - množství dat
  - rychlé zpracování
  - replikovatelnost
- nevýhody

Korpusové (textové) metody

- výhody
  - nezávislost na postoji badatele
  - množství dat
  - rychlé zpracování
  - replikovatelnost

- nevýhody
  - absence některých jevů
  - omezené možnosti sémantické analýzy

Korpusové (textové) metody

- korpus vs. text

- Čech, R., Kosek, P., Mačutek, J., Navrátilová, O. (2020). Proč (někdy) nemíchat texty aneb Text jako výchozí jednotka lingvistické analýzy. Naše řeč, 103, 24-36.

- blíže na některé z dalších přednášek

Behaviorální metody

- analýza chování **jedince** (resp. **skupiny lidí**)
- záměrné ovlivňování faktorů

- typicky psycholingvistika

Behaviorální metody

- velké množství metod

- měření reakční časů
- sledování očních pohybů
- elicitace jazykové produkce

Priming

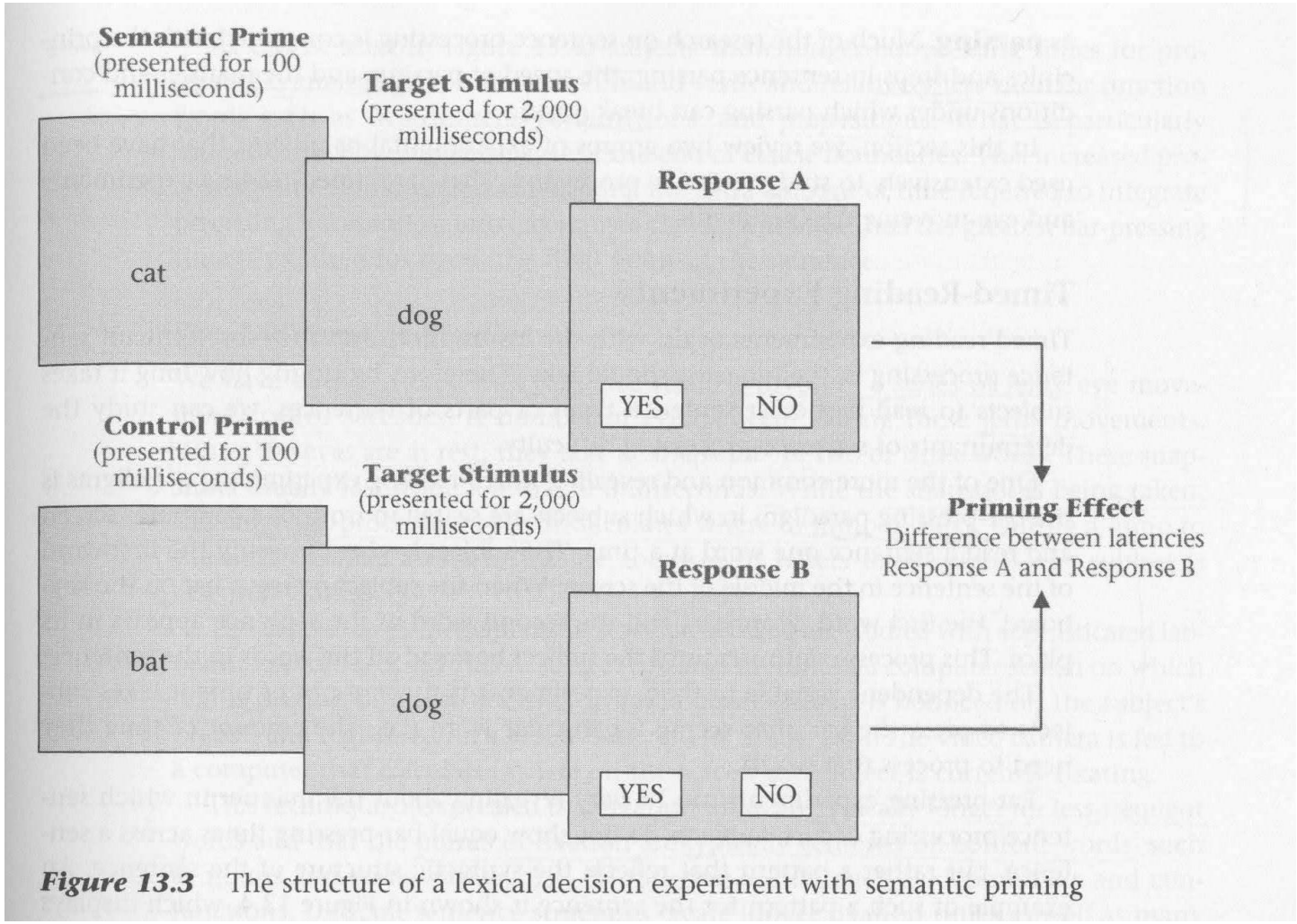


Figure 13.3 The structure of a lexical decision experiment with semantic priming

- priming poukazuje na to, jak jsou slova v mysli uspořádána → nejde o prostý soubor, ale o uspořádání do sítí

- stimul aktivuje nejen dané slovo, ale i slova, která jsou podobná (sémanticky → pes, kočka; foneticky → den, sen; ortograficky atd.)

- při akustickém vstupu do systému aktivujeme všechna lexikální hesla, která odpovídají tomu, co jsme slyšeli → jakmile vstup začne určité alternativy vylučovat, dochází k jejich deaktivaci

  více viz Altmann (2005, s. 84nn)

- srov. J. Chromý: https://www.youtube.com/watch?v=n9ZReisTcoA&list=PLZDL1lmSckjEoQse65LEG2-bv9PIe_Ey8&index=32

Time-Reading Experiment

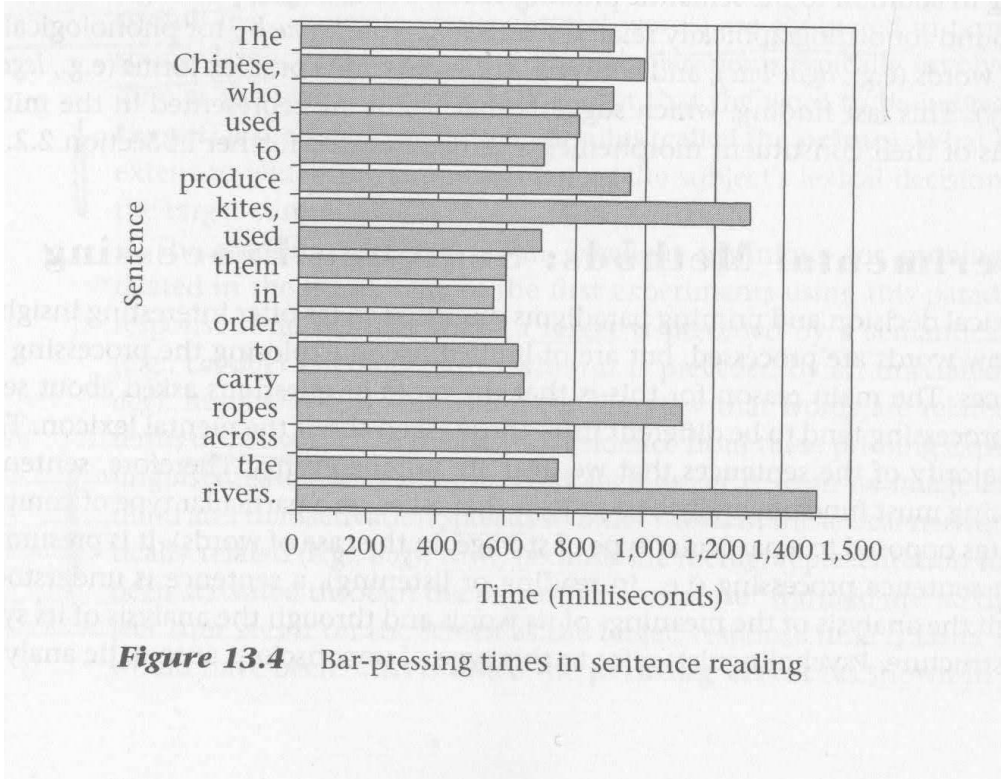


Figure 13.4 Bar-pressing times in sentence reading

- reflexe syntaktické struktury věty
- vnímání plnovýznamových slov časově náročnější než slov s významem gramatickým (ta tvoří uzavřenou skupinu, frekvence atd.)
- čas se prodlužuje i na konci jednotlivých vět

Behaviorální metody - shrnutí

- zpravidla analyzovány skupiny mluvčích → statistické vyhodnocení dat
- replikovatelnost
- dovoluje provádět analýzy jiného typu než analýzy jazykového chování

Data v lingvistice - shrnutí

- kromě expertní (individuální) introspekce
  - možnost komplexnějšího vyhodnocení → statistika
  - intersubjektivita
  - replikovatelnost

Kritická práce z daty - poznámka

- tento kurz = analýza jazykového chování (korpusy, texty)

Modely jazykového chování a jejich interpretace

- jazykové chování
  - dynamika, „nestabilita“
  - náhodné fluktuace
  - počínající tendence (srov. jazyková změna a její evoluce)

- **pravidlo**
  - tradičně pojato v deterministickém smyslu
    - jediná instance v rozporu s pravidlem = pravidlo neplatí

Pravidlo – deterministické pojetí

- slovesný přísudek se shoduje se subjektem

Petr zpíval × *Petr zpívala

Marie tancovala x *Marie tancoval

Pravidlo – nedeterministické/stochastické pojetí

- heslo SLOVOSLED NOMINÁLNÍ SKUPINY v NESČ“

Prepozice neshodného přívlastku je v principu **negramatická**

*Koupil mi nůžky na papír × *Na papír mi koupil nůžky*

(správná struktura ve čtení, při němž je na papír příslovečné určení).

Poměrně běžná je však prepozice genitivních přívlastků přivlastňovacích, a to zvláště v hovorovém stylu:

Mého pradědečka bratr padl v první světové válce
Našeho souseda zahrádka je plná krásných květin

Pravidlo – nedeterministické/stochastické pojetí

- heslo SLOVOSLED NOMINÁLNÍ SKUPINY v NESČ“

Prepozice neshodného přívlastku je v principu **negramatická**

*Koupil mi nůžky na papír × *Na papír mi koupil nůžky*

(správná struktura ve čtení, při němž je na papír příslovečné určení).

Poměrně běžná je však prepozice genitivních přívlastků přivlastňovacích, a to zvláště v hovorovém stylu:

Mého pradědečka bratr padl v první světové válce
Našeho souseda zahrádka je plná krásných květin

Modely jazykového chování a jejich interpretace

- jazykové chování
  - dynamika, „nestabilita“
  - náhodné fluktuace
  - počínající tendence (srov. jazyková změna a její evoluce)

- **stochastické** pojetí pravidel (a jazyka)
  - tendence
    - příklady?

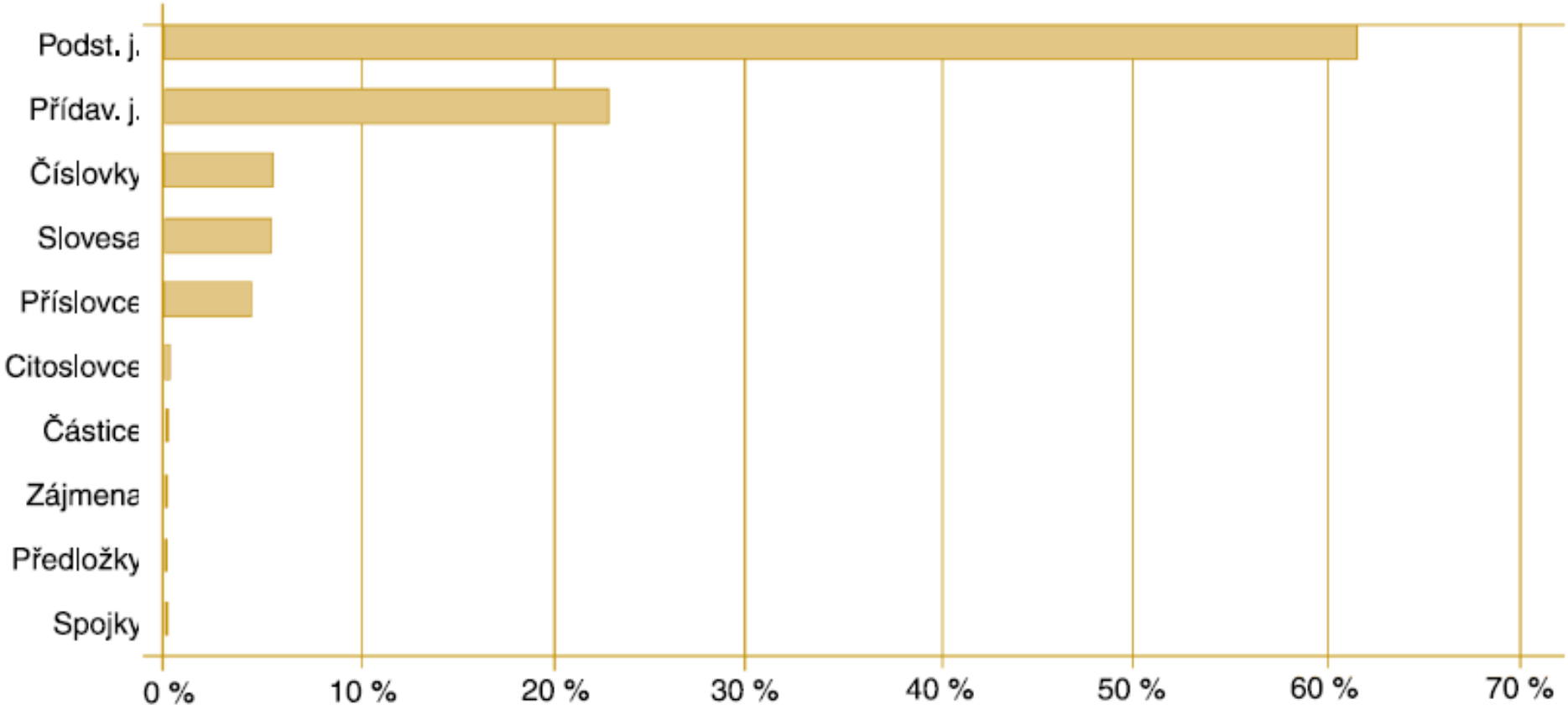
Modely jazykového chování a jejich interpretace

- jazykové chování
  - dynamika, „nestabilita“
  - náhodné fluktuace
  - počínající tendence (srov. jazyková změna a její evoluce)

- **stochastické** pojetí pravidel (a jazyka)
  - tendence
  - deterministické pravidlo je pak de facto extrémním případem stochastického pravidla -> vyskytuje se s pravděpodobností = 1
  - pravděpodobnost

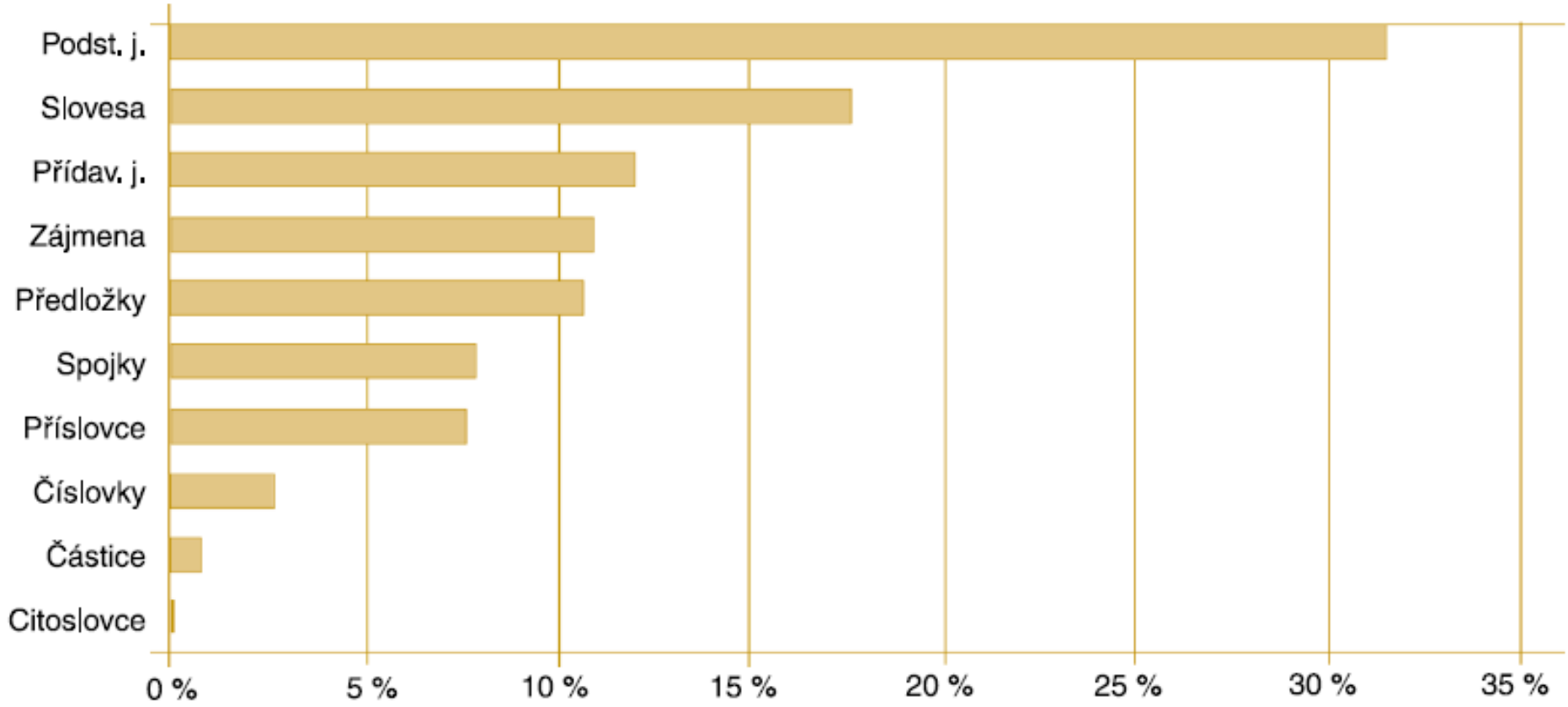
Frekvence

- **popis** jazykového systému, tak jak se projevuje v jazykovém chování
  - frekvenční charakteristiky jako další informace o povaze jazyka



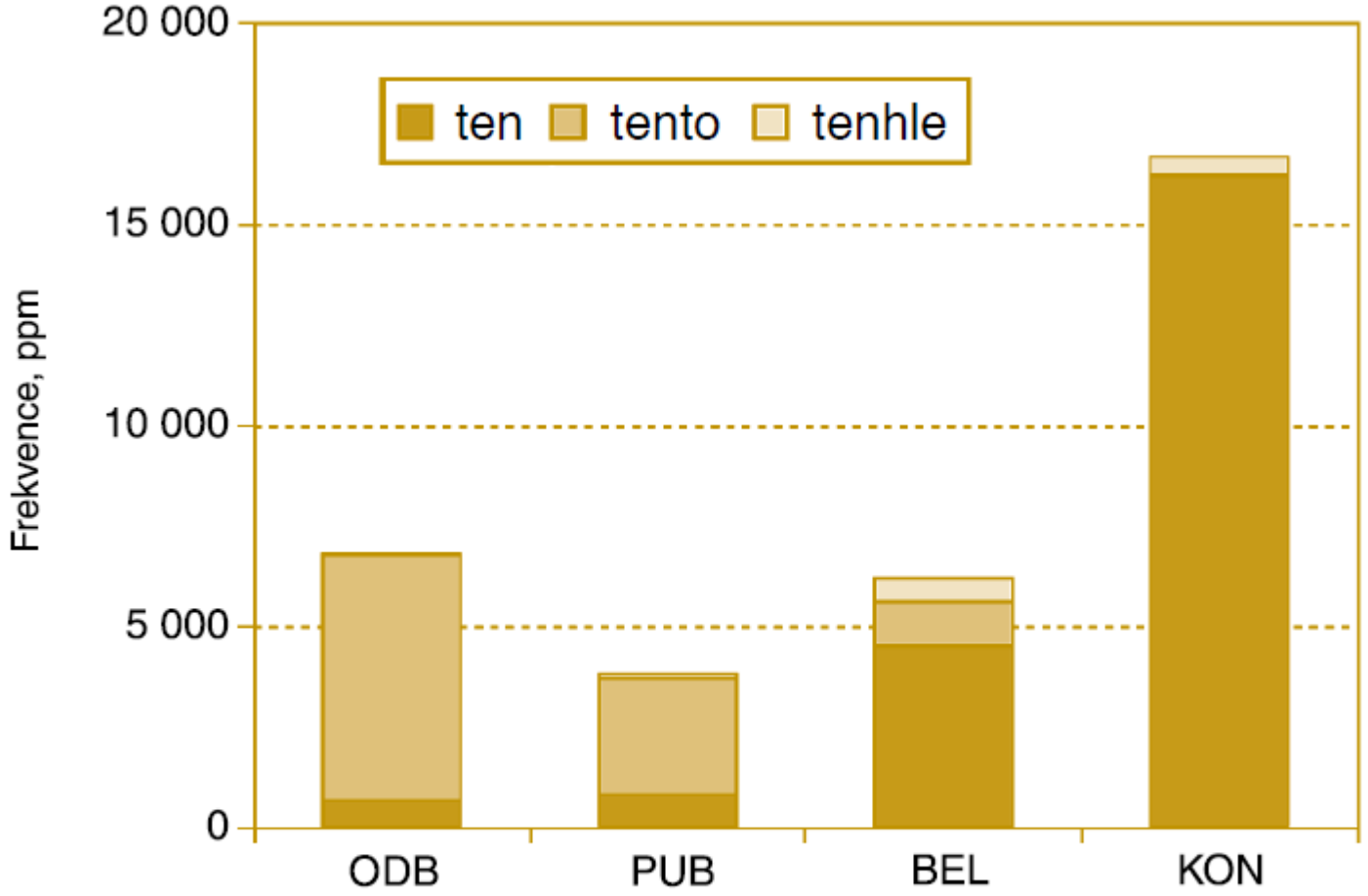
Obr. 7.1: Procentní zastoupení slovních druhů ve slovníku *(každé slovo počítáno jednou).*

- Cvrček et al. 2010



Obr. 7.2: Procentní zastoupení slovních druhů v textech *(počítán každý výskyt slova).*

- Cvrček et al. 2010



Obr. 7.4: Frekvence determinátorů *ten, tento* a *tenhle* v pozici před substantivem v různých textových typech.

