Kritická práce s daty 1

Radek Čech

Program

- obecný úvod
- deskriptivní statistika
- hypotézy a jejich testování
- možnosti textových analýz

Jazyková data & kvantitativní analýza

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

Jazyková data & kvantitativní analýza

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

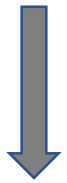
- teoretické problémy….
  - **co** se vlastně modeluje?

Modely jazykového chování a jejich interpretace

jazykové chování
(texty, promluvy)

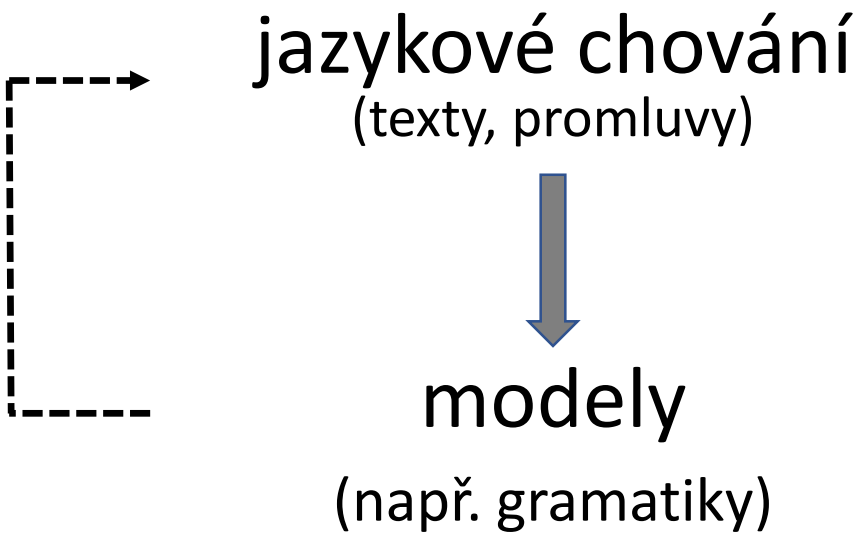
Modely jazykového chování a jejich interpretace

jazykové chování
(texty, promluvy)

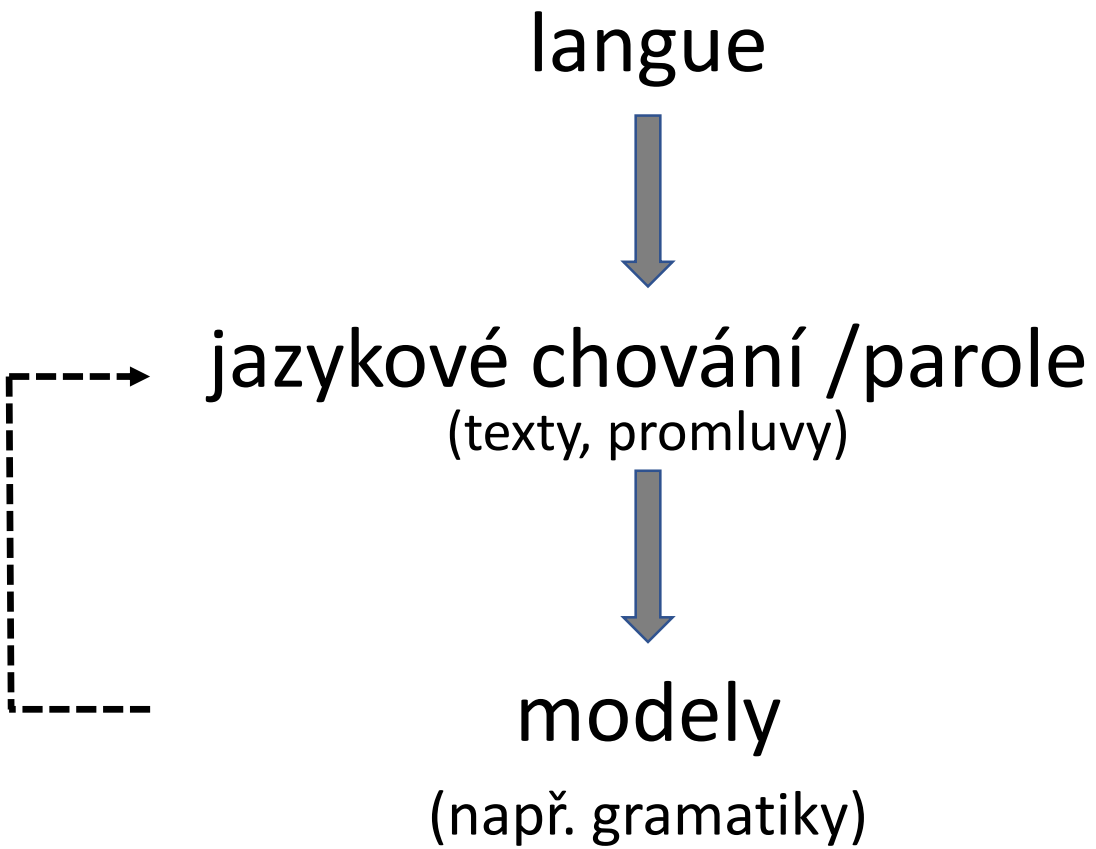


modely
(např. gramatiky)

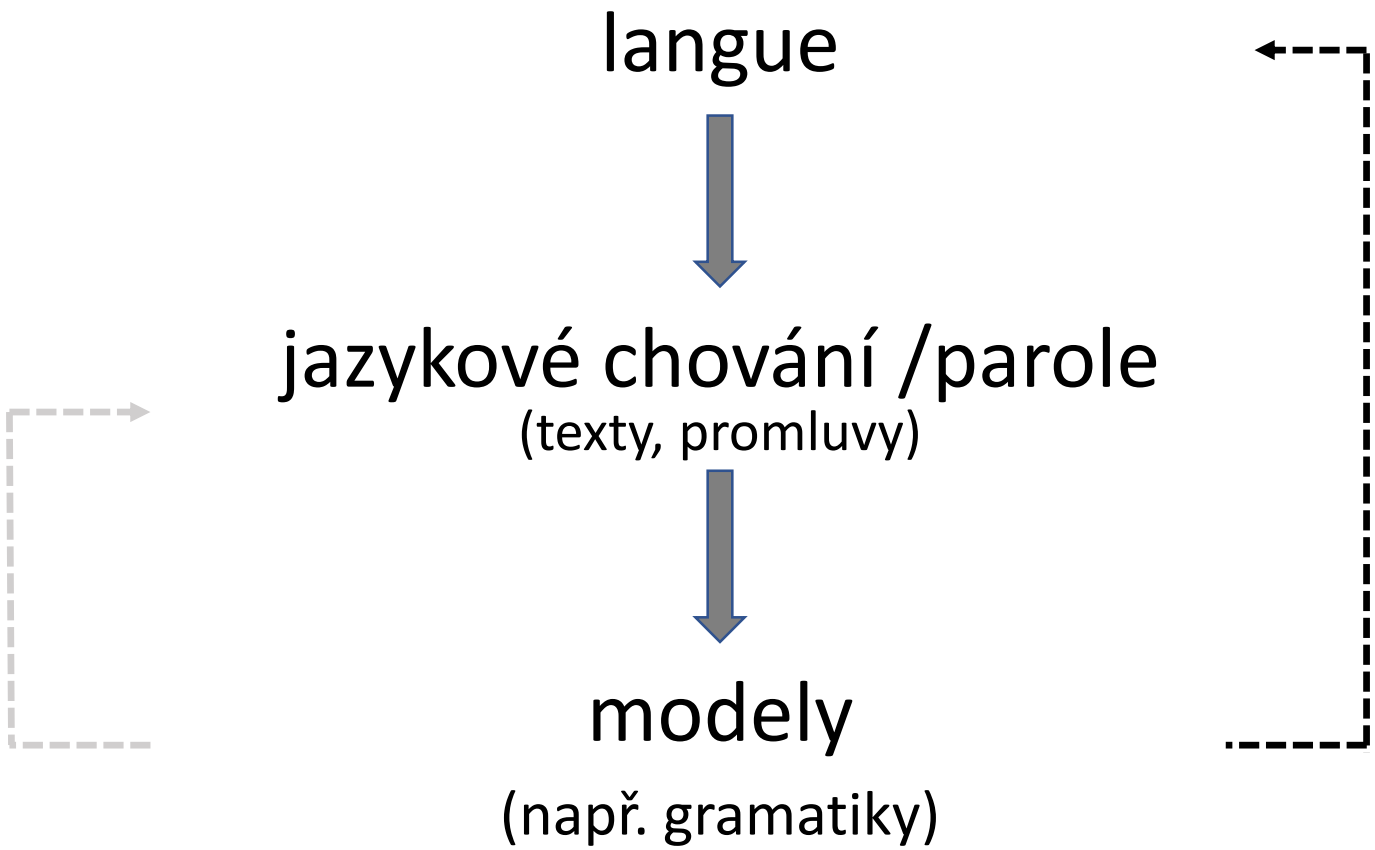
Modely jazykového chování a jejich interpretace



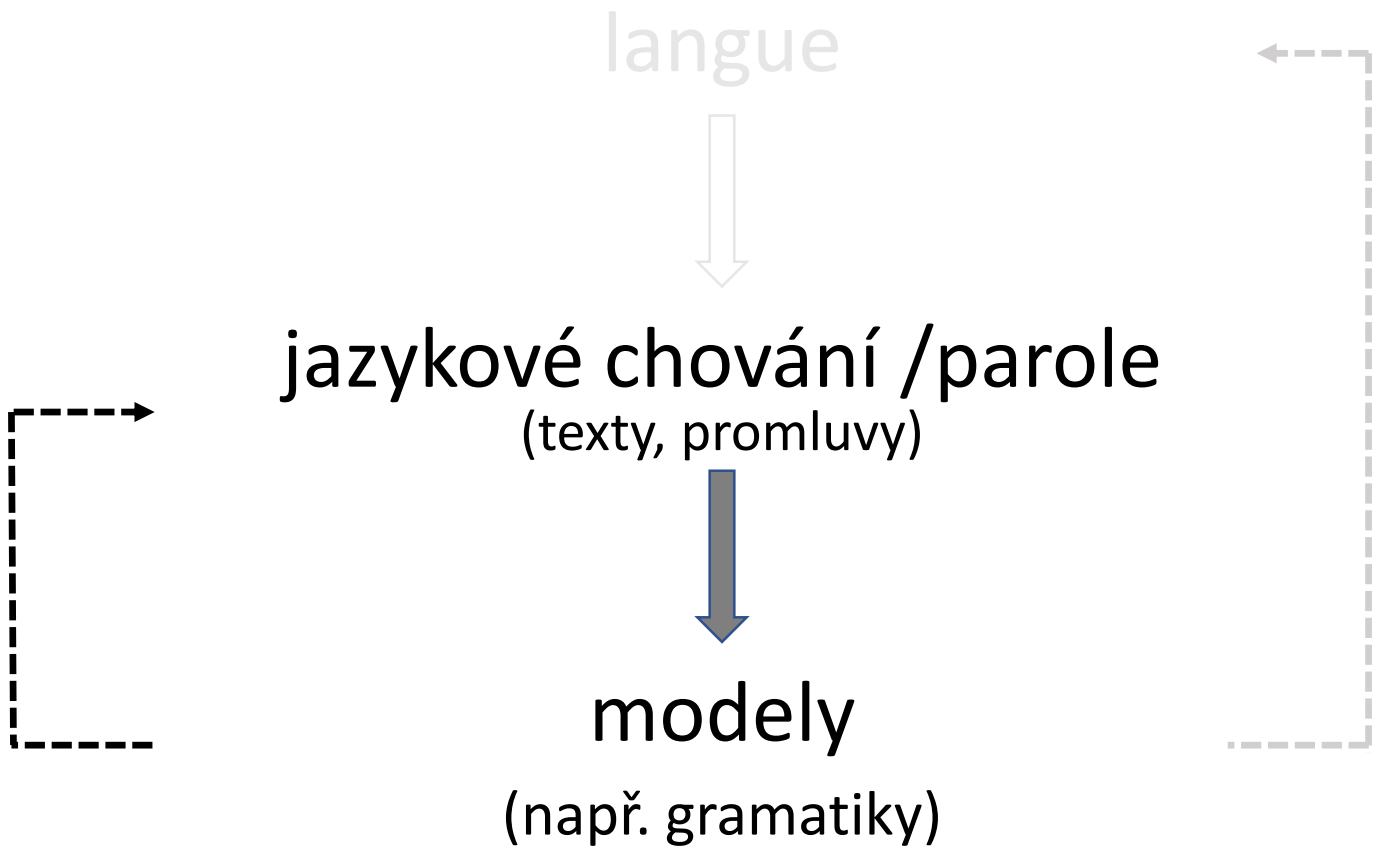
Modely jazykového chování a jejich interpretace



Modely jazykového chování a jejich interpretace



Modely jazykového chování a jejich interpretace



Modely jazykového chování a jejich interpretace

- jazykové chování
  - dynamika, „nestabilita“

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)

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

Stochastické pojetí jazyka

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

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.

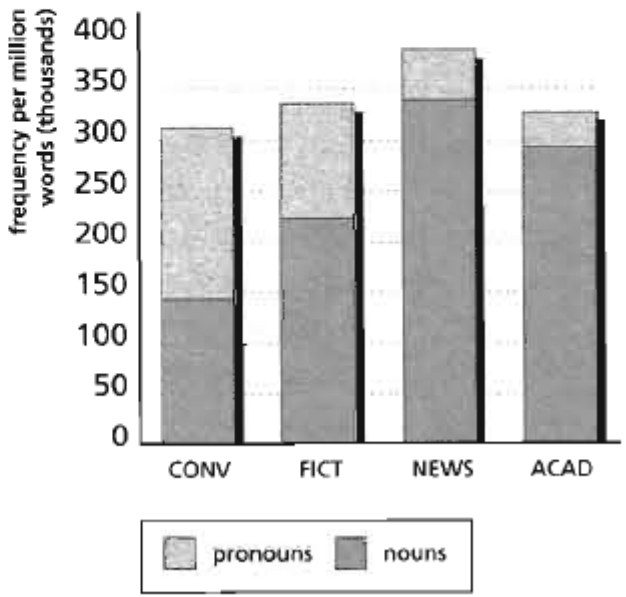
- The relative frequency of nouns is much higher in object position and as a complement or object of a preposition than in subject position.