- Cvrček et al. 2010

Biber et al. (1999): Longman Grammar of Spoken and Written English

The distribution of nouns and pronouns varies greatly depending upon register (2.3.5, 2.4.14). It further turns out that the use of pronouns v. full noun phrases varies in relation to syntactic role.

CORPUS FINDINGS [3,16]

Pronouns are slightly more common than nouns in conversation.

At the other extreme, nouns are many times more common than pronouns in news and academic prose.

The noun-pronoun ratio varies greatly depending upon syntactic role.

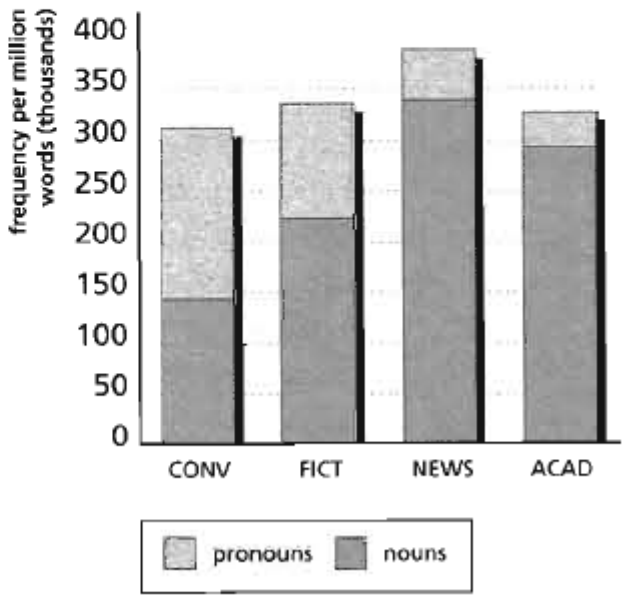
- The relative frequency of nouns is much higher in object position and as a complement or object of a preposition than in subject position.

DISCUSSION OF FINDINGS

As illustrated in 4.1.1, there are important differences in the reliance on nouns v. pronouns across registers. In

Figure 4.1

Distribution of nouns v. pronouns across registers



Stochastické pojetí jazyka

- **popis** jazykového systému, tak jak se projevuje v jazykovém chování
  - frekvenční charakteristiky jako další informace o povaze jazyka
  - distribuční rozdíly a jejich důvody (viz níže)

Stochastické pojetí jazyka

- **popis** jazykového systému, tak jak se projevuje v jazykovém chování
  - frekvenční charakteristiky jako další informace o povaze jazyka
  - distribuční rozdíly a jejich důvody (viz níže)
  - pravděpodobnostní modely reflektují charakter pozorovaných jevů (viz níže)

Stochastické pojetí jazyka

- **popis** jazykového systému, tak jak se projevuje v jazykovém chování
  - frekvenční charakteristiky jako další informace o povaze jazyka
  - distribuční rozdíly a jejich důvody (viz níže)
  - pravděpodobnostní modely reflektují charakter pozorovaných jevů (viz níže)

- **modely mechanismů** řídících jazykové chování
  - jejich platnost ověřována prostřednictvím empiricky testovatelných hypotéz

Stochastické pojetí jazyka

- **popis** jazykového systému, tak jak se projevuje v jazykovém chování
  - frekvenční charakteristiky jako další informace o povaze jazyka
  - distribuční rozdíly a jejich důvody (viz níže)
  - pravděpodobnostní modely reflektují charakter pozorovaných jevů (viz níže)

- **modely mechanismů** řídících jazykové chování
  - jejich platnost ověřována prostřednictvím empiricky testovatelných hypotéz
  - stochastické pojetí hypotéz
    - statistika

Stochastické pojetí jazyka

- teoretické rámce a jazykové teorie:
  - G. K. Zipf
  - emmergent grammar
  - synergetická lingvistika

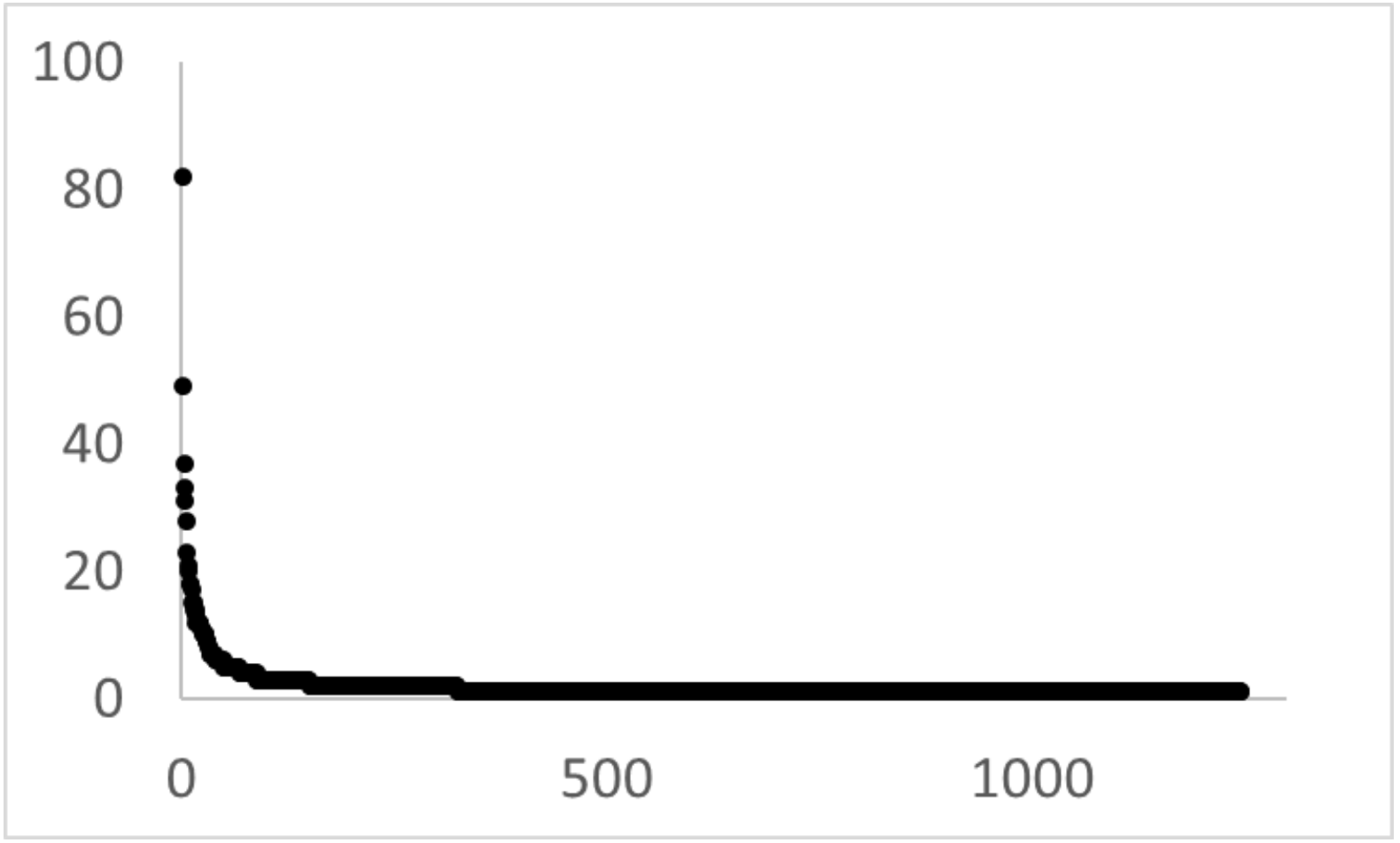
- relativně běžná současná praxe
  - ověřování mechanismů **bez** hlubšího teoretického rámce
    - komputační lingvistika
    - ad hoc analýzy

Frekvence

- smysluplná pouze jako „vztahová“ veličina
  - **distribuce** jednotek určitého typu
    - ranková frekvenční distribuce
    - frekvence délek slov/vět…
    - …
  - **vztah** frekvence a jiných vlastností
    - frekvence slovních druhů vs. typ textu
    - frekvence vs. délka slova
    - frekvence vs. polysémie
    - …

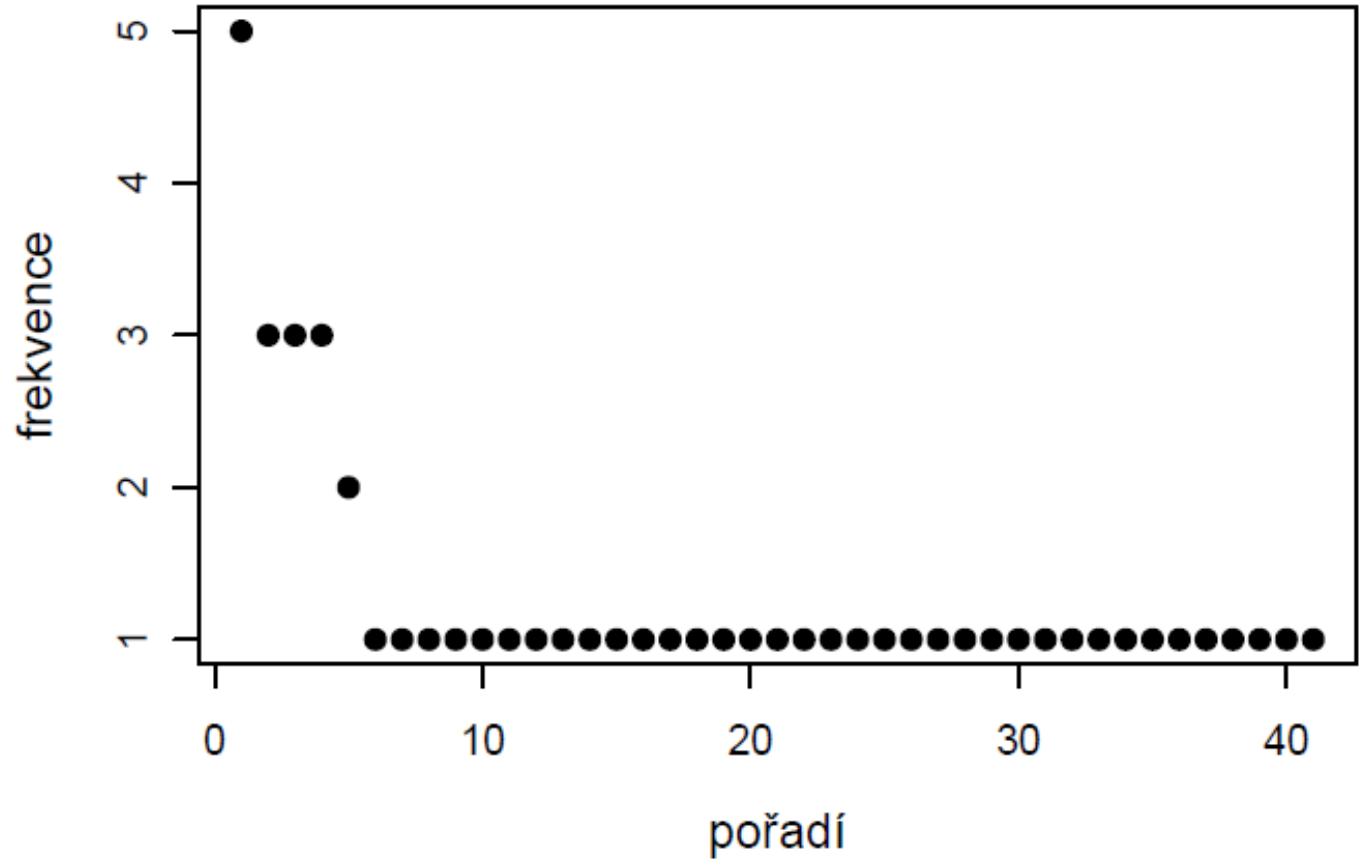
Distribuce jednotek

- Havel 1990: ranková frekvenční distribuce slov



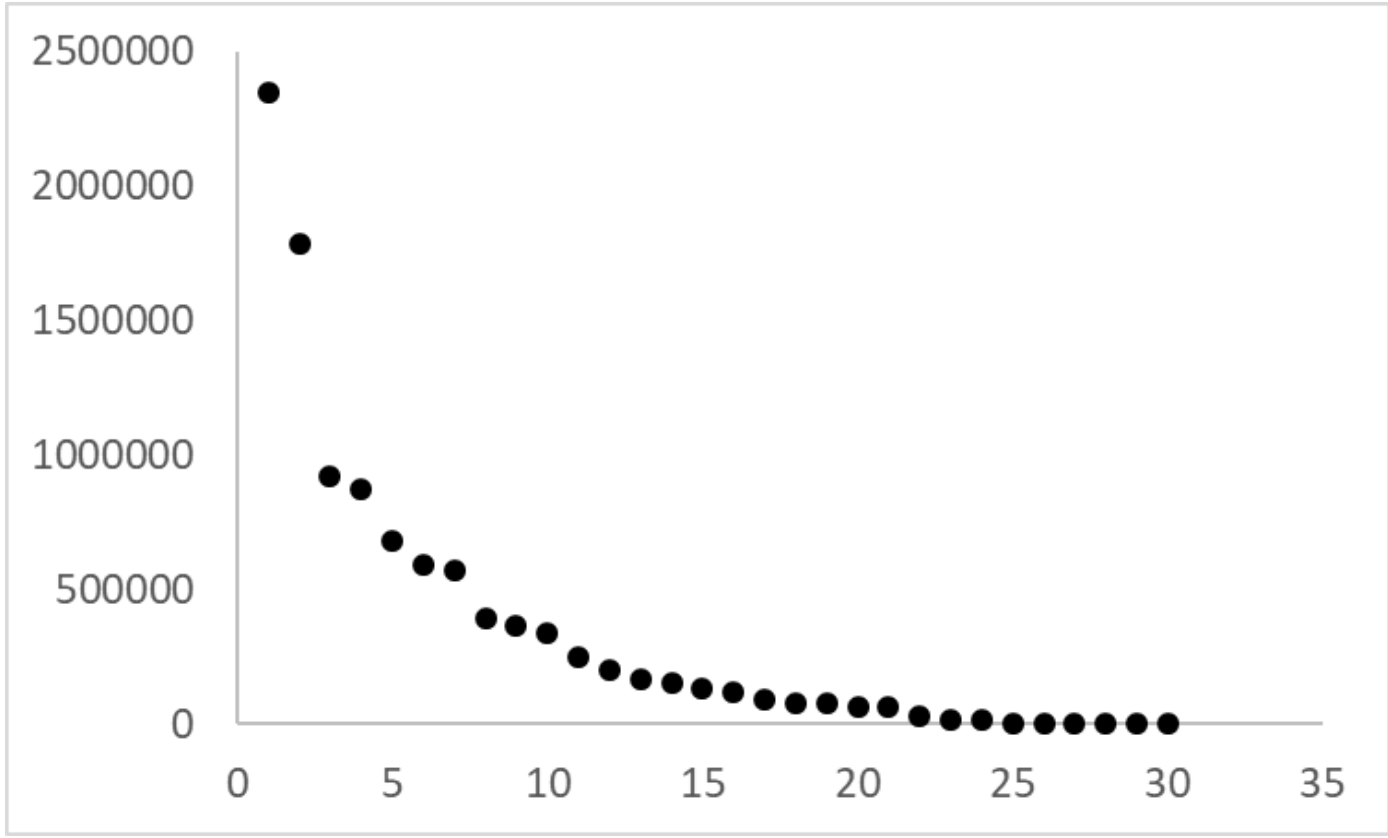
Distribuce jednotek

- Skácel: *Odvaha k tomu*: ranková frekvenční distribuce slov



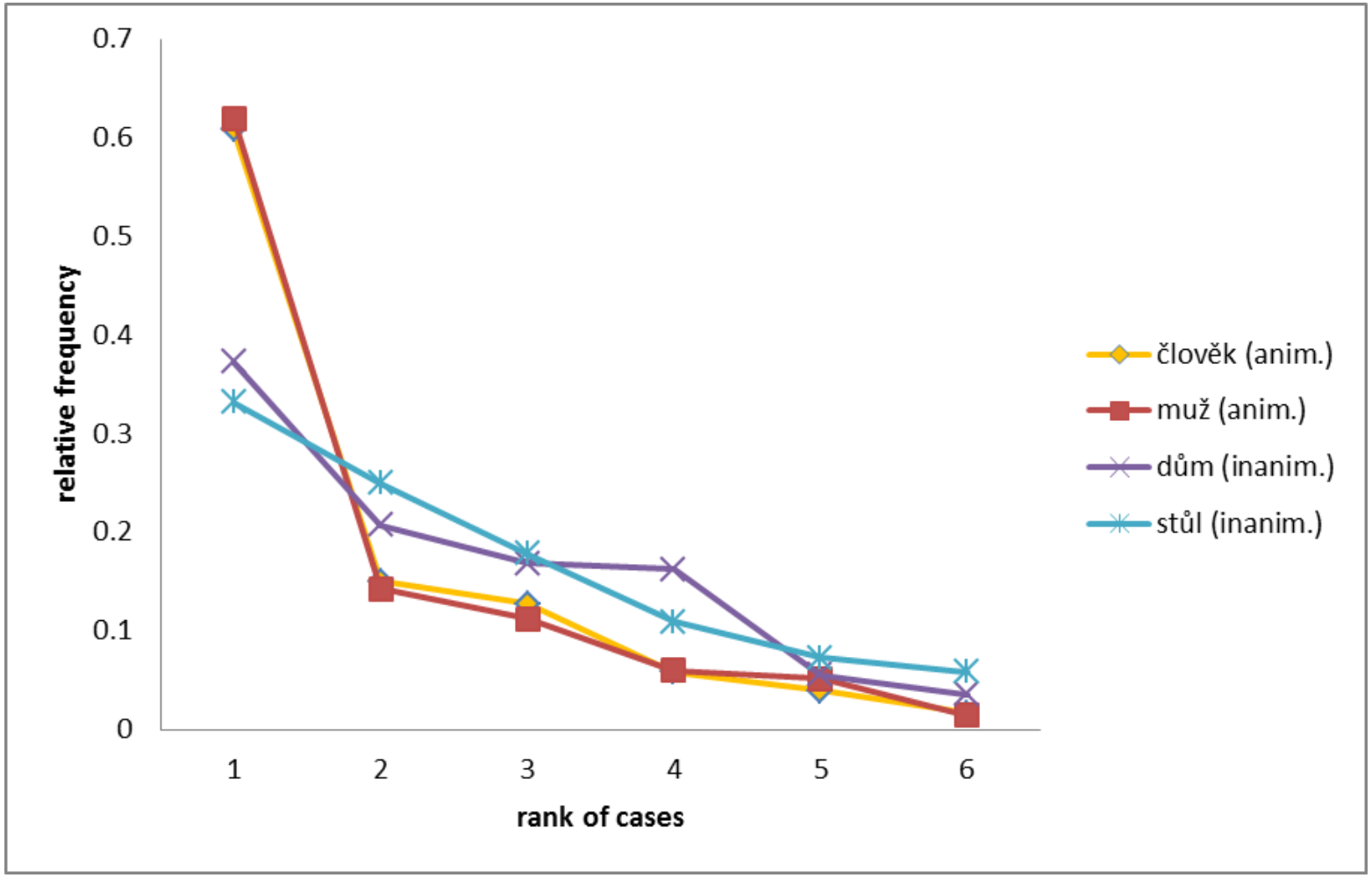
Distribuce jednotek

- SYN2005: : ranková frekvenční distribuce primárních předložek



Distribuce jednotek

- SYN2010: : ranková frekvenční pádů



Distribuce – modely a interpretace

- diverzifikovanost sytému:

- type-token ratio
- repeat rate
- entropie

- výsledkem jedna hodnota

Příklad

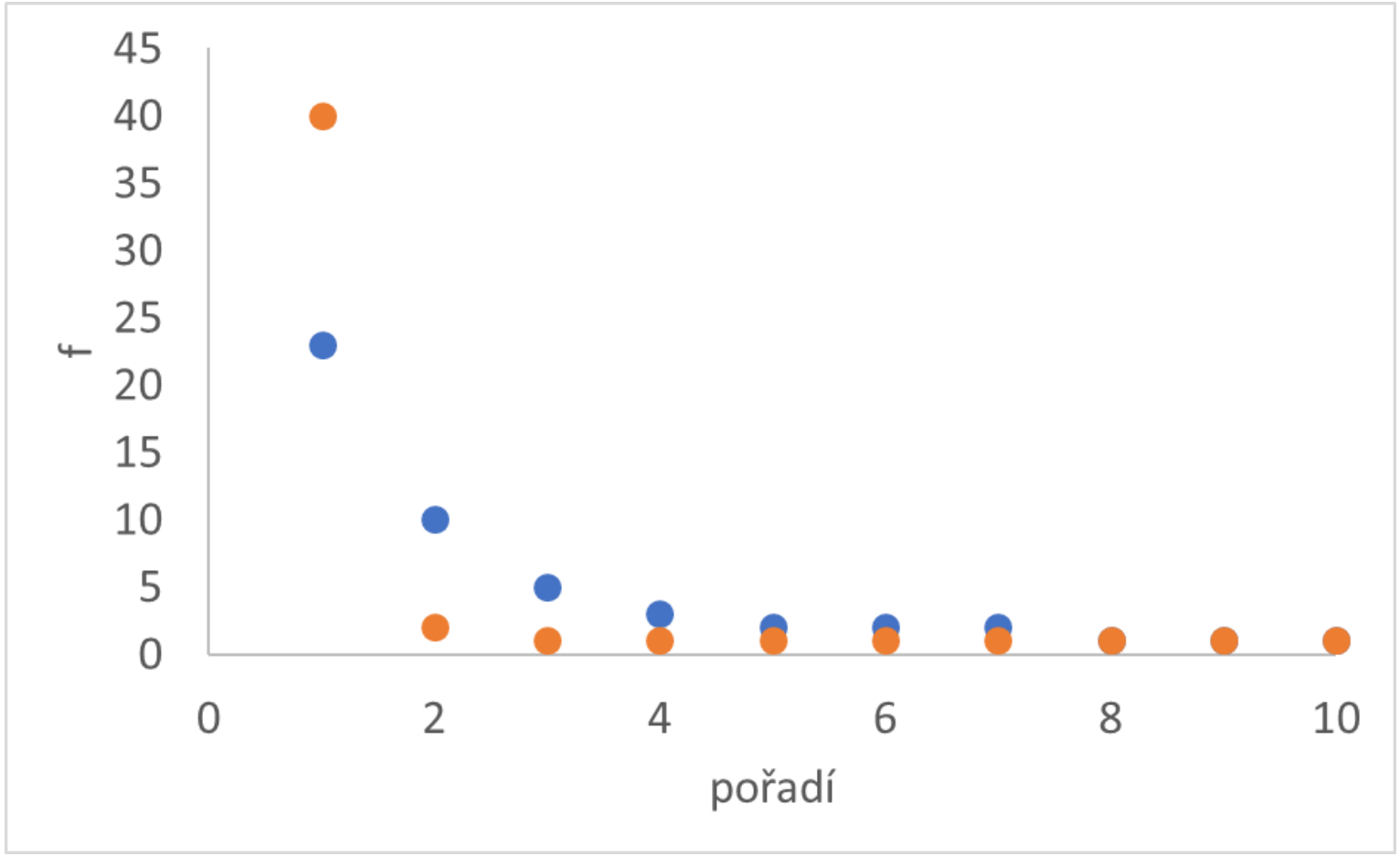
- V = 10 jednotek
- N = 50 výskytů

- dvě různé distribuce

Příklad

pořadí	f		pořadí	f
1	23		1	40
2	10		2	2
3	5		3	1
4	3		4	1
5	2		5	1
6	2		6	1
7	2		7	1
8	1		8	1
9	1		9	1
10	1		10	1

Příklad



Příklad – type-token poměr

- diverzifikovanost/slovní bohatství

$$TTR = \frac{V}{N}$$

- jaká bude teoreticky nejvyšší a nejnižší hodnota TTR u textu (souboru), který bude mít délku N = 50 slov?

Příklad – type-token poměr

- nejvyšší hodnota = každé slovo pouze jednou

$$TTR_{max} = \frac{V}{N}$$

Příklad – type-token poměr

- nejvyšší hodnota = každé slovo pouze jednou

$$TTR_{max} = \frac{V}{N} = \frac{50}{50} = 1$$

Příklad – type-token poměr

- nejvyšší hodnota = každé slovo pouze jednou

$$TTR_{max} = \frac{V}{N} = \frac{50}{50} = 1$$

- nejnižší hodnota = jedno slovo v celém texu

Příklad – type-token poměr

- nejvyšší hodnota = každé slovo pouze jednou

$$TTR_{max} = \frac{V}{N} = \frac{50}{50} = 1$$

- nejnižší hodnota = jedno slovo v celém texu

$$TTR_{min} = \frac{V}{N} = \frac{1}{50} = 0.02$$

Příklad – type-token poměr

- a co naše hypotetická data?
- liší se jejich TTR?

pořadí	f		pořadí	f
1	23		1	40
2	10		2	2
3	5		3	1
4	3		4	1
5	2		5	1
6	2		6	1
7	2		7	1
8	1		8	1
9	1		9	1
10	1		10	1

Příklad – type-token poměr

- a co naše hypotetická data?
- liší se jejich TTR?

$$TTR_{příklad} = \frac{V}{N} = \frac{10}{50} = 0.2$$

pořadí	f		pořadí	f
1	23		1	40
2	10		2	2
3	5		3	1
4	3		4	1
5	2		5	1
6	2		6	1
7	2		7	1
8	1		8	1
9	1		9	1
10	1		10	1

Příklad – index opakování (repeat rate)

- míra koncentrace jednotek (např. slov) v souboru

$$RR = \sum_{r=1}^{V} p_r^2$$

$$p_r = \frac{f_r}{N}$$

$$RR = \frac{1}{N^2} \sum_{r=1}^{V} f_r^2$$

Příklad – index opakování (repeat rate)

- nejvyšší koncentrace = jedno slovo v celém texu

$$RR_{max} = \frac{f_r^2}{N^2} = \frac{50^2}{50^2} = \frac{2500}{2500} = 1$$

- nejnižší koncentrace = každé slovo pouze jednou

$$RR_{min} = \frac{f_r^2}{N^2} = \frac{1^2 + 1^2 + 1^2 \ldots + 1^2}{50^2} = \frac{50}{2500} = 0.02$$

Příklad – repeat rate

- a co naše hypotetická data?
- liší se jejich RR?

pořadí	f		pořadí	f
1	23		1	40
2	10		2	2
3	5		3	1
4	3		4	1
5	2		5	1
6	2		6	1
7	2		7	1
8	1		8	1
9	1		9	1
10	1		10	1

Příklad – repeat rate

- a co naše hypotetická data?
- liší se jejich RR?

- Excel

pořadí	f		pořadí	f
1	23		1	40
2	10		2	2
3	5		3	1
4	3		4	1
5	2		5	1
6	2		6	1
7	2		7	1
8	1		8	1
9	1		9	1
10	1		10	1