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

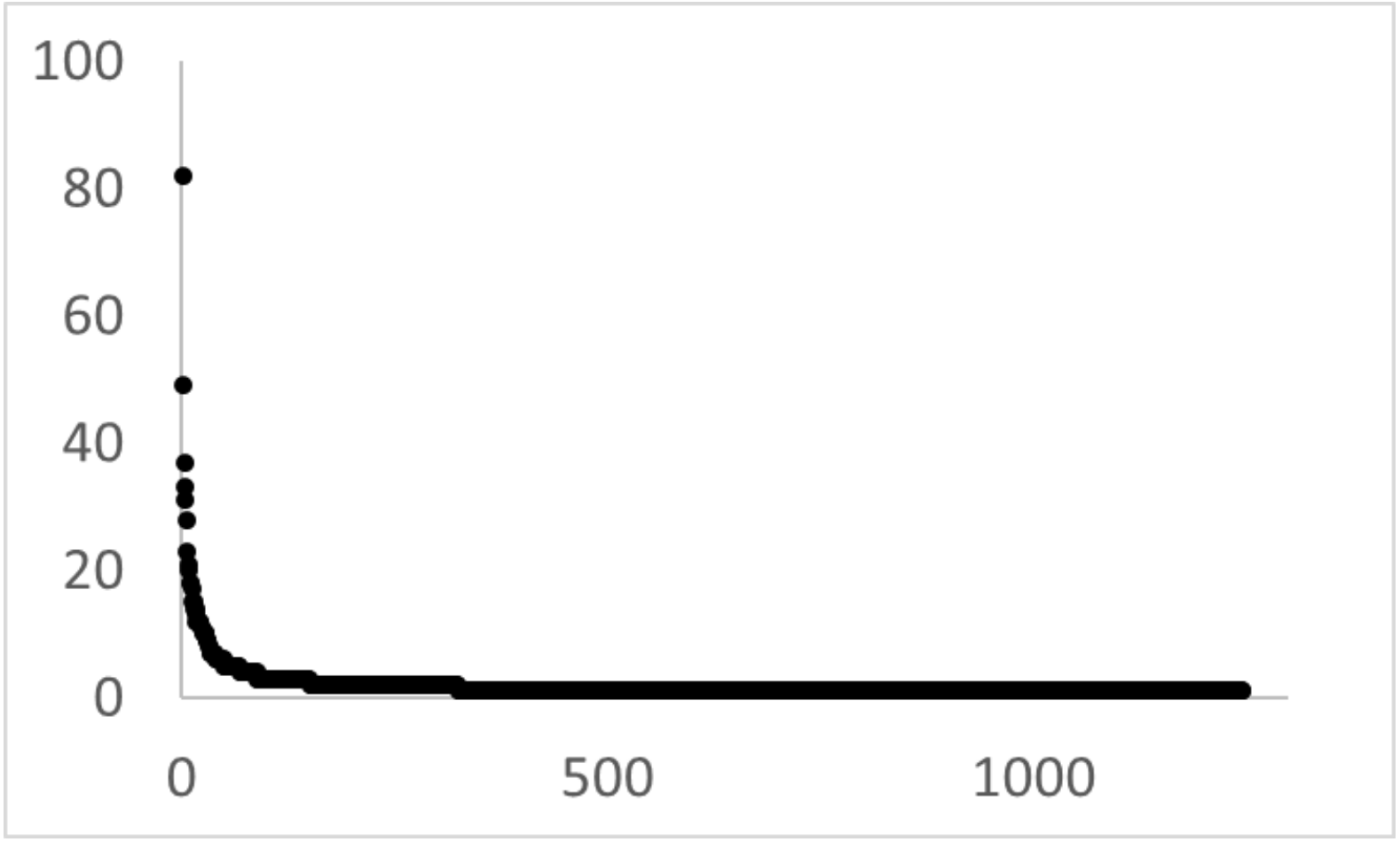
- relativně běžná současná praxe
  - ověřování mechanismů **bez** hlubšího teoretického rámce
    - komputační lingvistika
    - ad hoc analýzy

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
    - …

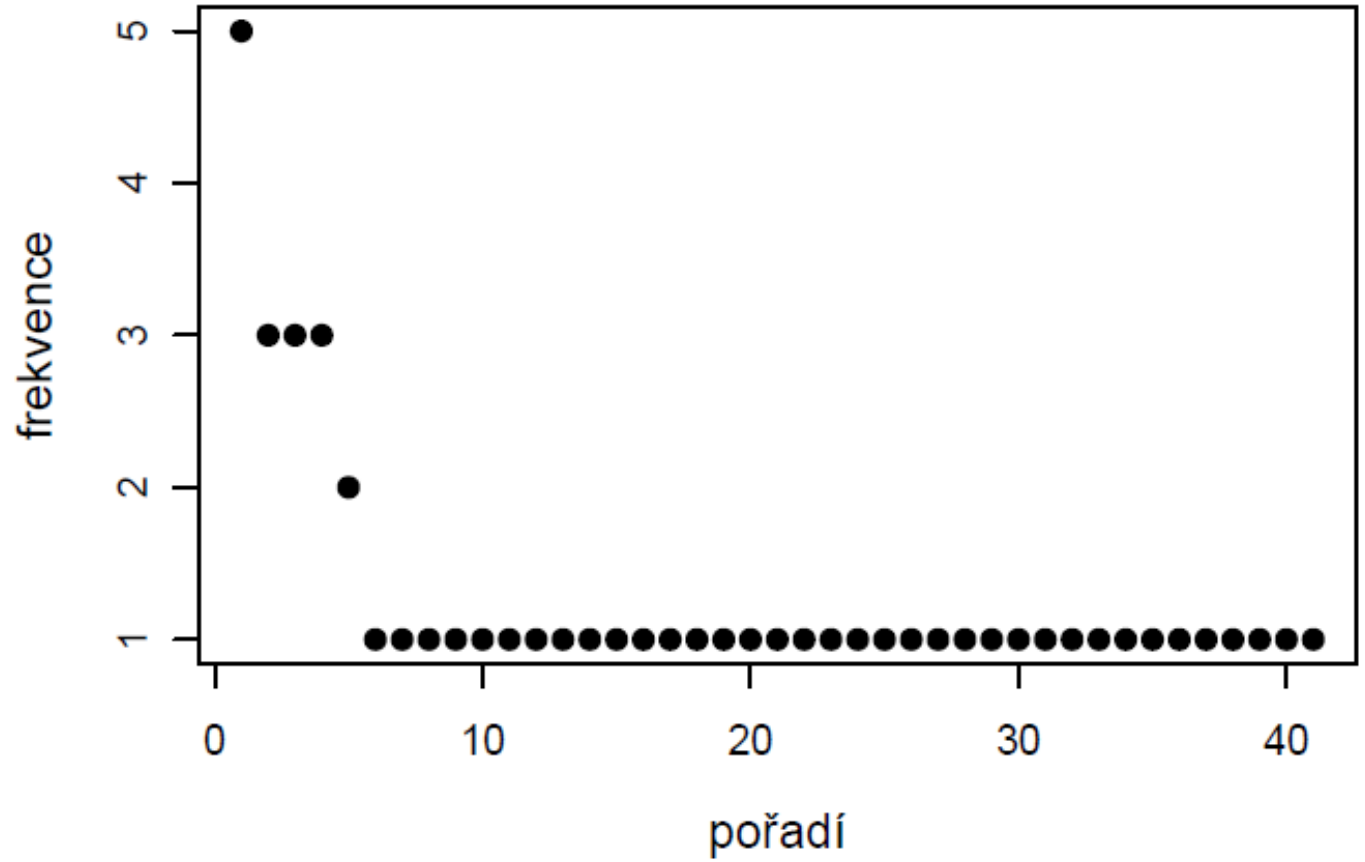
Distribuce jednotek

- Havel 1990: ranková frekvenční distribuce slov



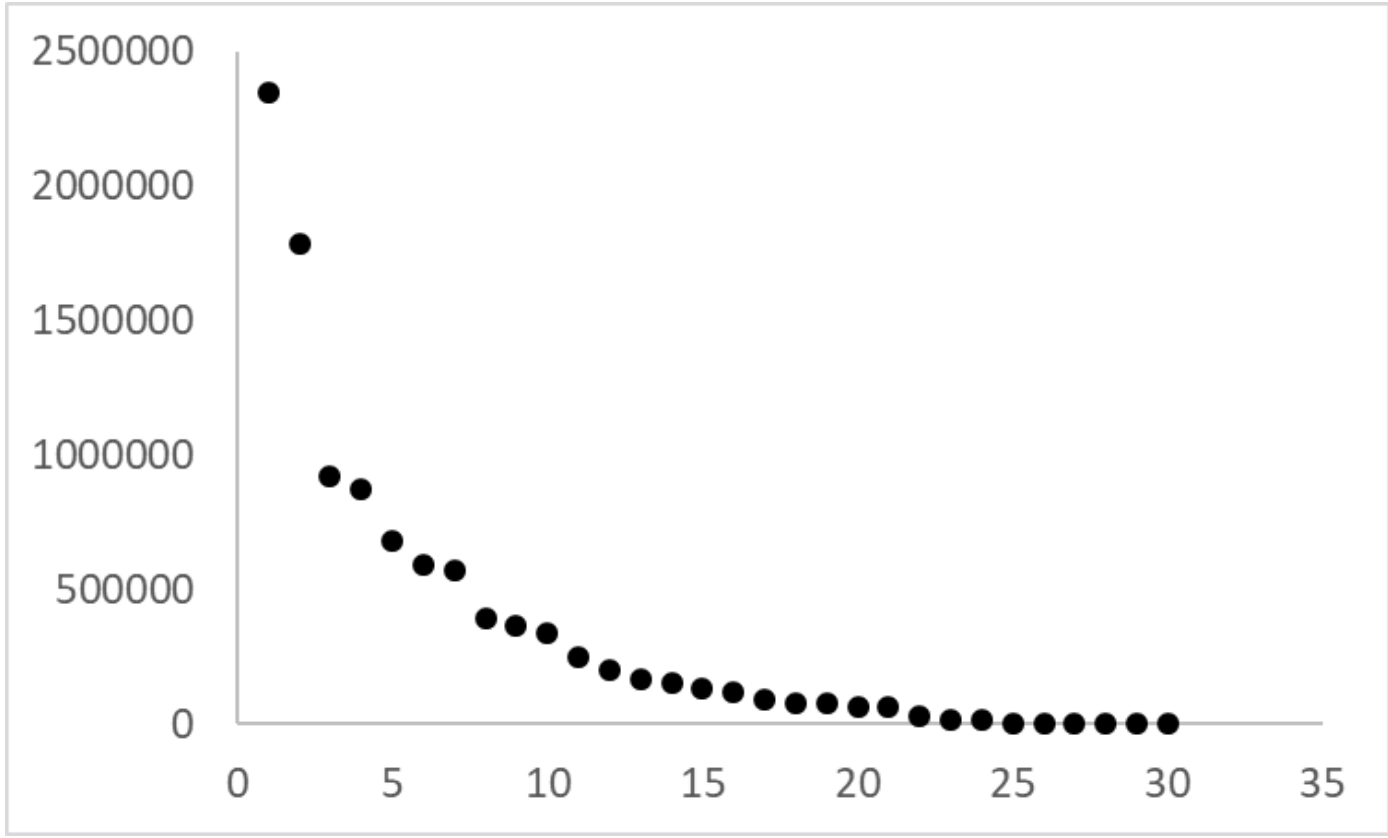
Distribuce jednotek

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



Distribuce jednotek

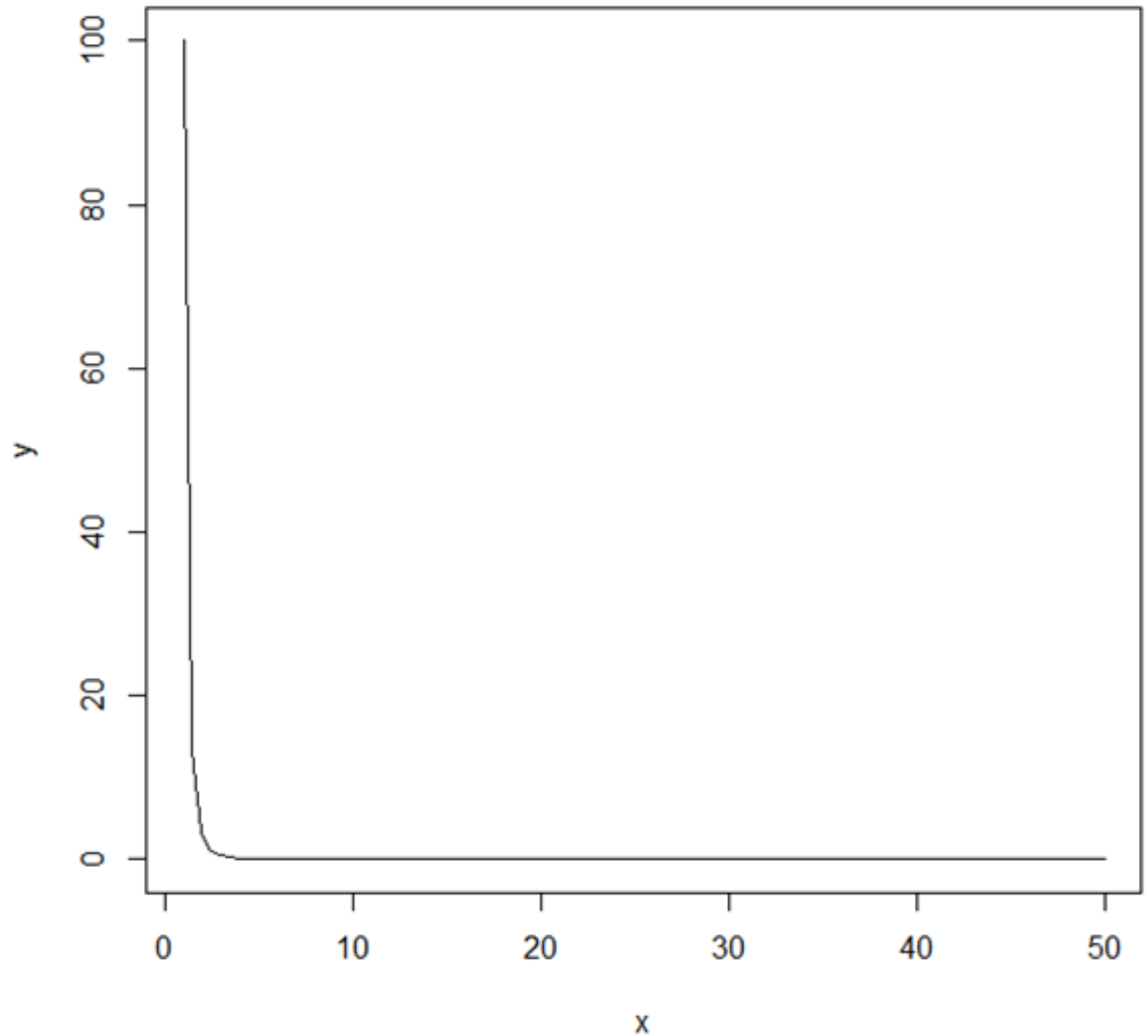
- SYN2005: : ranková frekvenční distribuce primárních předložek