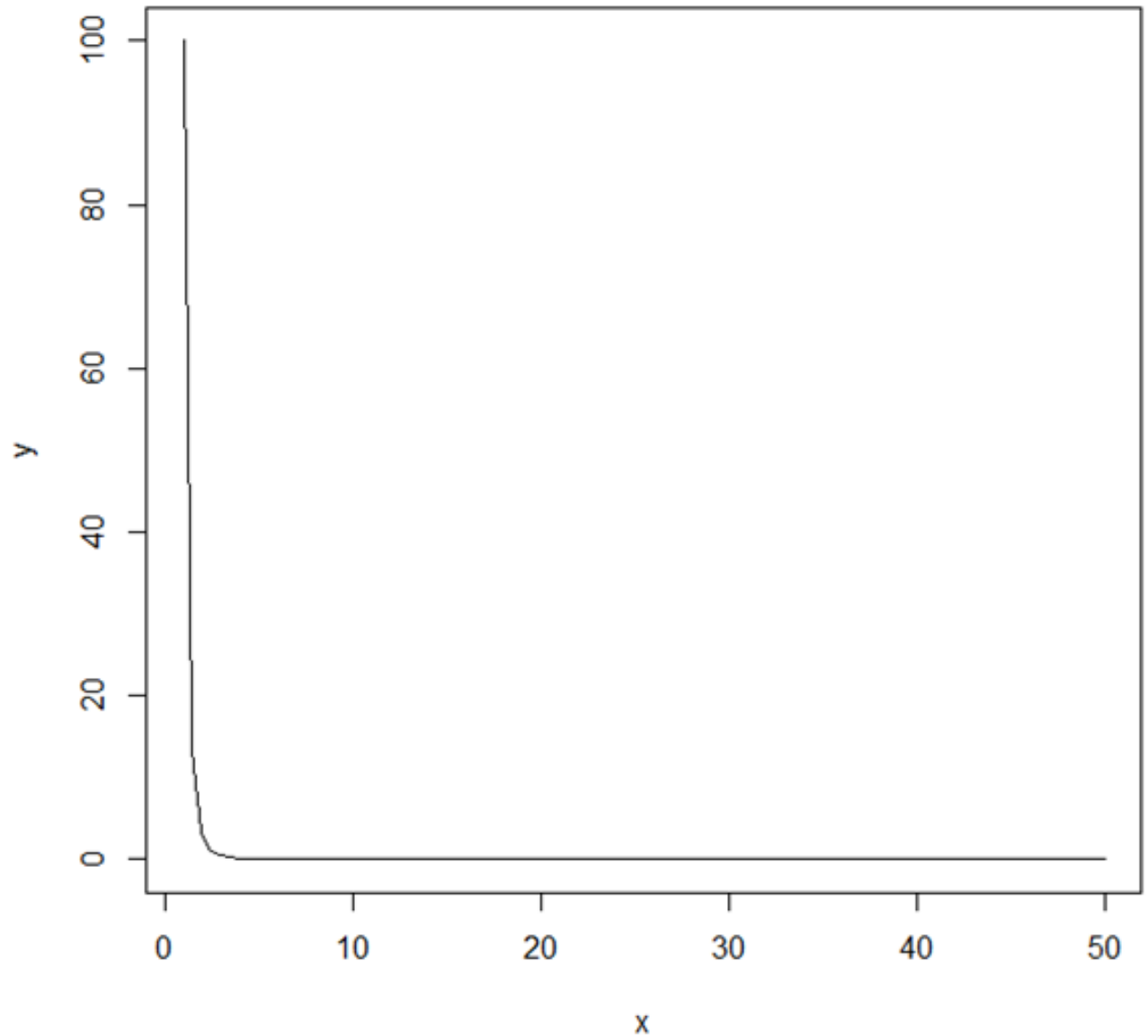
Distribuce – modely a interpretace

- distribuční funkce
  - diskrétní veličiny
  - spojité veličiny

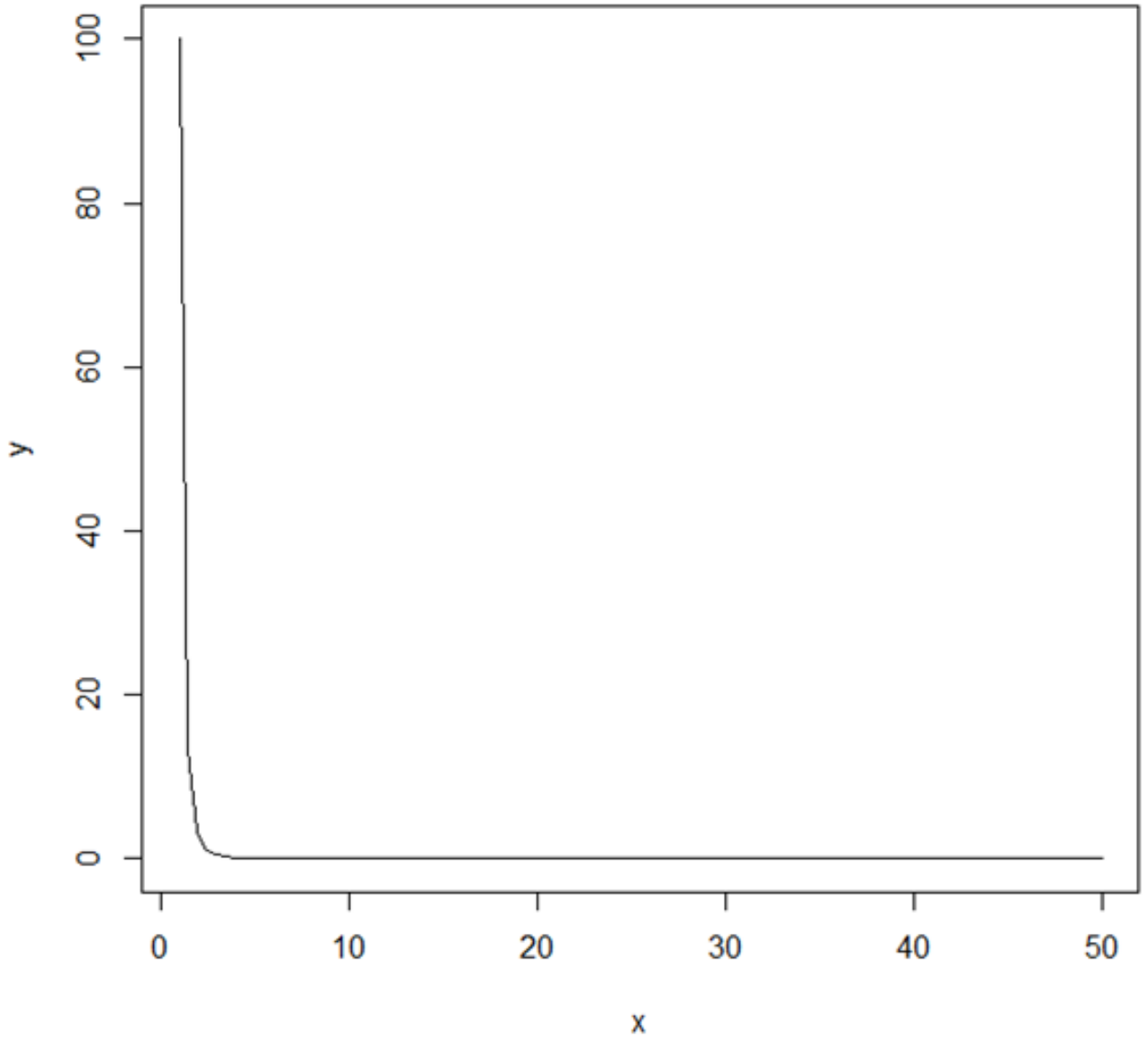
Model – mocninná funkce

$$y = ax^{-b}$$

$$y = 100x^{-5}$$

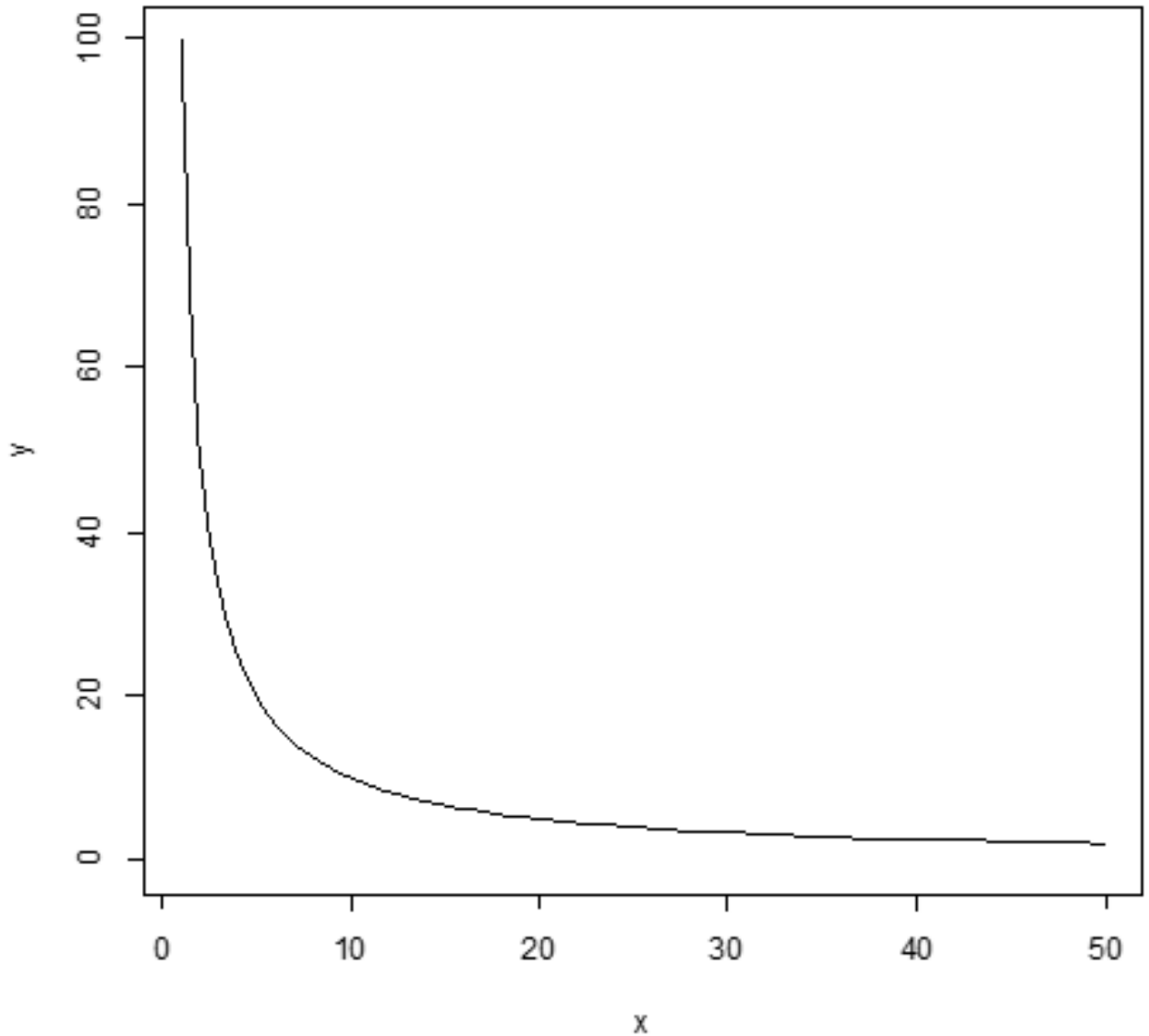


$$y = 100x^{-5}$$

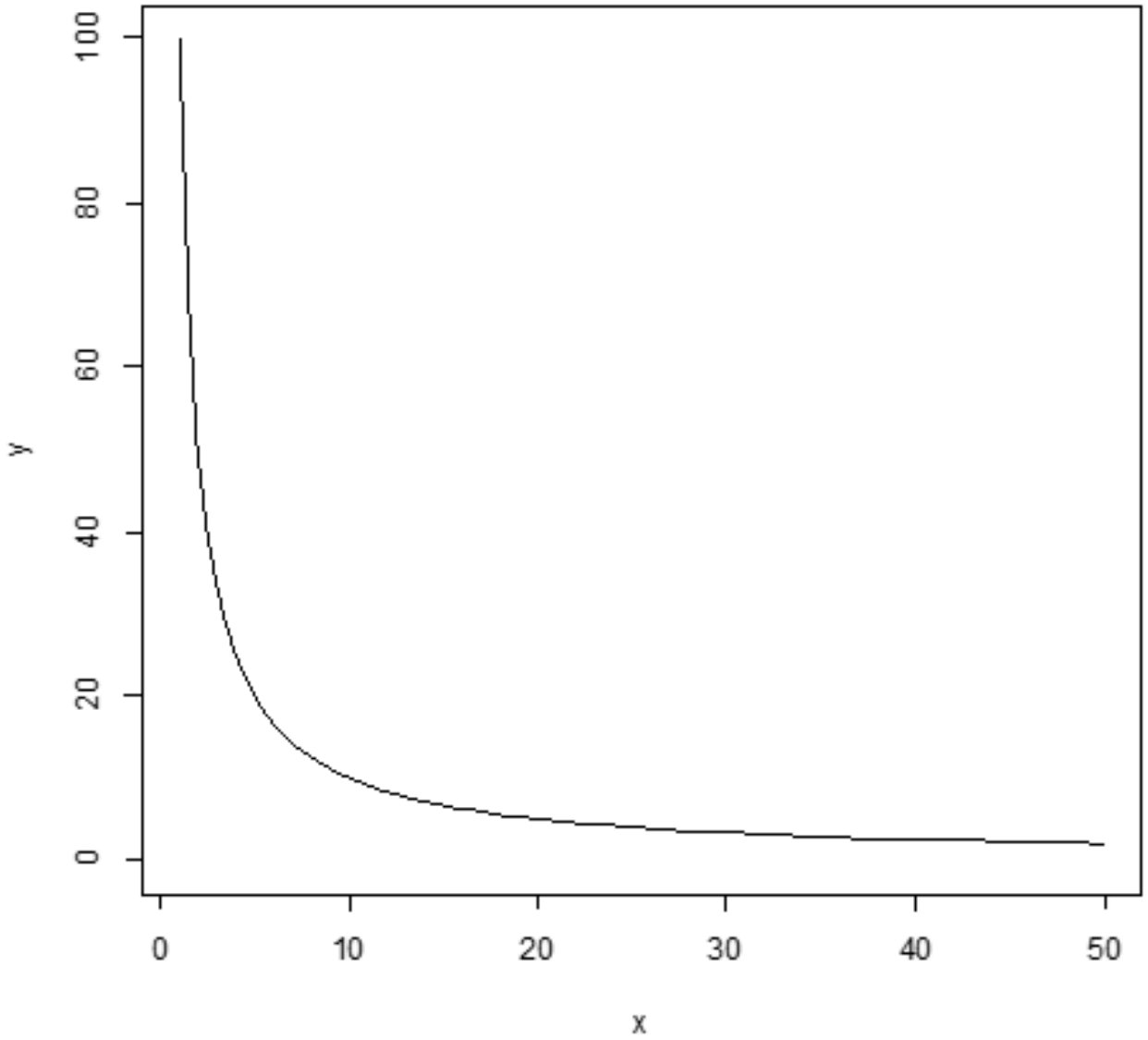


- bohatý slovník
  - většina slov se neopakuje

$$y = 100x^{-1}$$

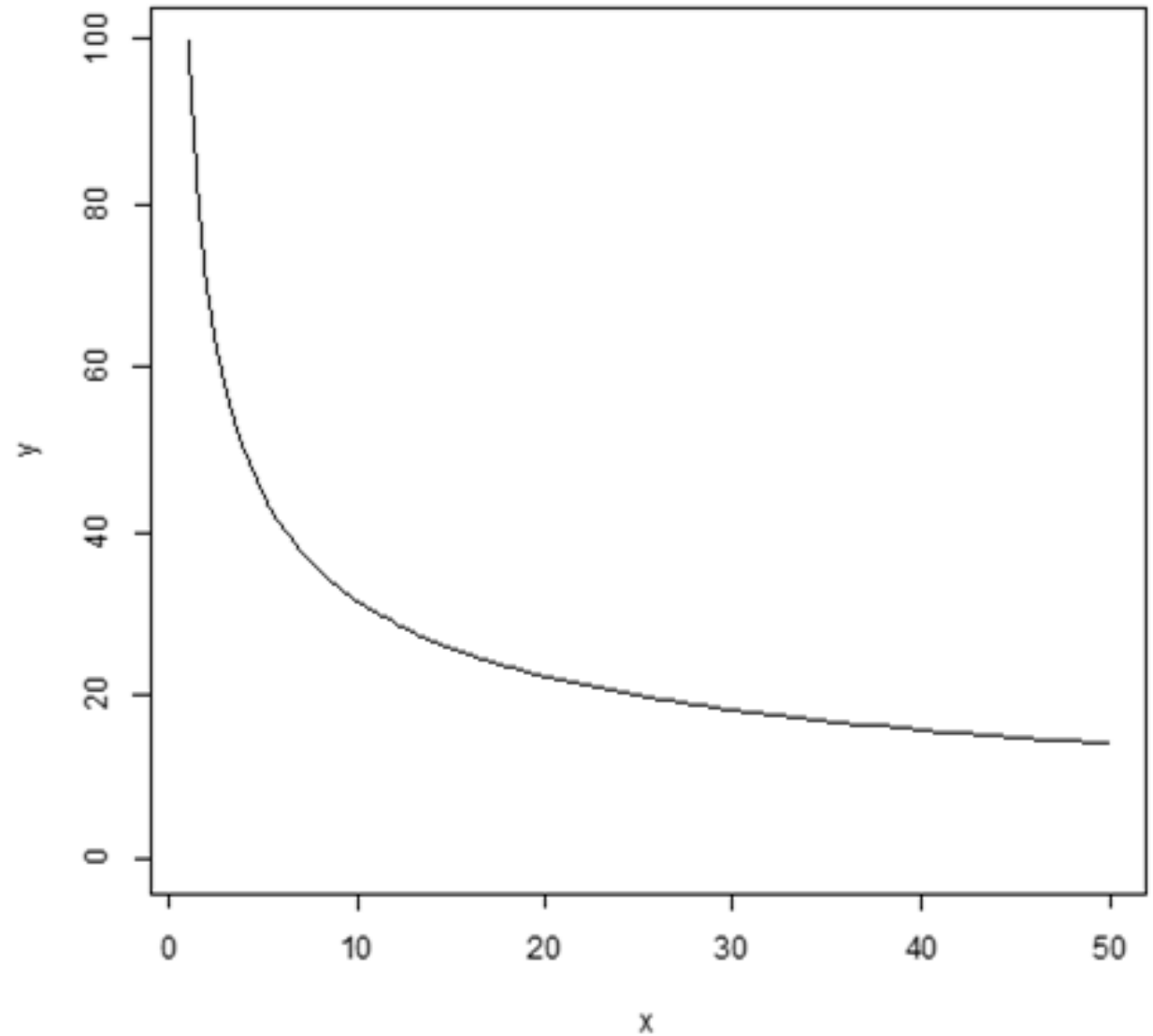


$$y = 100x^{-1}$$

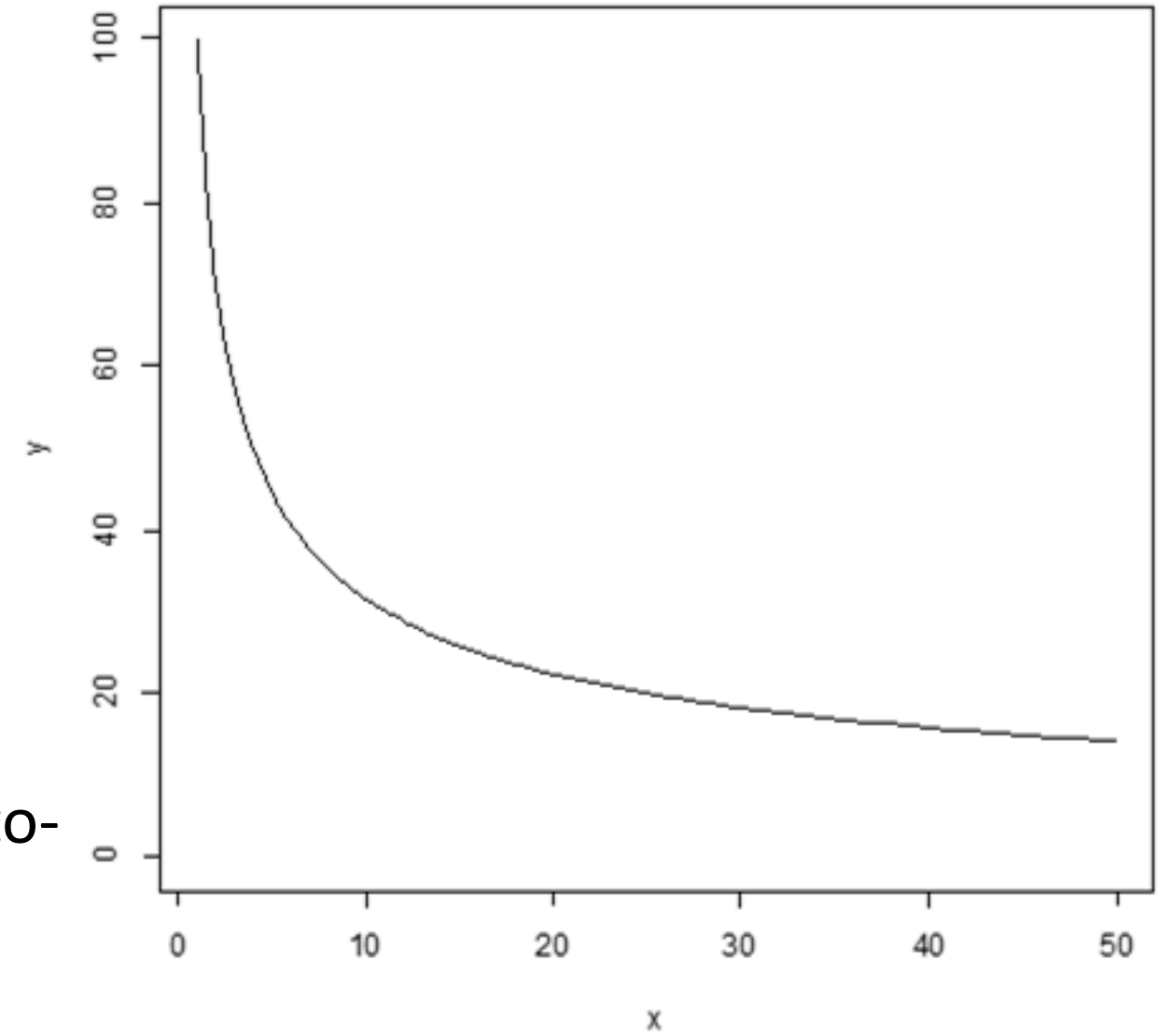


- chudší slovník
  - slova se častěji opakují

$$y = 100x^{-0.5}$$



$$y = 100x^{-0.5}$$



- nejchudší slovník (z prezento-vaných příkladů)
  - slova se opakují ještě častěji

Distribuce příslovečných určení

- Čech, Uhlířová (2014)

Adverbial	*r*	*f*	f_r
Place	1	273	27.3
Time	2	204	20.4
Manner	3	172	17.2
Means	4	68	6.8
Aspect	5	61	6.1
Condition	6	59	5.9
Measure	7	52	5.2
Cause	8	30	3.0
Result	9	18	1.8
Origin	10	18	1.8
Purpose	11	17	1.7
Concession	12	16	1.6
Originator	13	12	1.2
∑		1 000	100

Adverbial	Noun	Adverb	Clause
Place	263	9	1
Time	96	104	4
Manner	79	75	18
Means	68	-	-
Aspect	46	13	2
Condition	30	-	29
Measure	21	30	1
Cause	11	-	19
Result	18	-	-
Origin	18	-	-
Purpose	10	-	7
Concession	4	-	12
Originator	12	-	-
∑	676	231	93
R^2	0.98	1	0.96

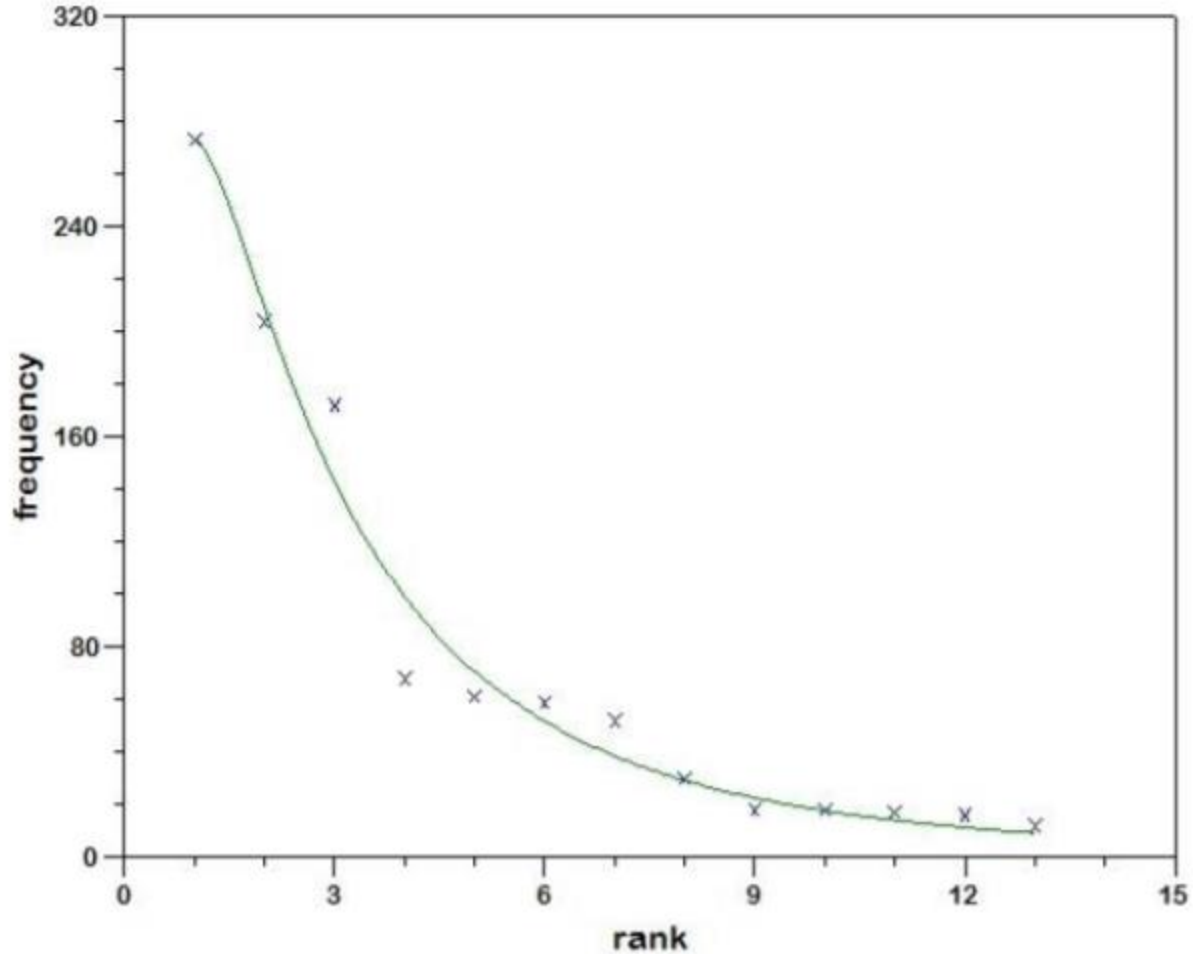


Figure 1. The distribution of all adverbials and the result of the fitting of the Zipf-Alekseev function to the data.

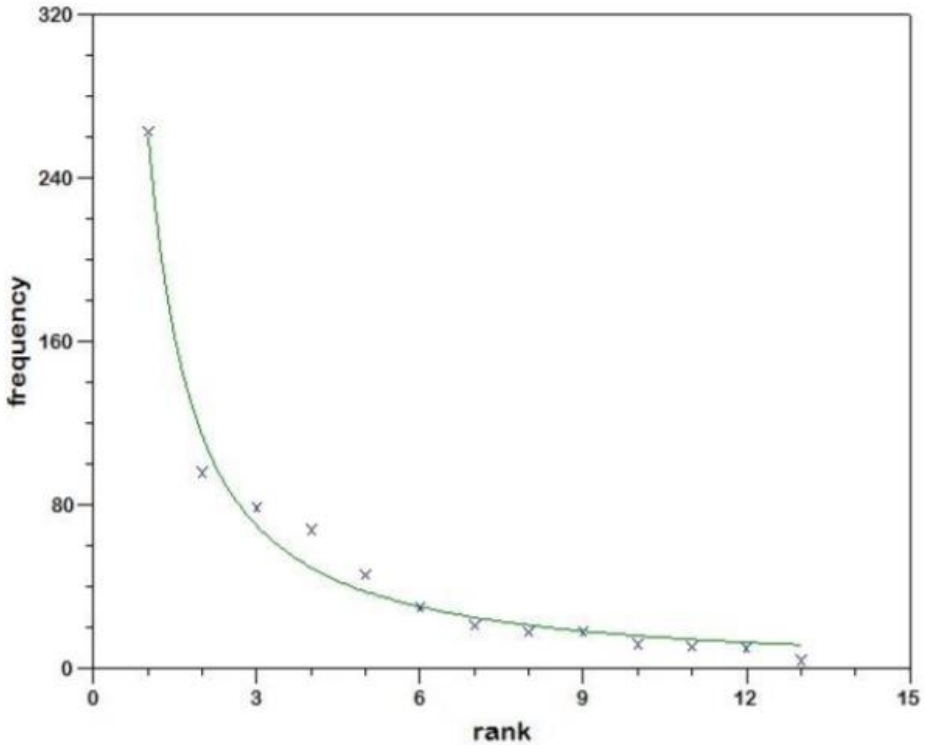


Figure 2. The distribution of adverbials expressed by nouns and the resu fitting of the Zipf-Alekseev function to the data.