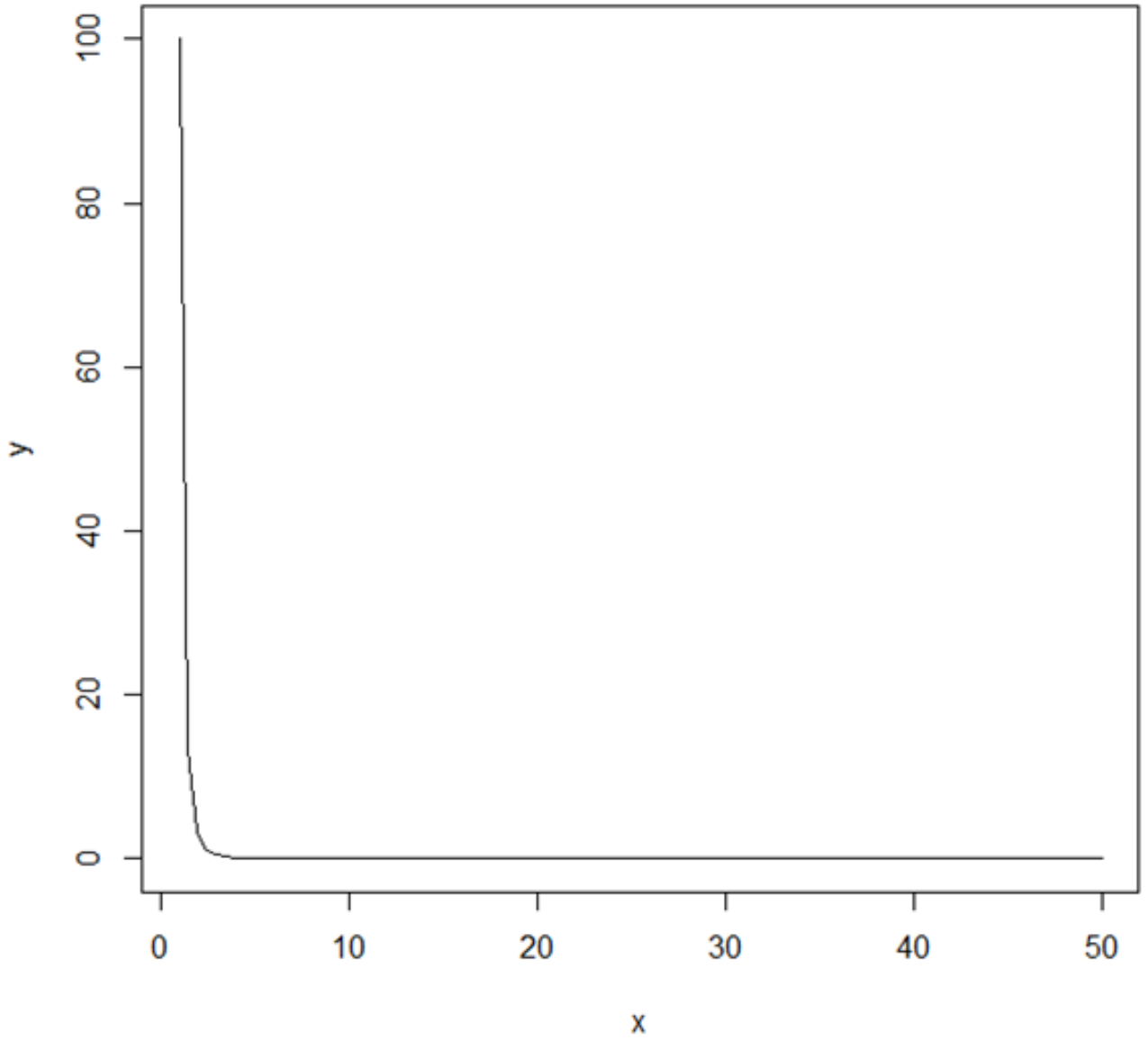
Model – mocninná funkce

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

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

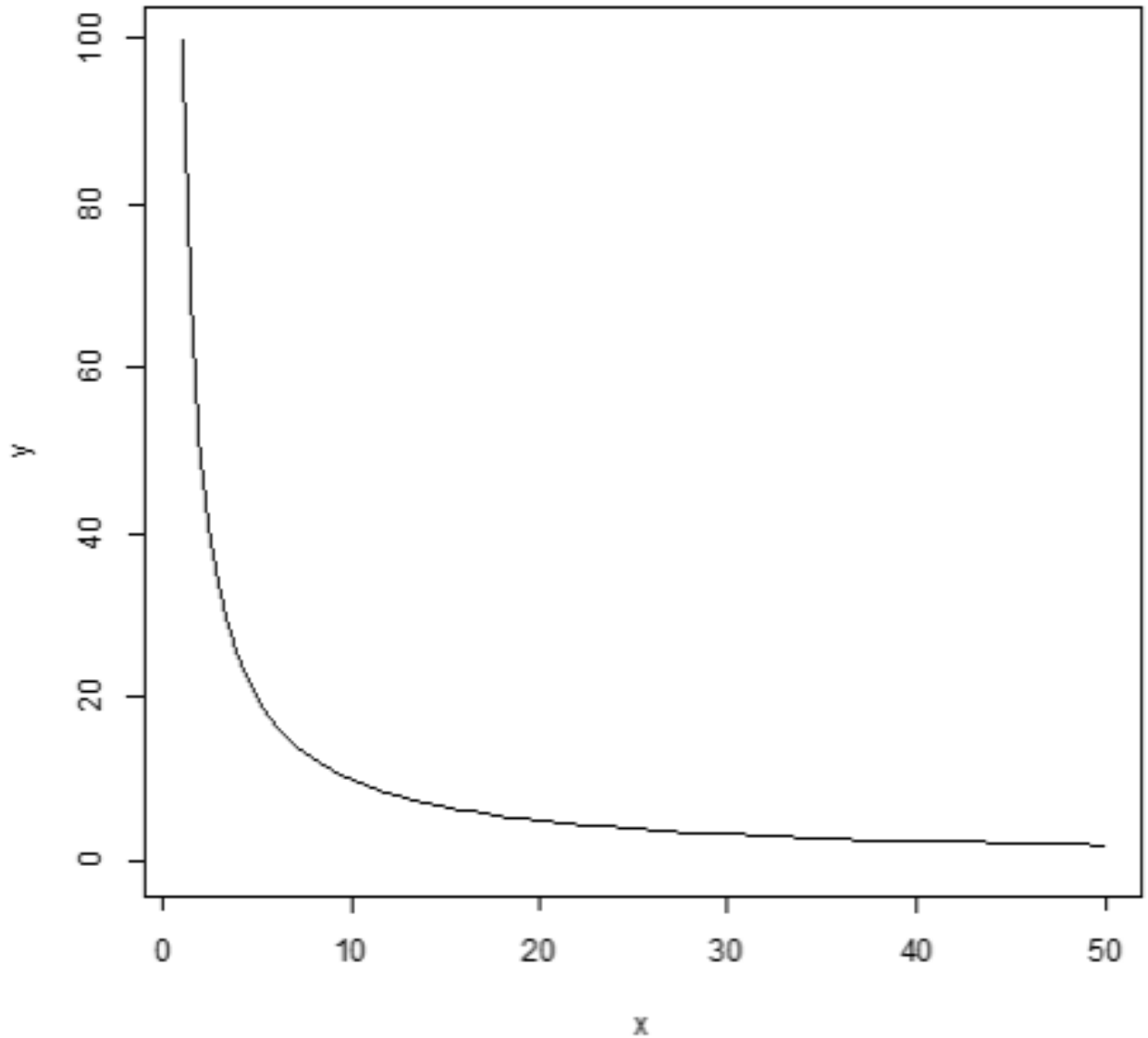


$$y = 100x^{-5}$$

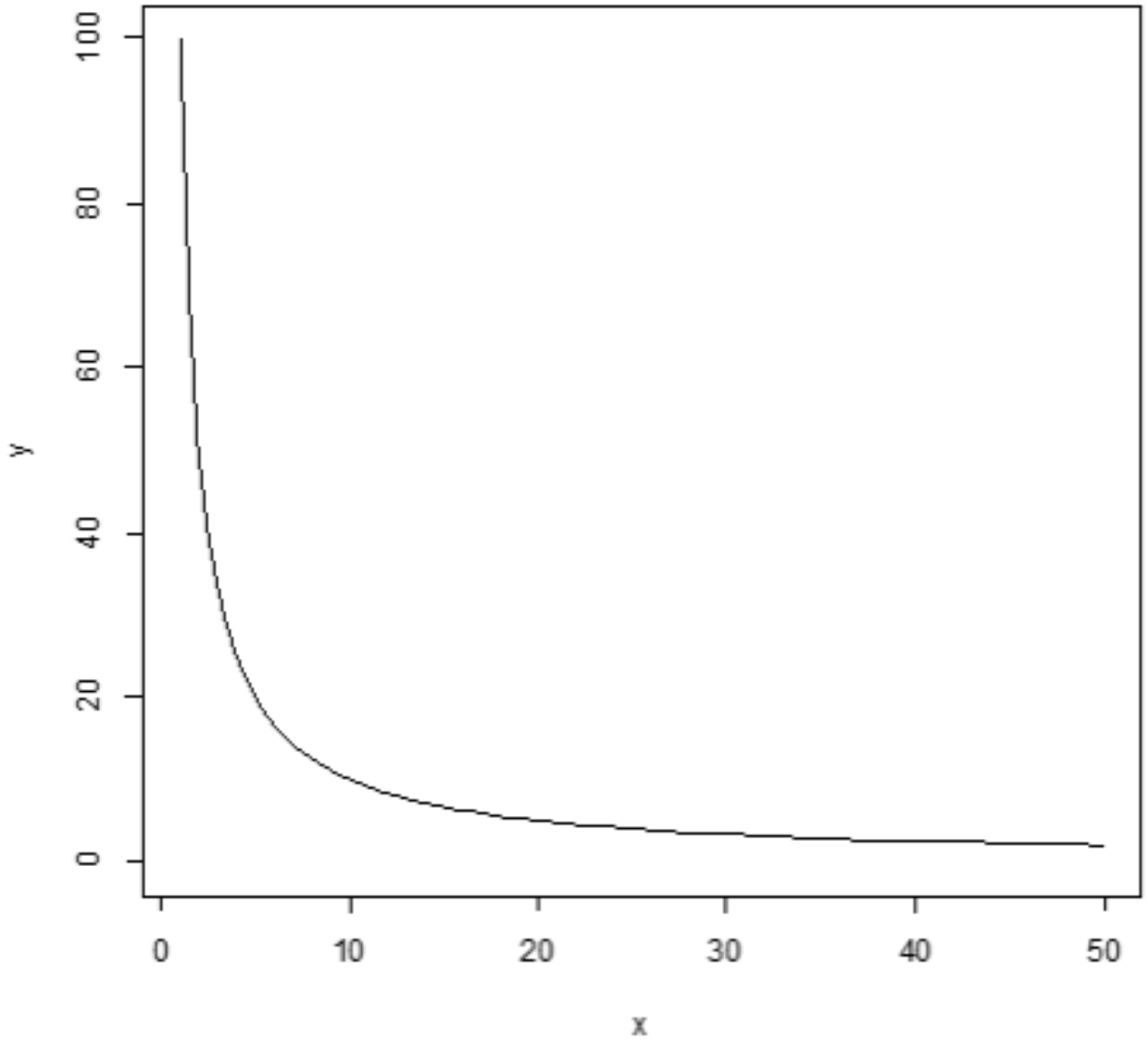


- bohatý slovník
  - většina slov se neopakuje