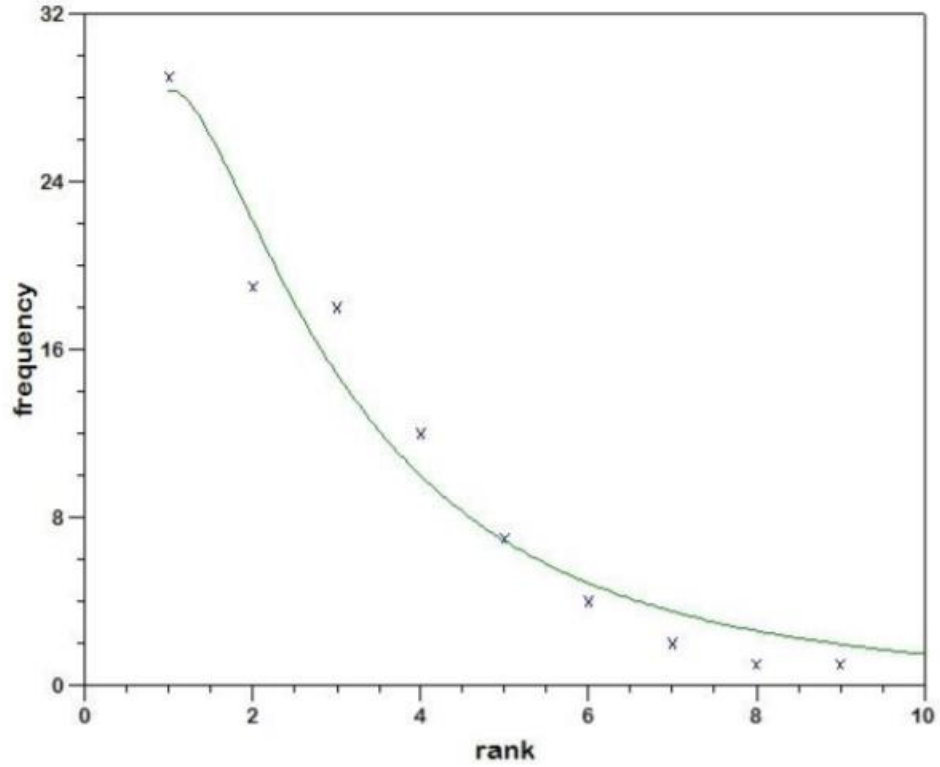


Figure 4. The distribution of adverbials expressed by clauses and the result of the fitting of the Zipf-Alekseev function to the data.

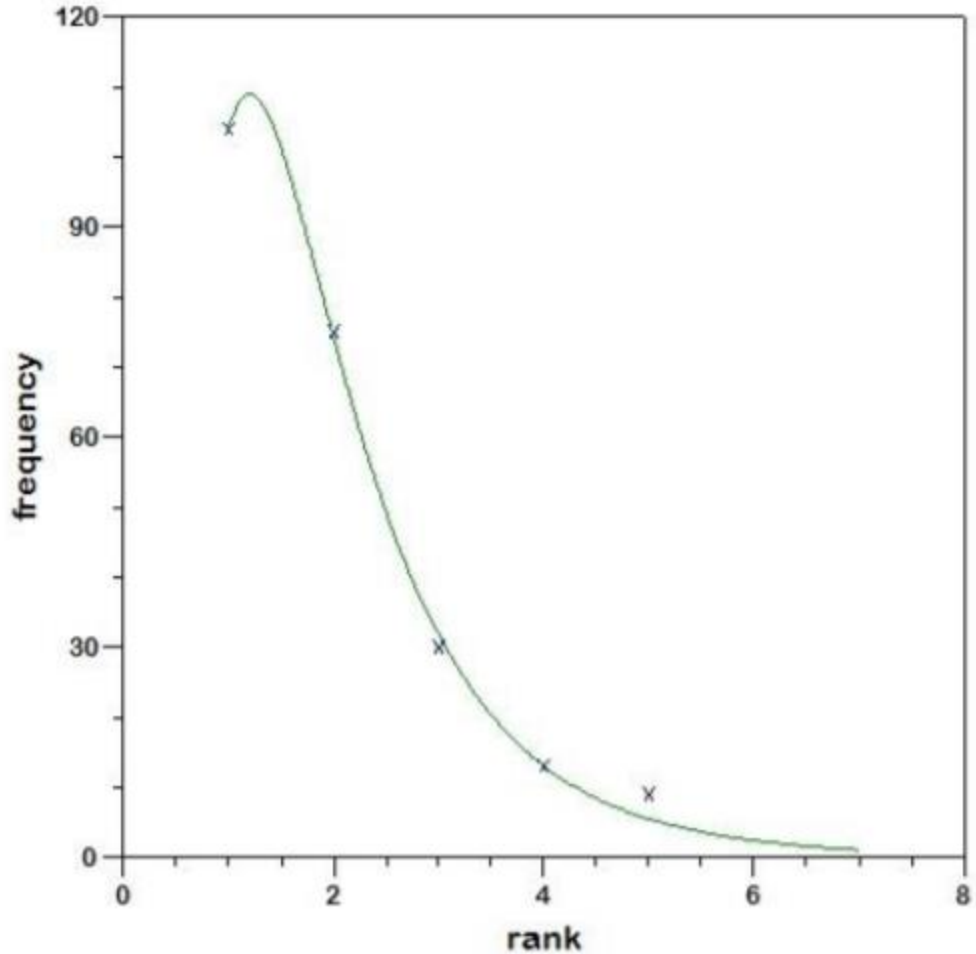
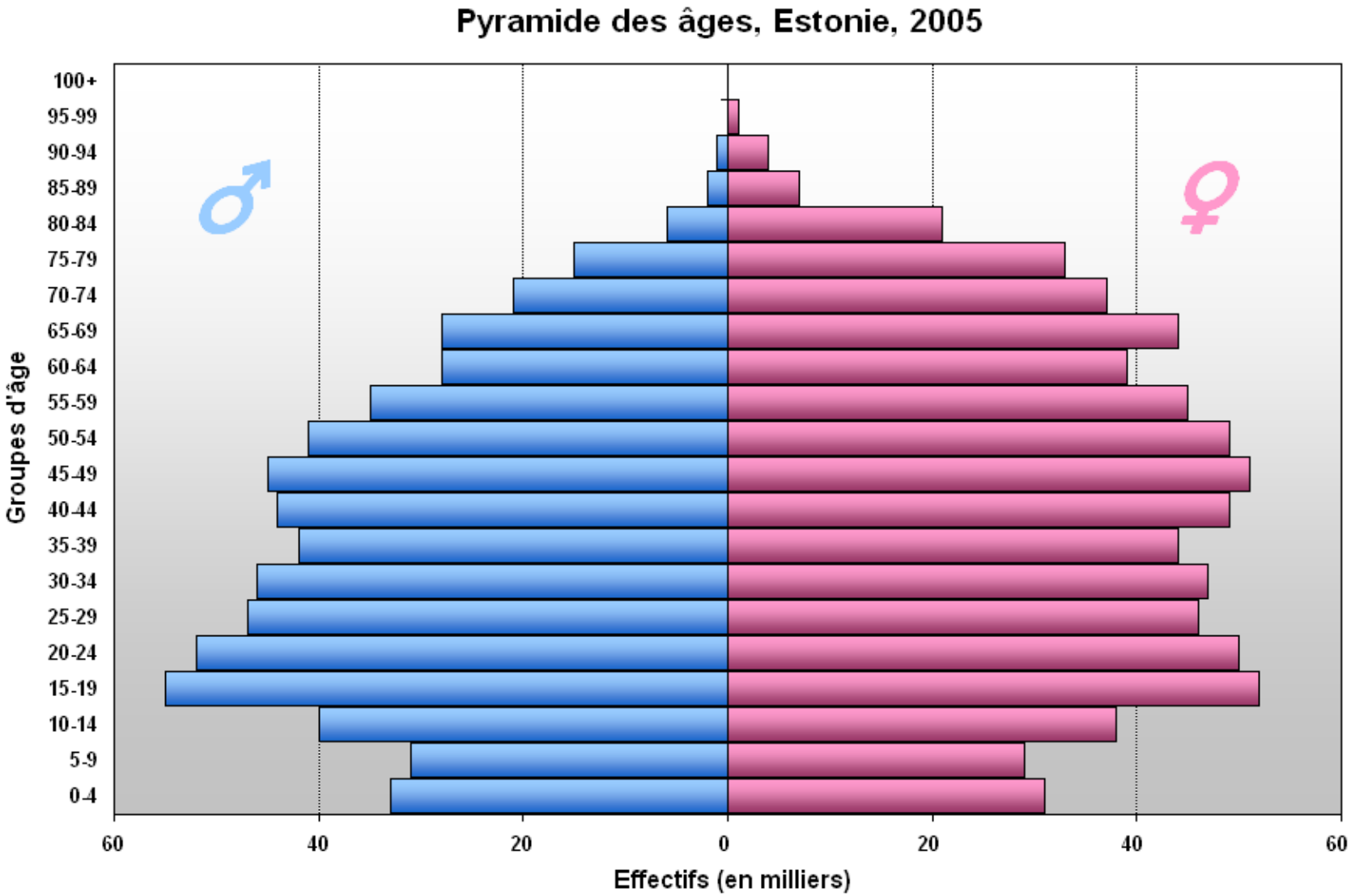
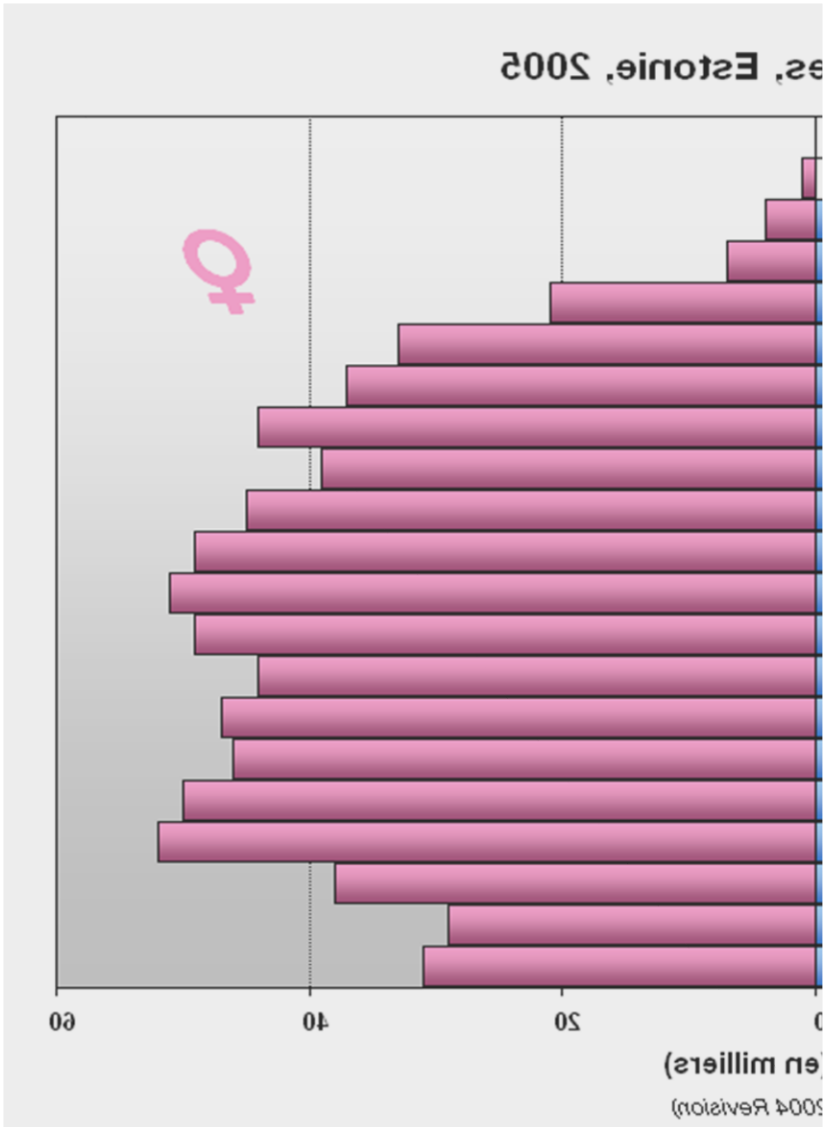
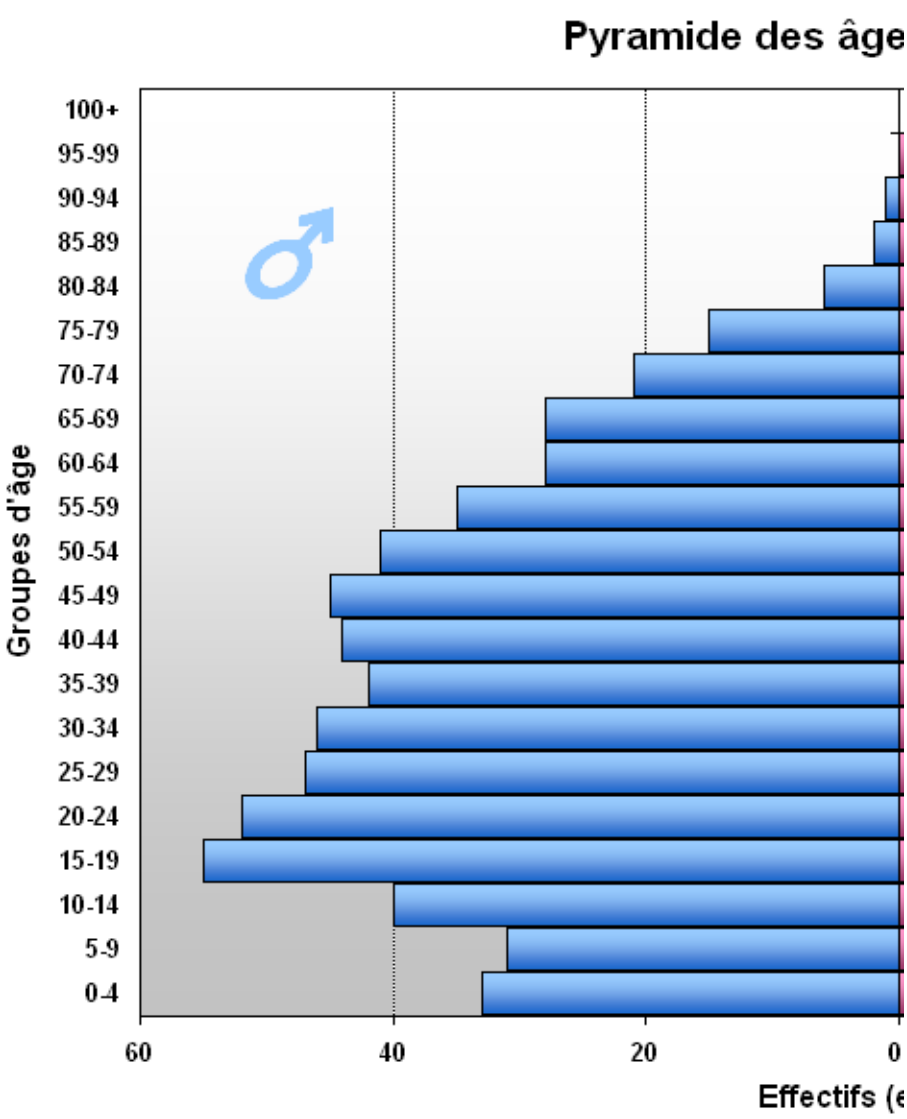


Figure 3. The distribution of adverbials expressed by adverbs and the result of the fitting of the Zipf-Alekseev function to the data.



Source: Organisation des Nations Unies (World Population Prospects: The 2004 Revision)





Source: Organisation des Nations Unies (World Population Prospects: The 2

Quantitative Linguistics, an Invitation

Karl-Heinz Best

Otto Rottmann

2017

RAM-Verlag

Case study

- Radek Čech, Emmerich Kelih, Jan Mačutek: Impact of semantics on case diversification

Porovnání délek – a jeho interpretace

- problémy
  - délka slova (S) délka mluvního taktu (MT)
  - délka slova (S) a vliv typu textu
  - délka mluvního taktu (MT) a vliv typu textu

Slovo vs. mluvní takt

- v čem je rozdíl?

Slovo vs. mluvní takt

- stůl
- na stole
- v domě
- napil se
- napil jsem se
- podali jsme jim ho
- řekl, že přijde

Slovo vs. mluvní takt

- stůl
- na stole
- v domě
- napil se
- napil jsem se
- podali jsme jim ho
- řekl, že přijde

- stůl
- nastole
- vdomě
- napilse
- napiljsemse
- podalijsmejimho
- řekl žepřijde

Porovnání délek – a jeho interpretace

- problémy
  - délka slova (S) délka mluvního taktu (MT)
  - délka slova (S) a vliv typu textu
  - délka mluvního taktu (MT) a vliv typu textu

- jazykový materiál
  - Ukradený kaktus (K. Čapek)
  - Žánrové a stylové proměny veřejné jazykové komunikace (J. Kraus)

Porovnání délek – a jeho interpretace

- problémy
  - délka slova (S) délka mluvního taktu (MT)
  - délka slova (S) a vliv typu textu
  - délka mluvního taktu (MT) a vliv typu textu

- jazykový materiál
  - Ukradený kaktus (K. Čapek)
  - Žánrové a stylové proměny veřejné jazykové komunikace (J. Kraus)
- segmentace
  - S jako grafická jednotka, délka (L) měřena v počtu slabik
  - MT vymezen podle Palkové (2004), délka (L) měřena v počtu slabik

Porovnání délek – a jeho interpretace

- očekávání
  - S budou kratší než MT
  - délka S a MT bude delší v odborných textech než v beletrii

Porovnání délek – a jeho interpretace

- očekávání
  - S budou kratší než MT
  - délka S a MT bude delší v odborných textech než v beletrii

- jak měřit?

Výsledky – průměrné délky

	L_S	L_{MT}
bel	2	2,89
odb	2,83	3,51

Průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

Průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

! citlivý na od odlehlé hodnoty

{2,2,3,3,4,5} průměr = 3,17

Průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

! citlivý na od odlehlé hodnoty

{2,2,3,3,4,5} průměr = 3,17

{2,2,3,3,4,20}

Průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

! citlivý na od odlehlé hodnoty

{2,2,3,3,4,5} průměr = 3,17

{2,2,3,3,4,20} průměr = 5,67

Průměr

$$\bar{x} = \frac{(x_1 + x_2 + x_3 + \ldots + x_n)}{n}$$

! citlivý na od odlehlé hodnoty

{2,2,3,3,4,5} průměr = 3,17

{2,2,3,3,4,20} průměr = 5,67

{5,5,6,6,6,6} průměr = 5,67

Variabilita dat – směrodatná odchylka

- rozptyl
  - střední hodnota kvadrátů odchylek od střední hodnoty

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \ldots + (x_N - \bar{x})^2}{N - 1}$$

Variabilita dat – směrodatná odchylka

{2,2,3,3,4,5} průměr = 3,17

$$\sigma^2 = \frac{(2-3{,}17)^2 + (2-3{,}17)^2 + (3-3{,}17)^2 + (3-3{,}17)^2}{6-1} +$$

$$+ \frac{(4-3{,}17)^2 + (5-3{,}17)^2}{5} =$$

$$= \frac{1{,}3689 + 1{,}3689 + 0{,}0289 + 0{,}0289 + 0{,}6889 + 3{,}3489}{5} = \frac{6{,}8334}{5} =$$

$$= 1{,}367$$

Variabilita dat – směrodatná odchylka

směrodatná odchylka

{2,2,3,3,4,5} průměr = 3,17

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2} = 1{,}169$$

Variabilita dat – směrodatná odchylka

{2,2,3,3,4,5} průměr = 3,17 SD =1,17

{2,2,3,3,4,20} průměr = 5,67 SD = 7,06

{5,5,6,6,6,6} průměr = 5,67 SD = 0,52

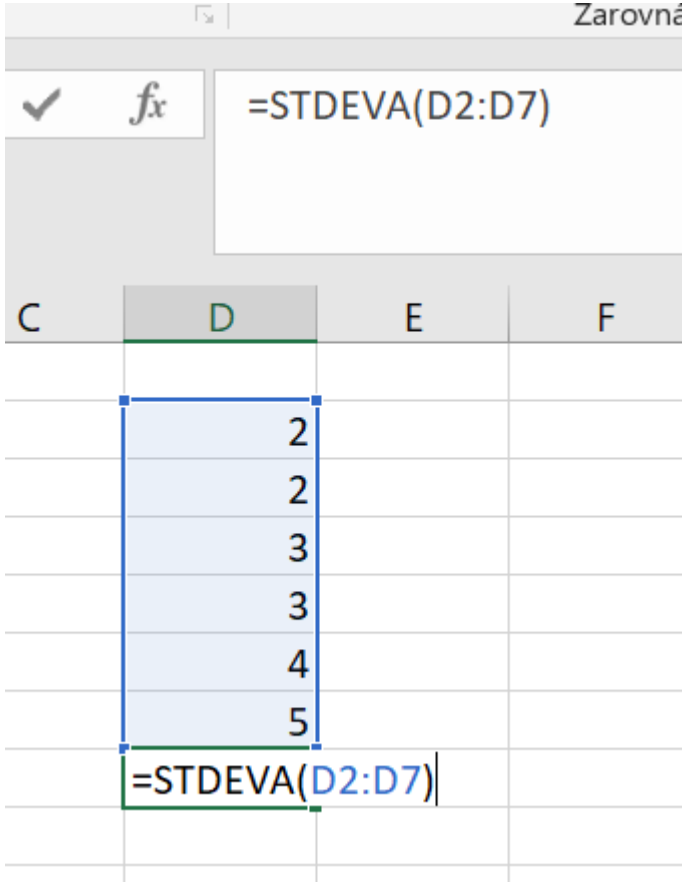
Variabilita dat – směrodatná odchylka

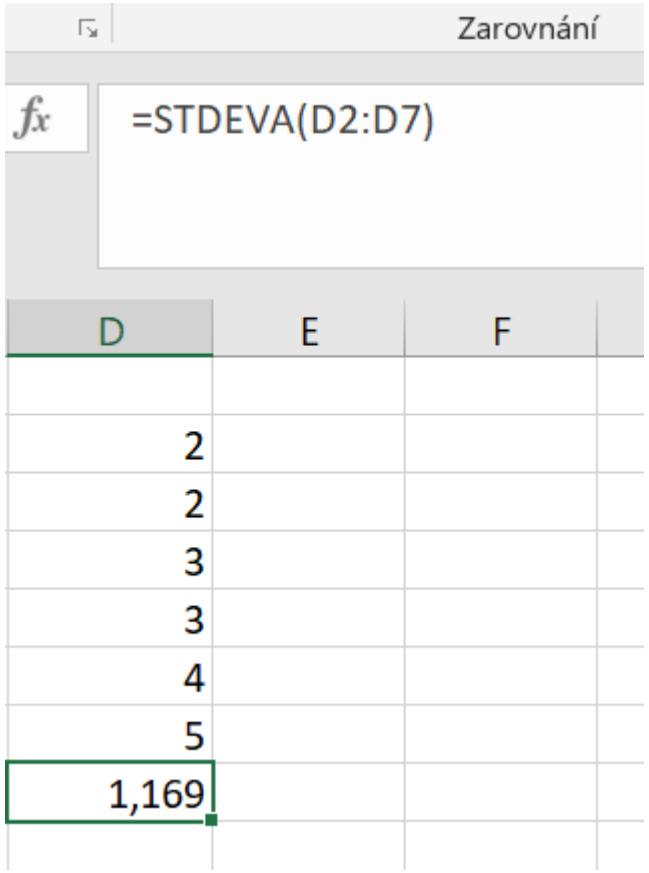
{2,2,3,3,4,5} průměr = 3,17 SD = 1,17

{2,2,3,3,4,20} průměr = 5,67 SD = 7,06

{5,5,6,6,6,6} průměr = 5,67 SD = 0,52

SD v Excelu



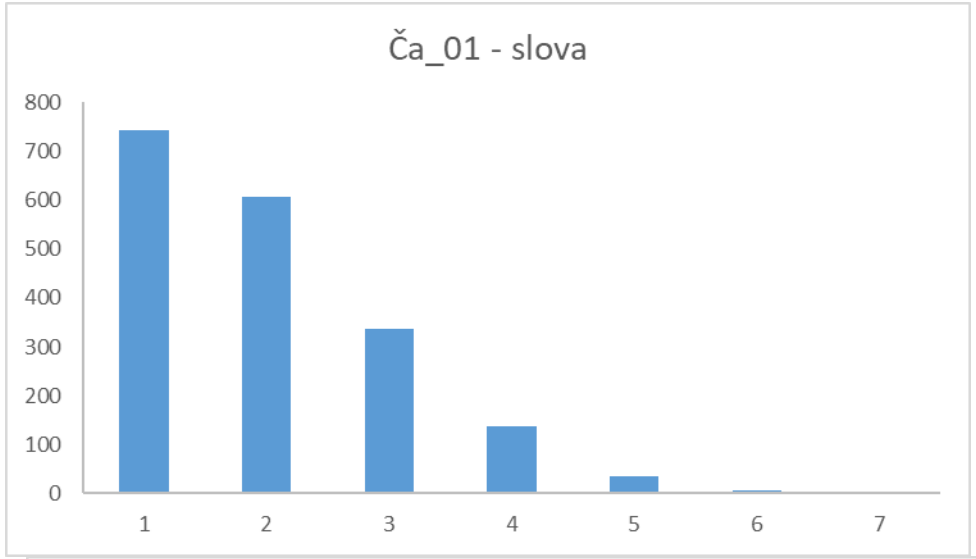


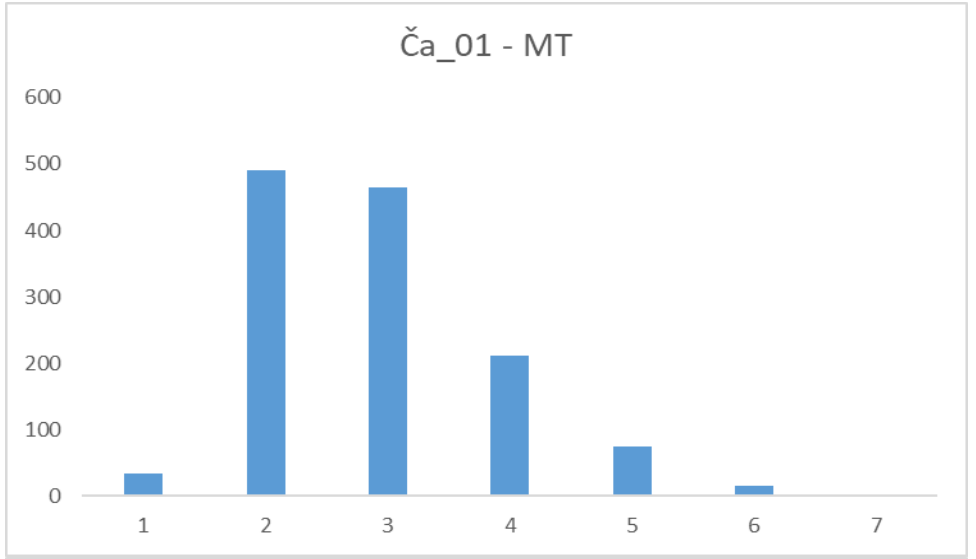
více viz https://support.office.com/cs-cz/article/stdeva-funkce-5ff38888-7ea5-48de-9a6d-11ed73b29e9d

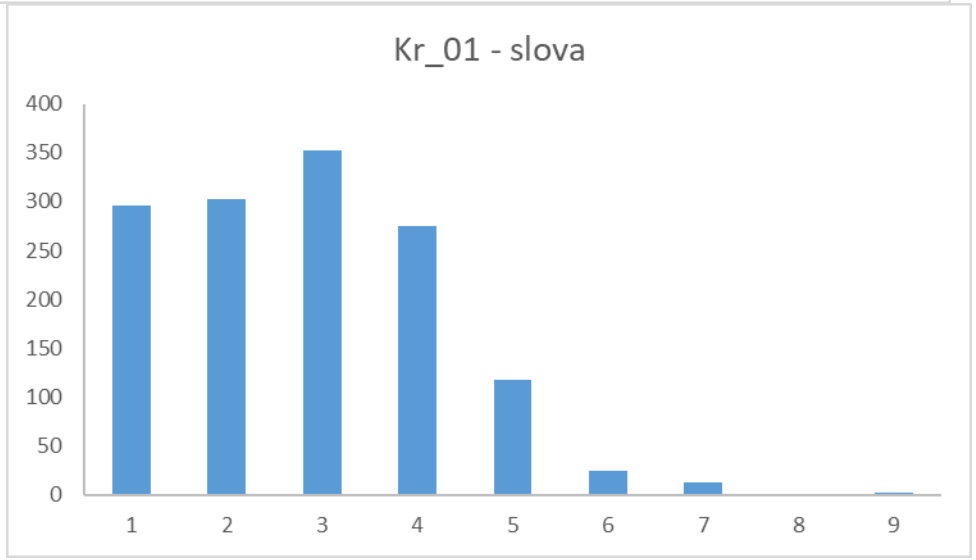
Výsledky – průměrné délky a SD

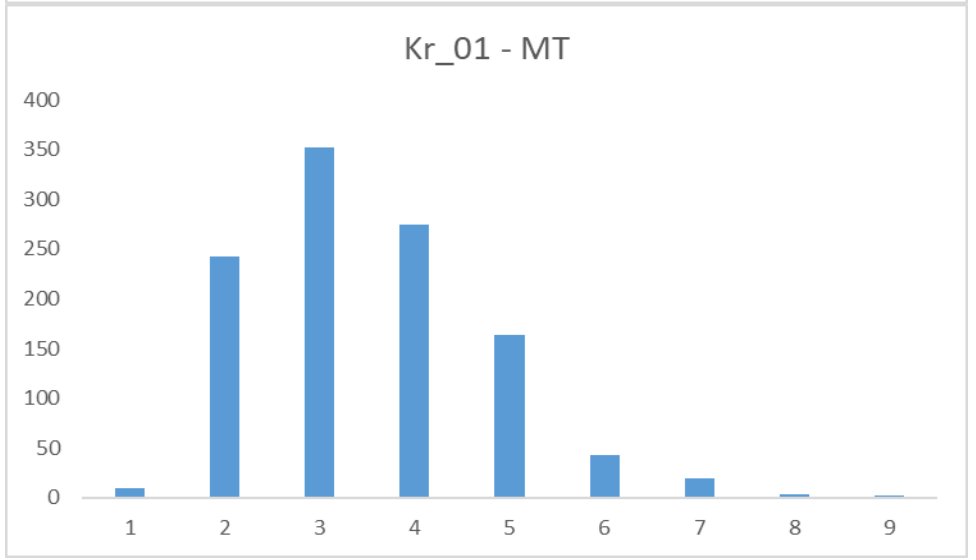
	L_S	SD_S	L_{MT}	SD_{MT}
bel	2	1,05	2,89	1
odb	2,83	1,4	3,51	1,23