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

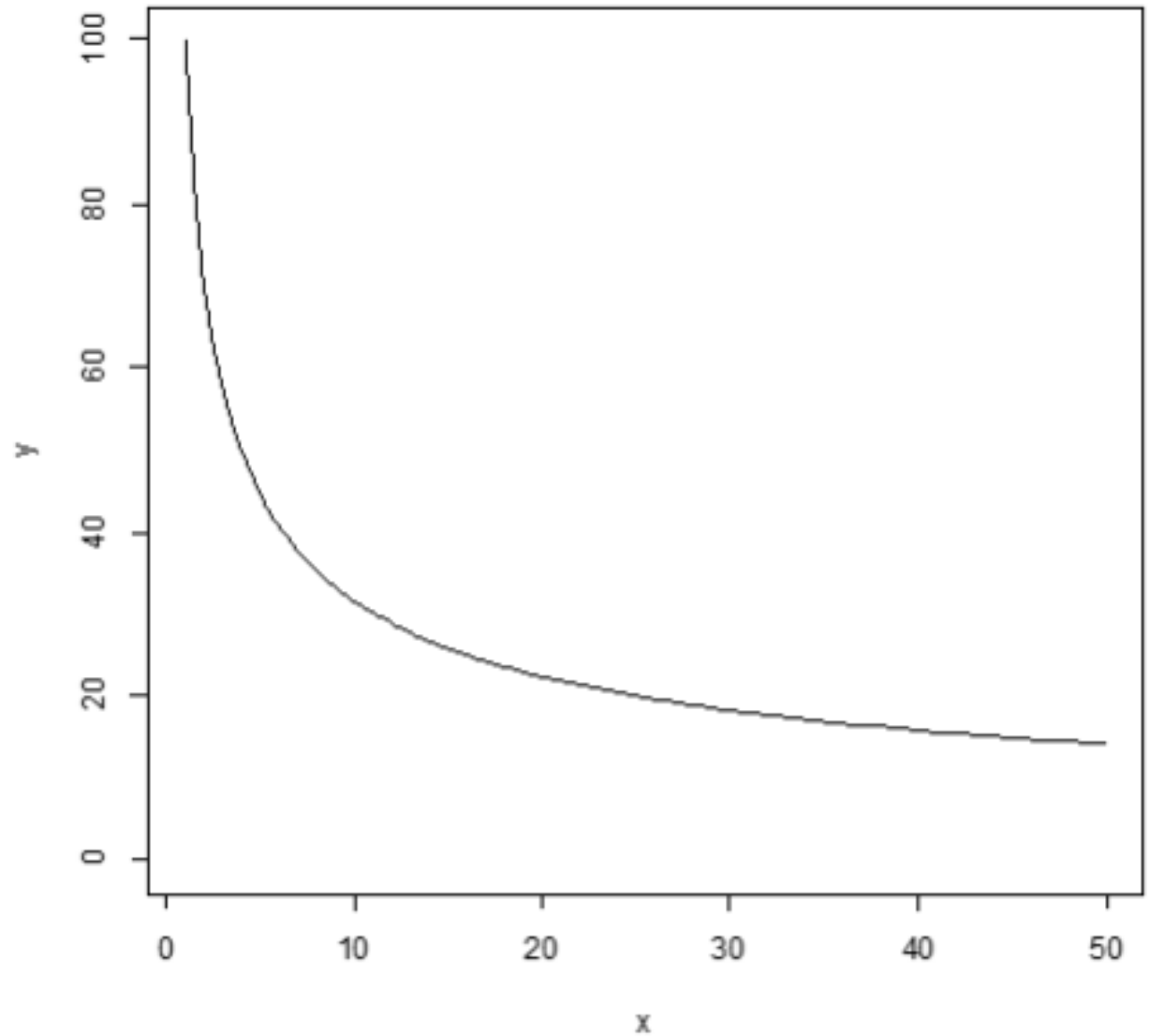


$$y = 100x^{-1}$$

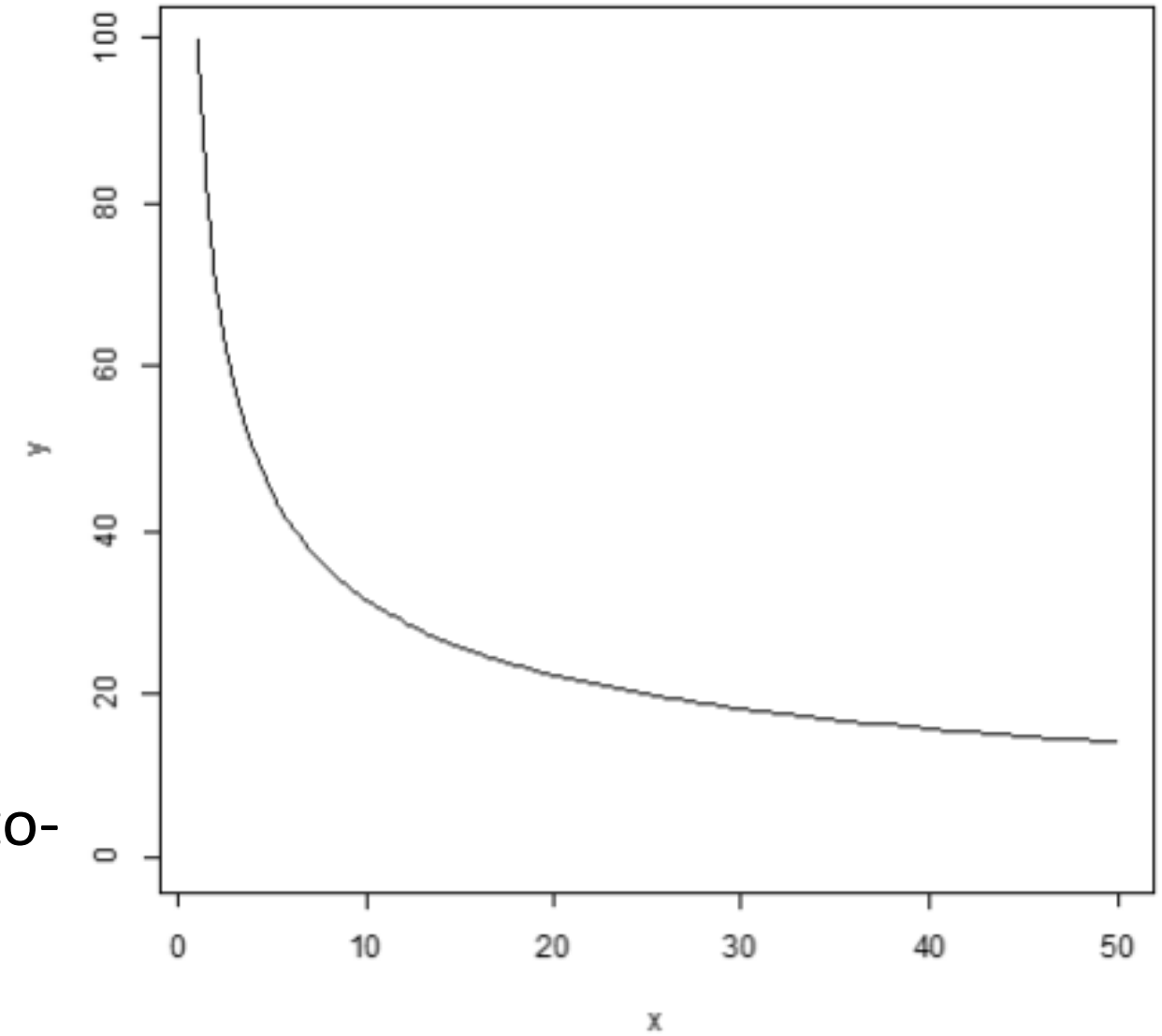


- chudší slovník
  - slova se častěji opakují

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



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



- 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