Porovnání délek – jeho interpretace & grafické znázornění



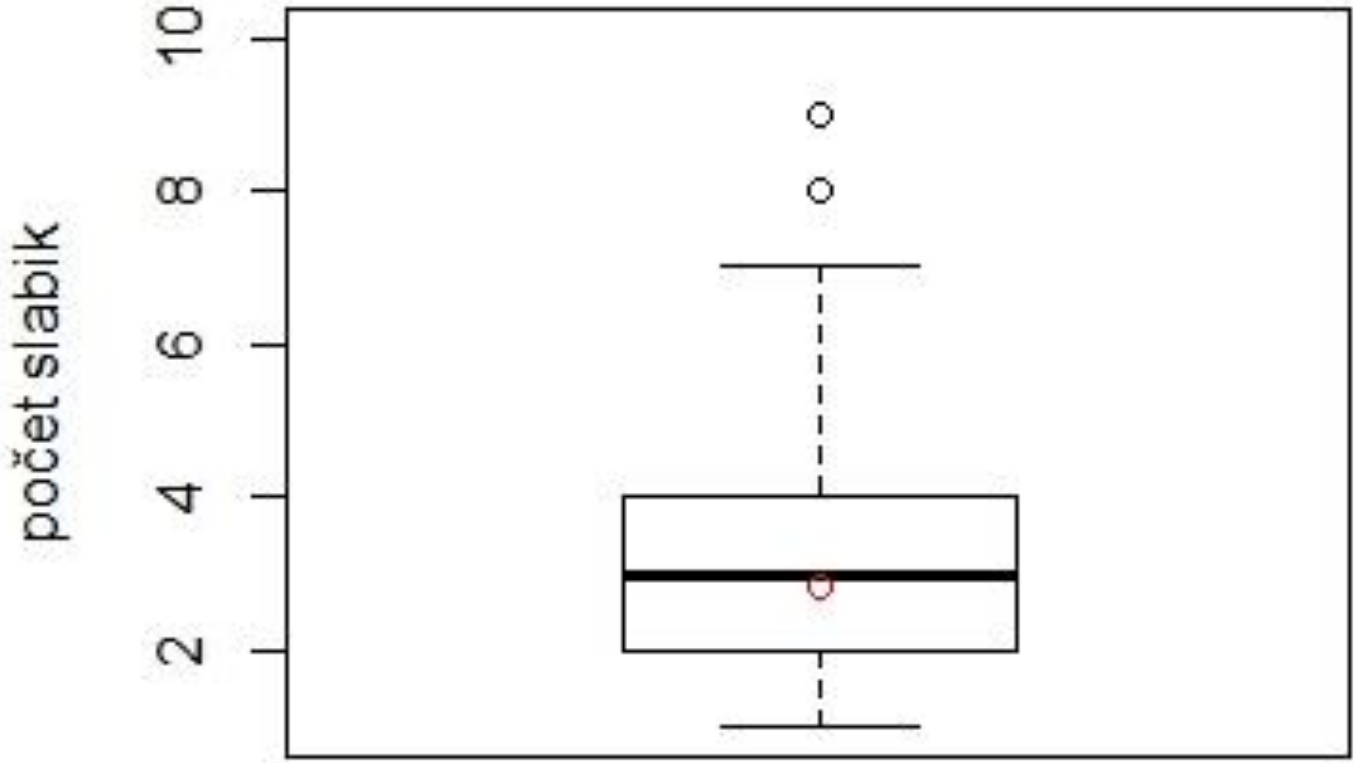




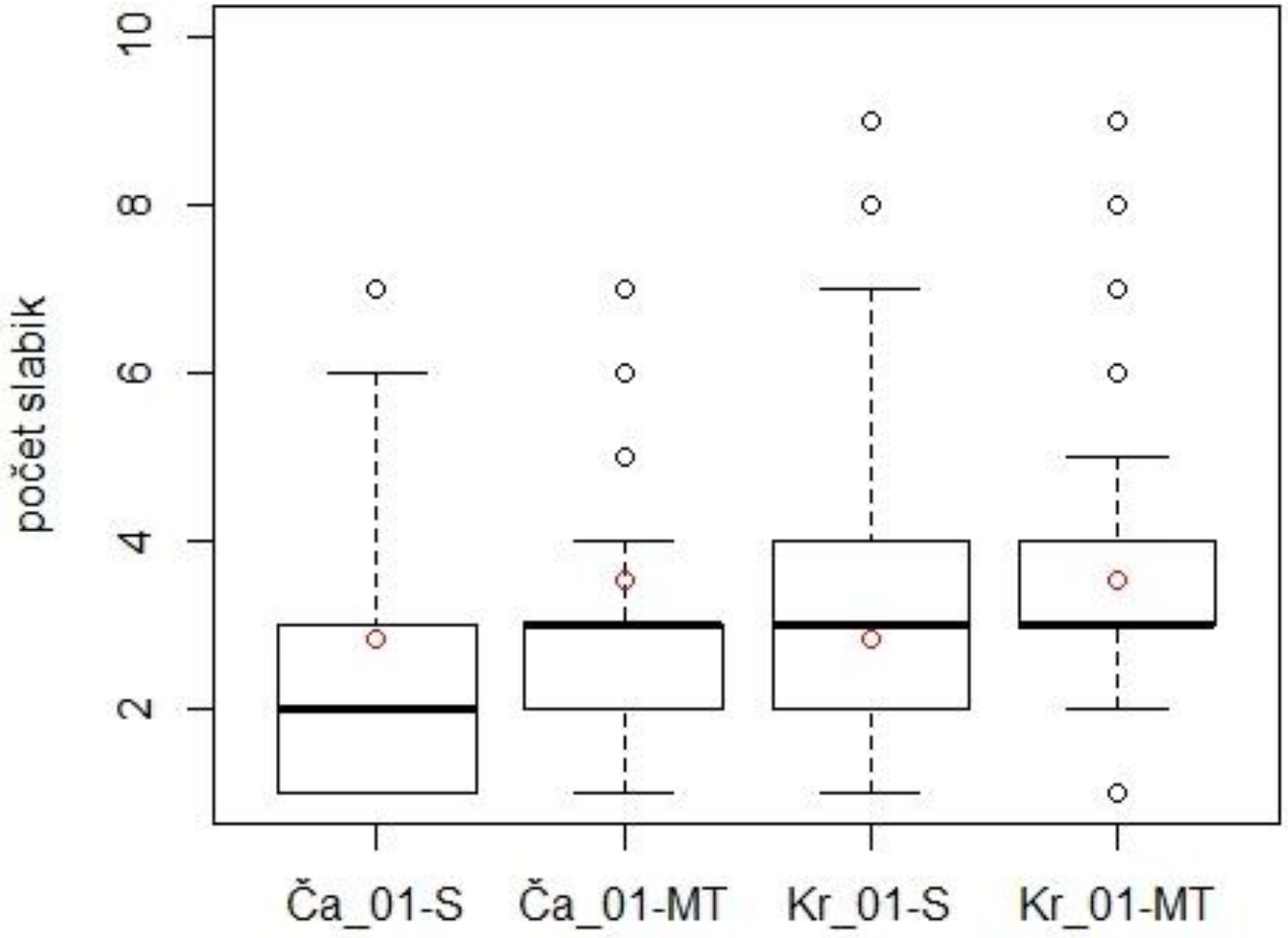


Porovnání délek – jeho interpretace & grafické znázornění

- Kr_01 – MT



Porovnání délek – jeho interpretace & grafické znázornění



Medián

- rozděluje soubor na dvě stejné poloviny
- není ovlivněn extrémními hodnotami

Medián

- rozděluje soubor na dvě stejné poloviny
- není ovlivněn extrémními hodnotami

{2,2,3,3,4,5} průměr = 3,17 medián = 3

{2,2,3,3,4,20} průměr = 5,67 medián = 3

{5,5,6,6,6,6} průměr = 5,67 medián = 6

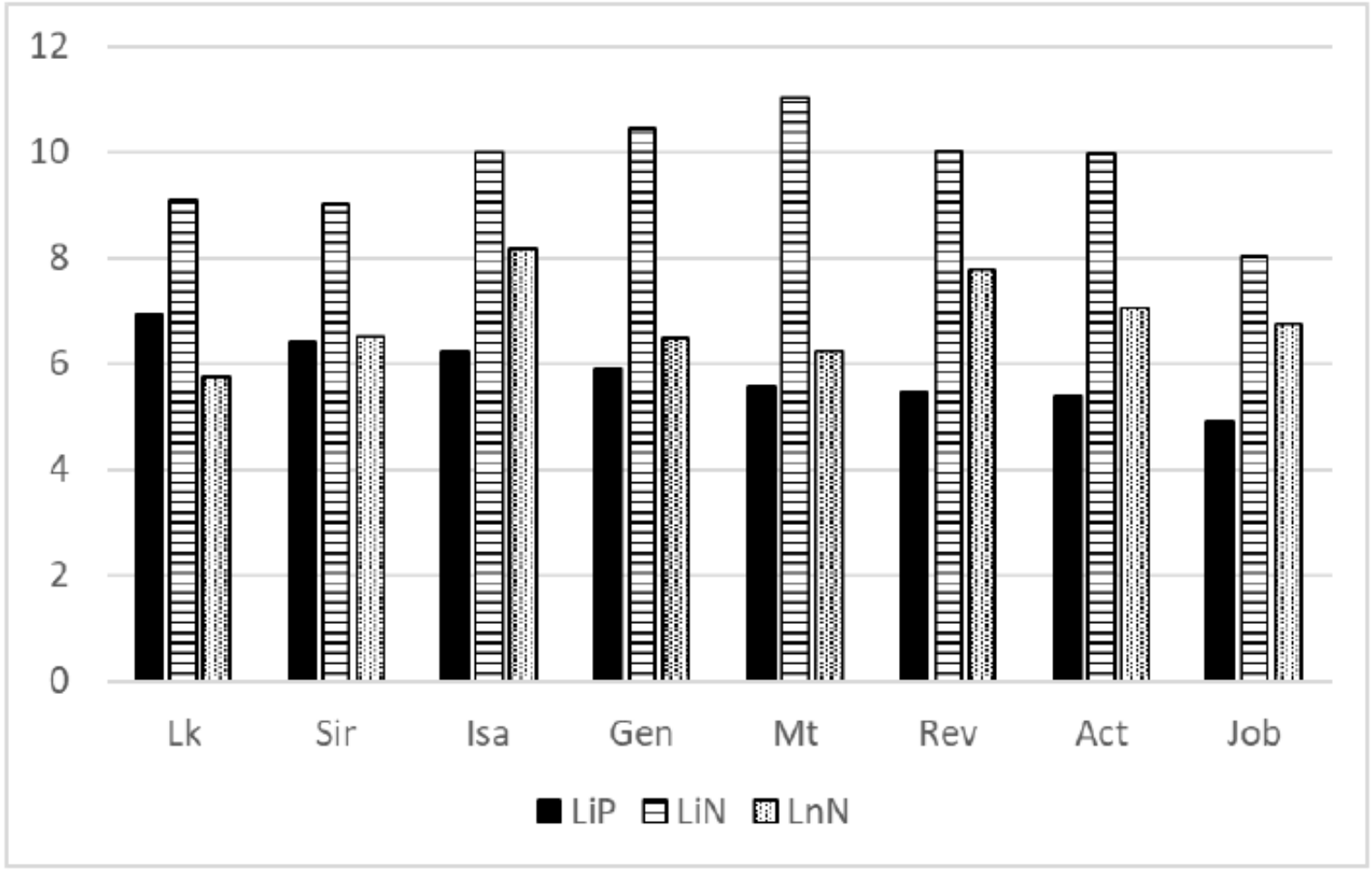
Výsledky – průměrné délky, SD, medián

	L_S	SD_S	M_S	L_{MT}	SD_{MT}	M_{MT}
bel	2	1,05	2	2,89	1	3
odb	2,83	1,4	3	3,51	1,23	3

Vztah délky syntaktické fráze a pozice enklitik

- délka fráze měřena v počtu písmen
- enklitika *sě, mi*
- fráze s enklitikem v postiniciální pozici by měla být v průměru kratší než fráze bez enklitika

Vztah délky syntaktické fráze a pozice enklitik



Vztah délky syntaktické fráze a pozice enklitik

	Lk	Sir	Isa	Gen	Mt	Rev	Act	Job	mean	sd
L_iP	6.94	6.41	6.23	5.91	5.58	5.45	5.4	4.9	**5.9**	2.6
L_iN	9.1	9.02	10	10.45	11.01	10.01	9.96	8.02	**10**	6.7
L_nN	5.75	6.52	8.18	6.48	6.23	7.77	7.06	6.74	**6.9**	3.1

Table 10 Average length of analyzed phrases of *sě*

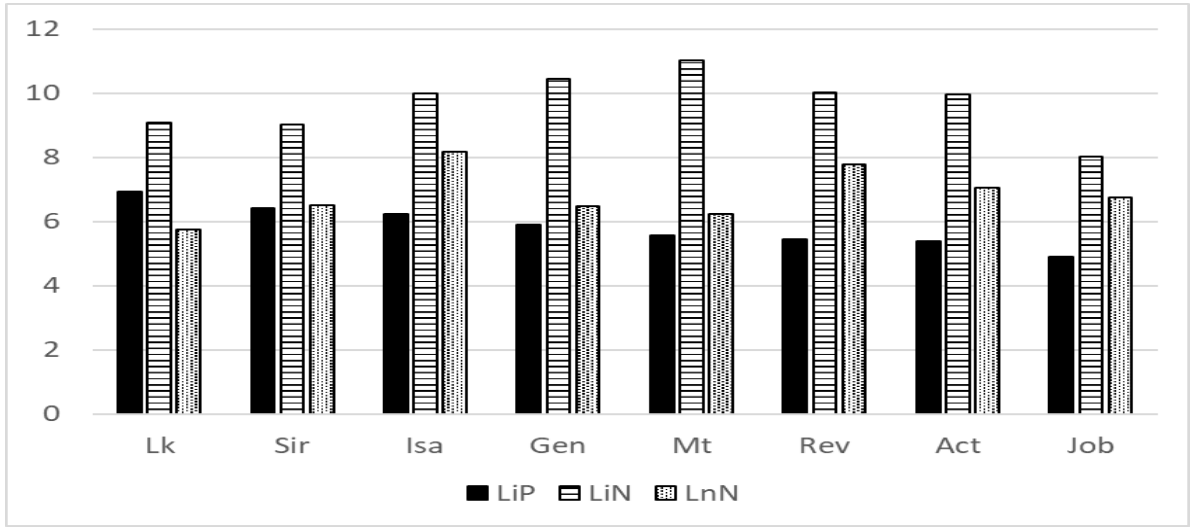


Figure 2 Average length of phrases of *sě* presented in Table 4.

Vztah délky syntaktické fráze a pozice enklitik

	Lk+Sir+Isa+Gen+Mt+Rev+Act+Job	
	mean	**sd**
L_iP	4.82	2.43
L_iN	9.54	6.23
L_nN	6.42	2.04

Table 11 Average length of analyzed phrases of *mi*

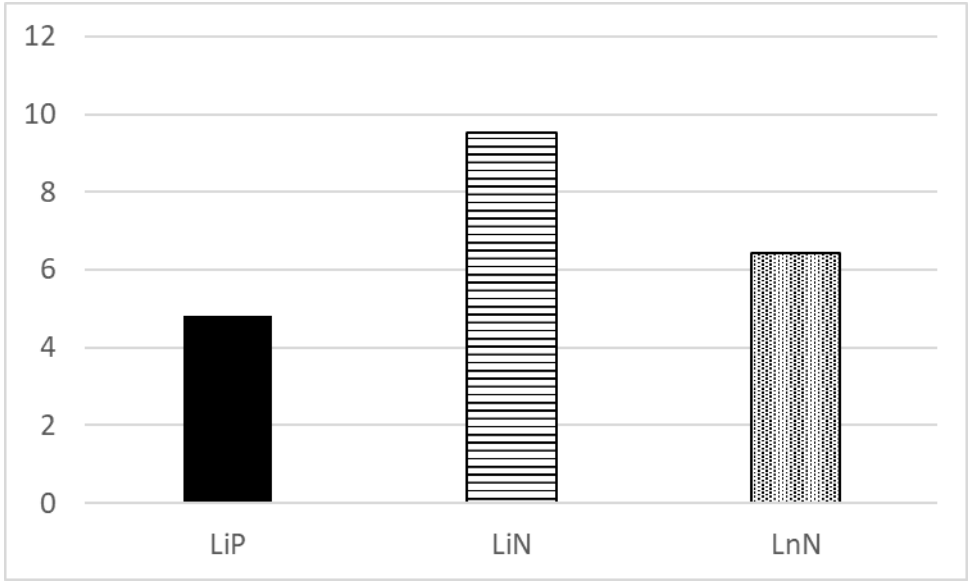
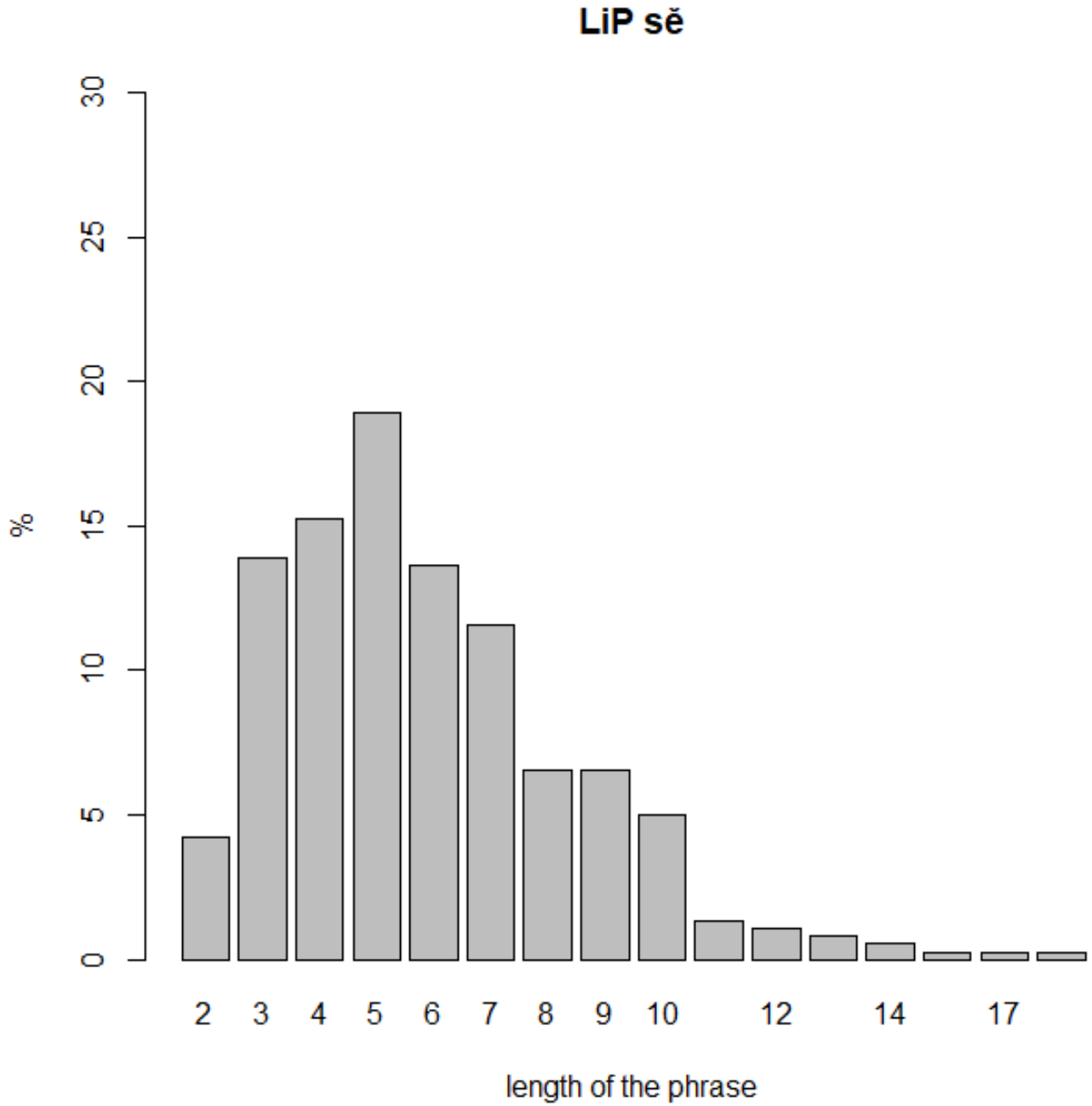
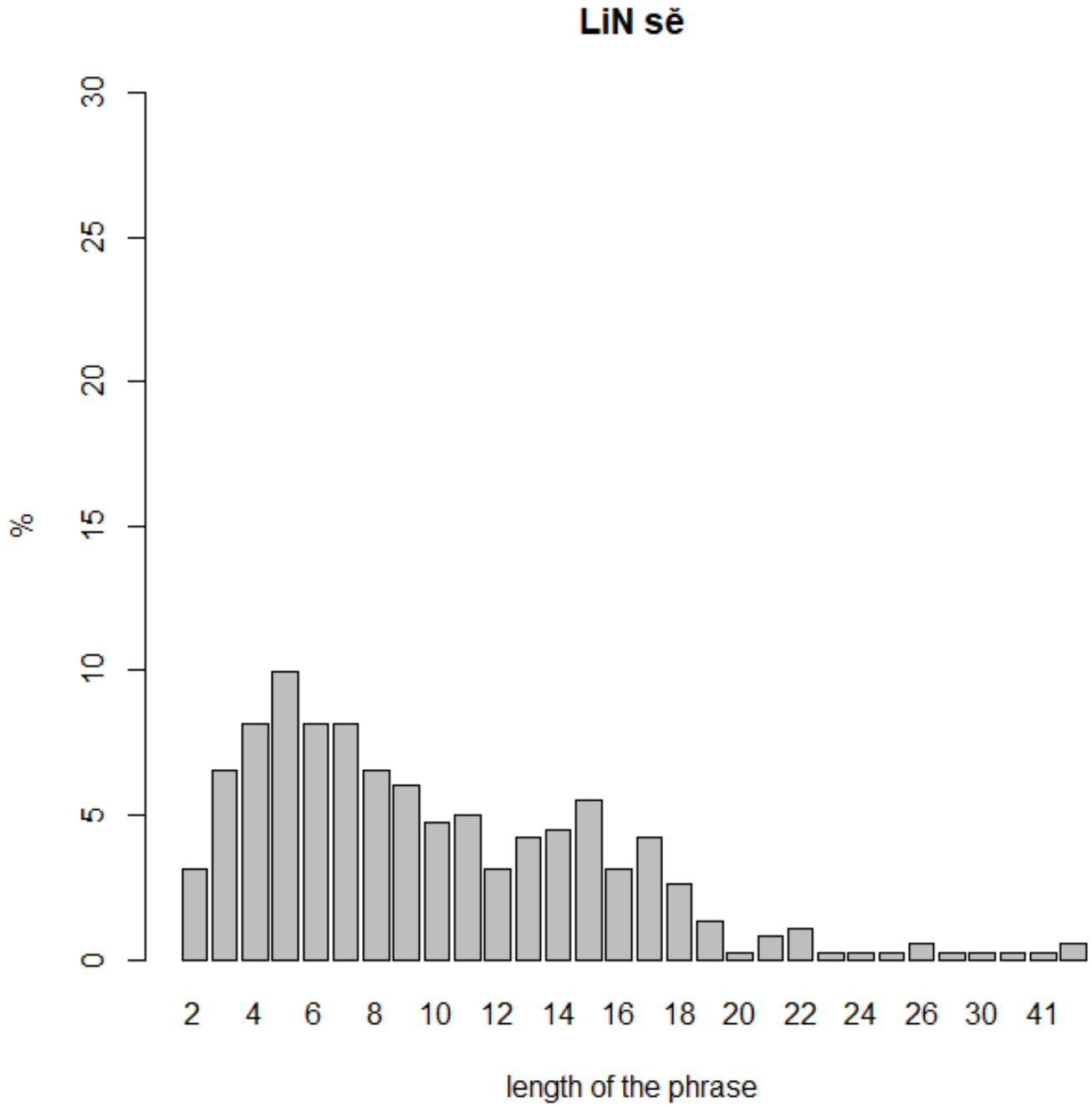
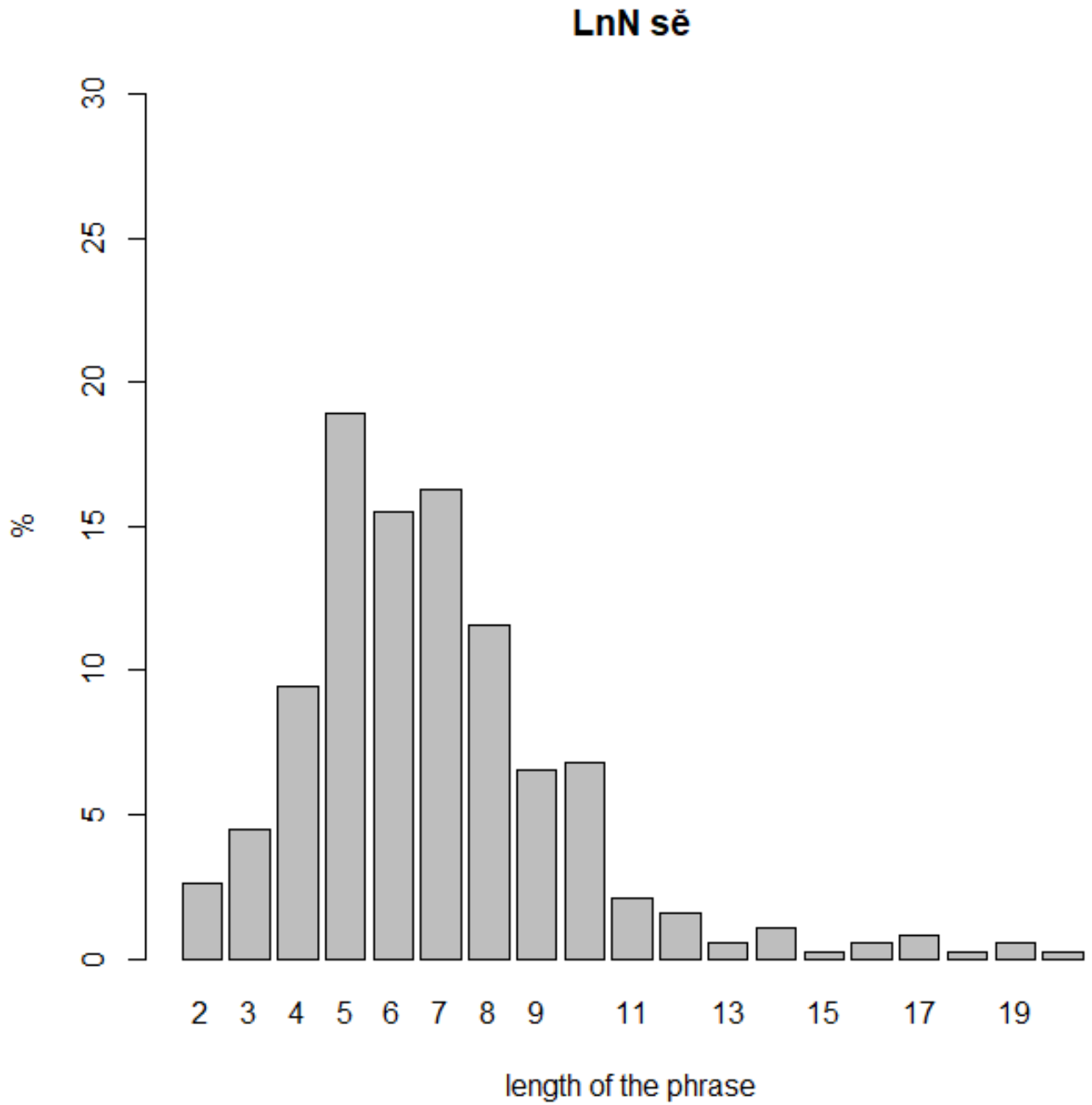
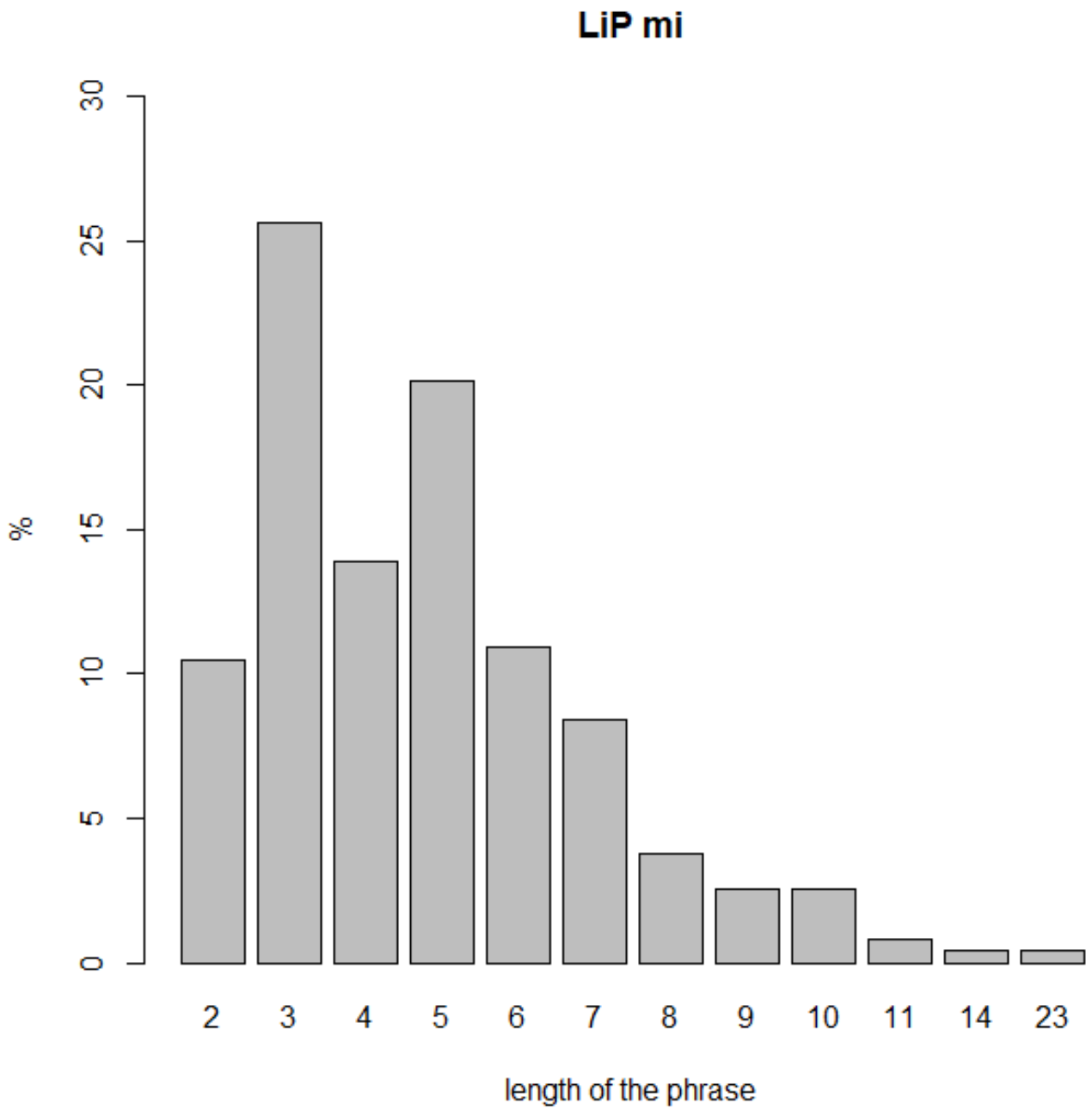


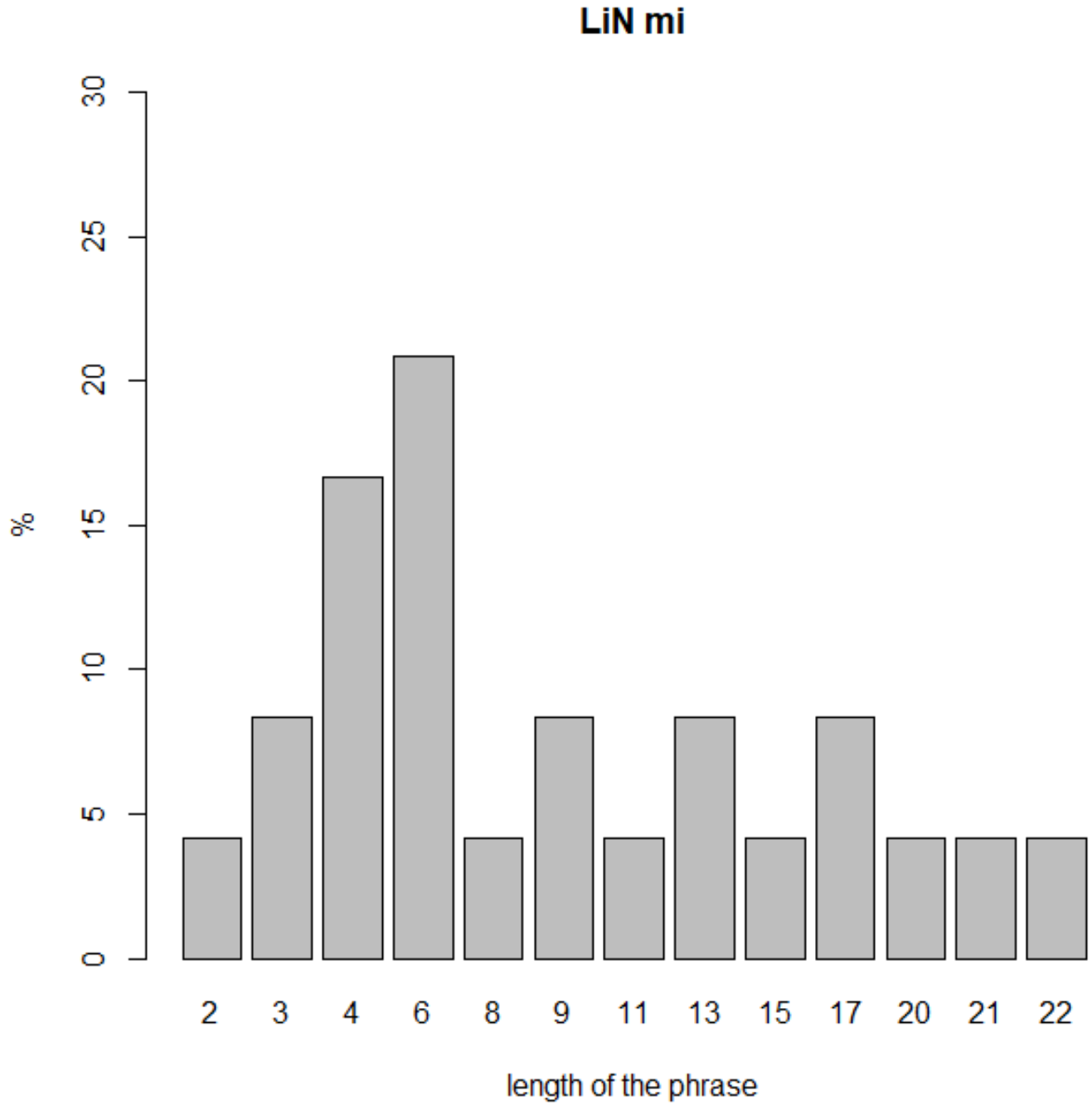
Figure 3 Average length of phrases of *mi* presented in Table 11

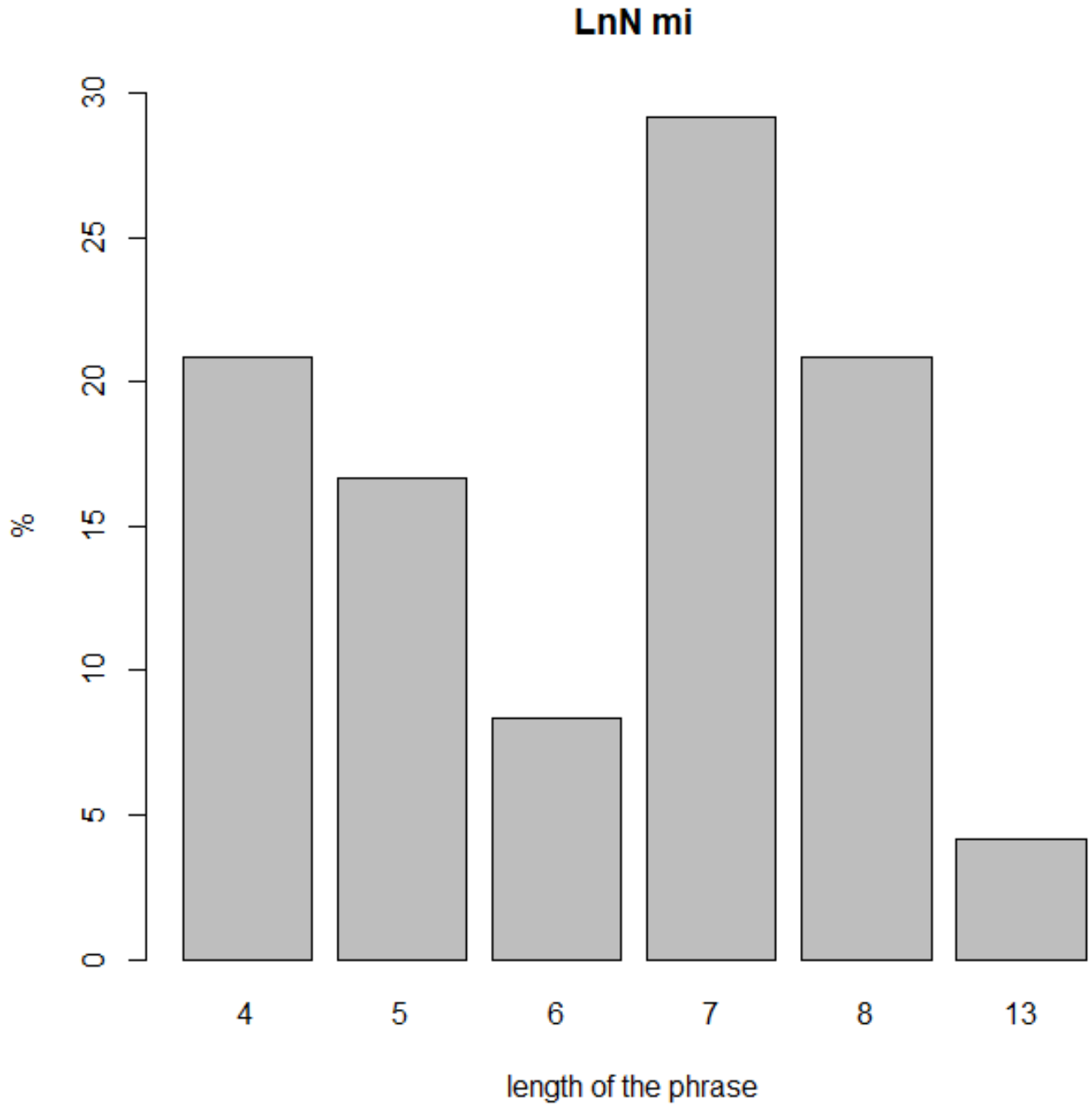

LiP sě
%
0
5
10
15
20
25
30
2
3
4
5
6
7
8
9
10
12
14
17
length of the phrase


LiN sě
%
0
5
10
15
20
25
30
2
4
6
8
10
12
14
16
18
20
22
24
26
30
41
length of the phrase


LnN sĕ
%
0
5
10
15
20
25
30
2
3
4
5
6
7
8
9
11
13
15
17
19
length of the phrase


LiP mi
%
0
5
10
15
20
25
30
2
3
4
5
6
7
8
9
10
11
14
23
length of the phrase


LiN mi
%
0
5
10
15
20
25
30
2
3
4
6
8
9
11
13
15
17
20
21
22
length of the phrase


LnN mi
%
0
5
10
15
20
25
30
4
5
6
7
8
13
length of the phrase

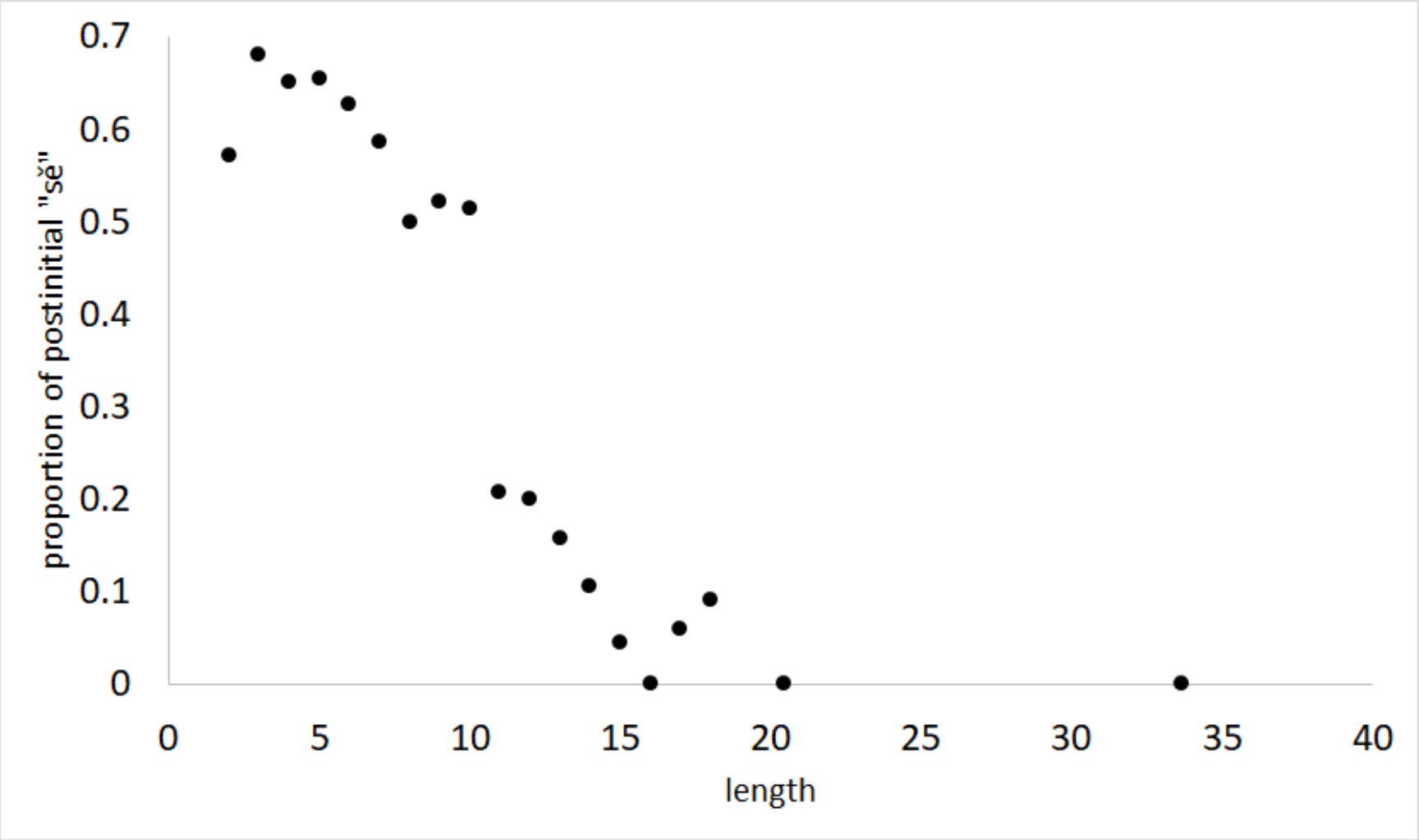
Za hranice popisu…

Za hranice popisu… k testování hypotéz

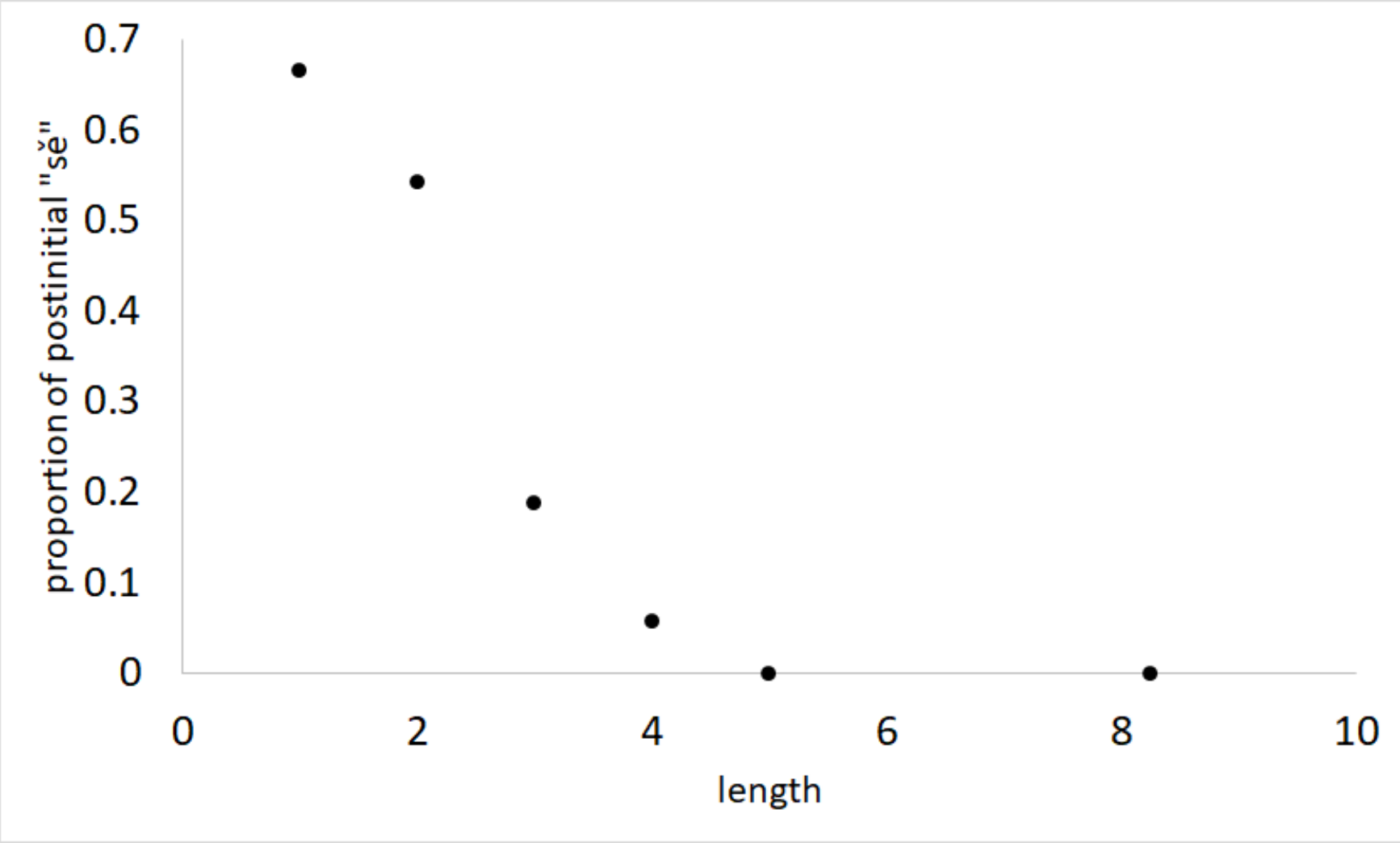
- teoretická zdůvodnění

- hypotéza: čím je iniciální fráze delší, tím menší je pravděpodobnost, že se za ní vyskytne enklitikon

Results - letters



Results - words



Porovnání délek – jeho interpretace

- test…