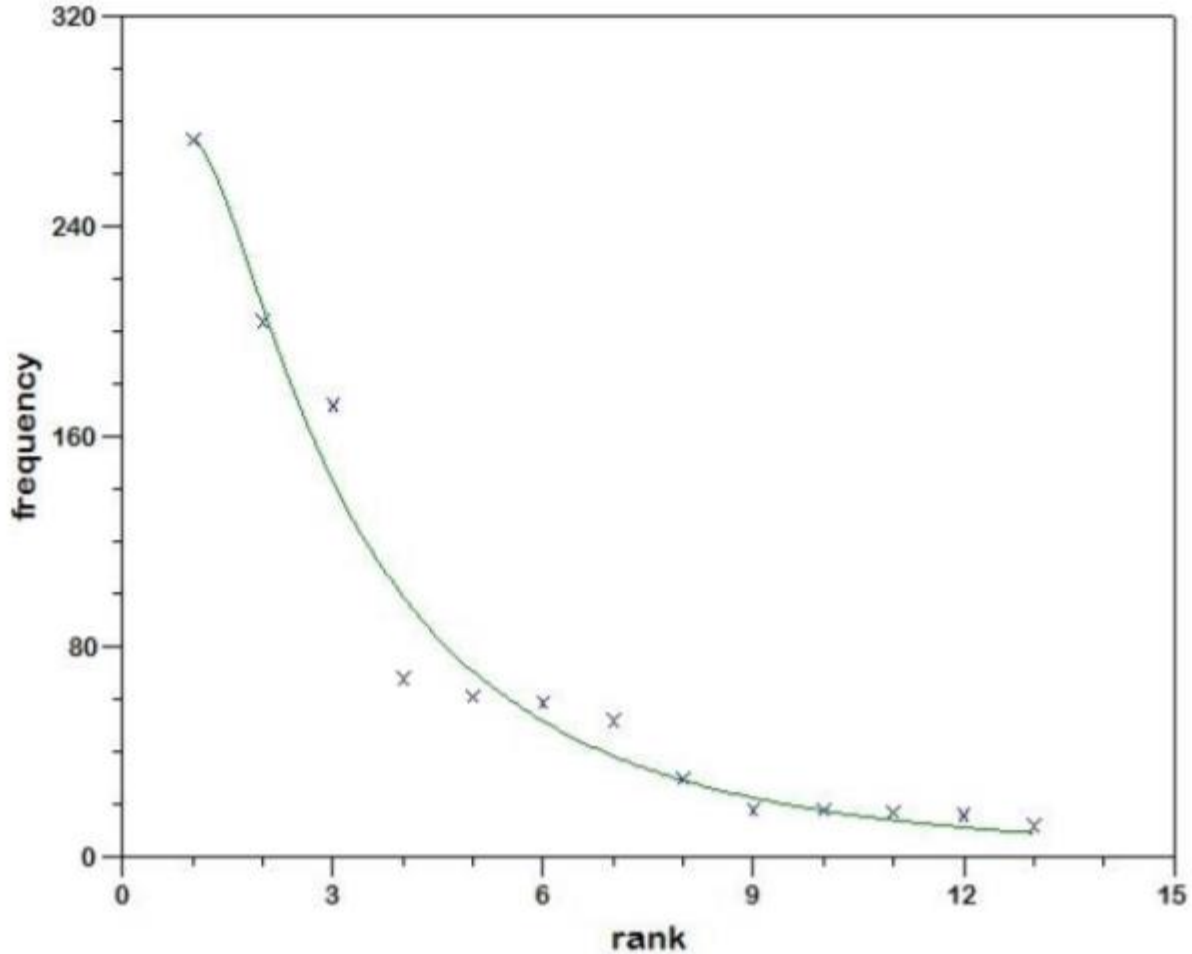


Figure 1. The distribution of all adverbials and the result of the fitting of the Zipf-Alekseev function to the data.

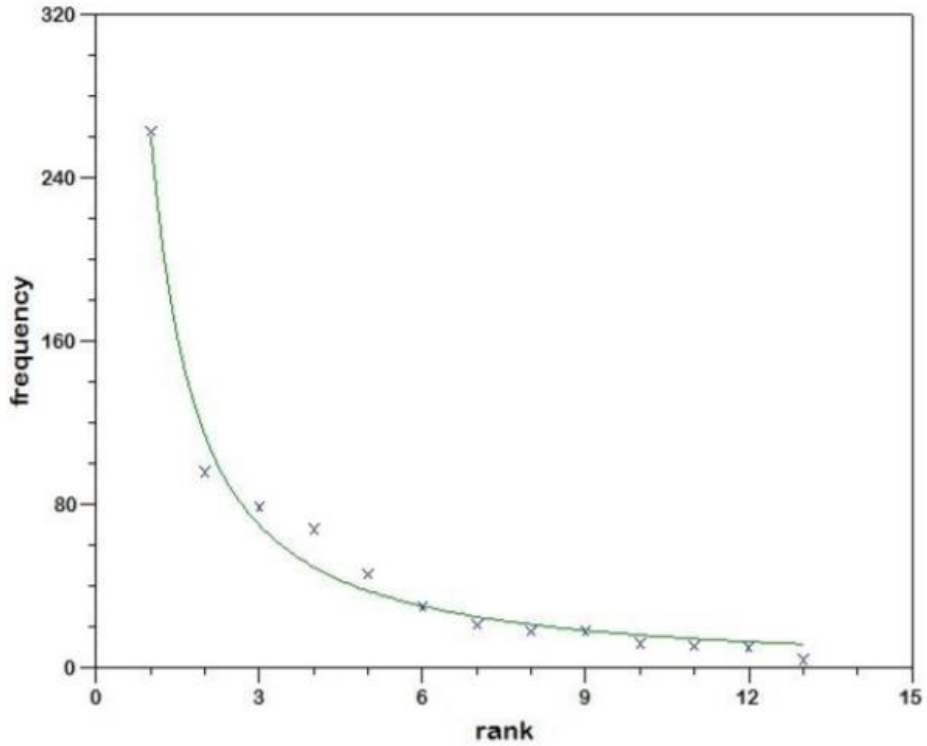


Figure 2. The distribution of adverbials expressed by nouns and the resu
fitting of the Zipf-Alekseev function to the data.

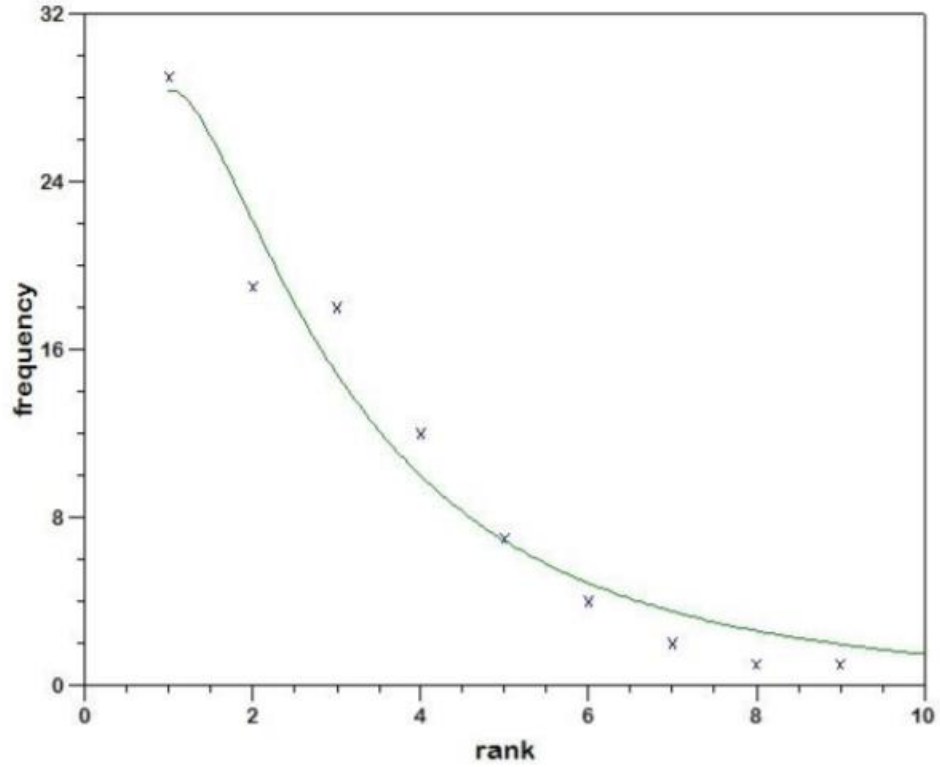


Figure 4. The distribution of adverbials expressed by clauses
and the result of the fitting of the Zipf-Alekseev function to the data.

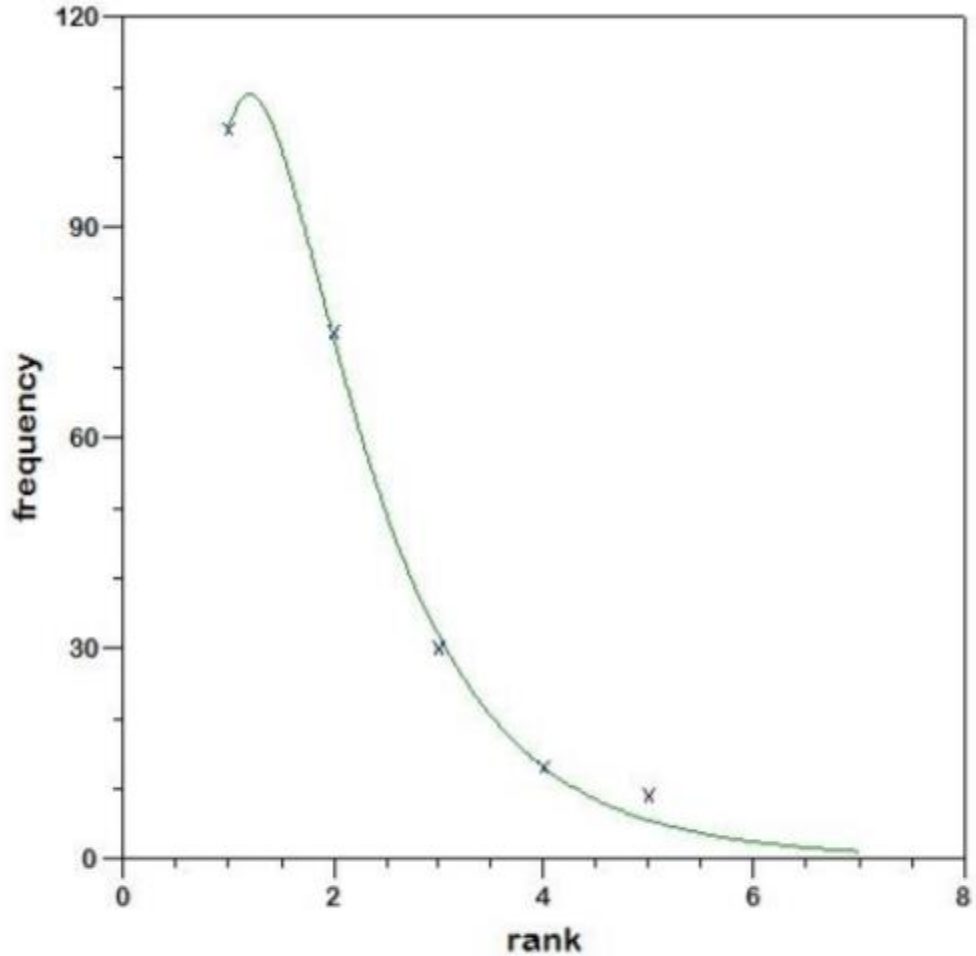
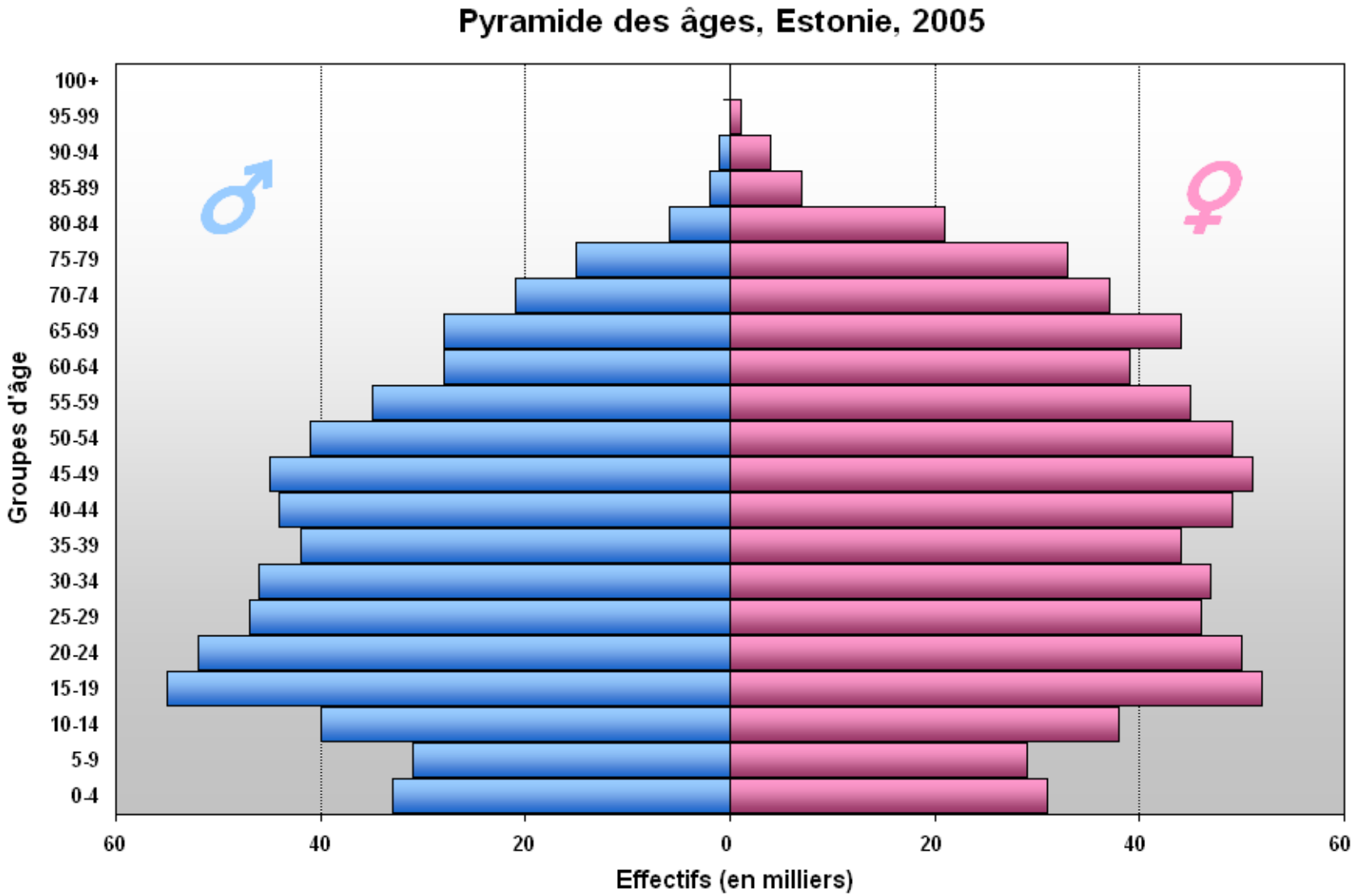
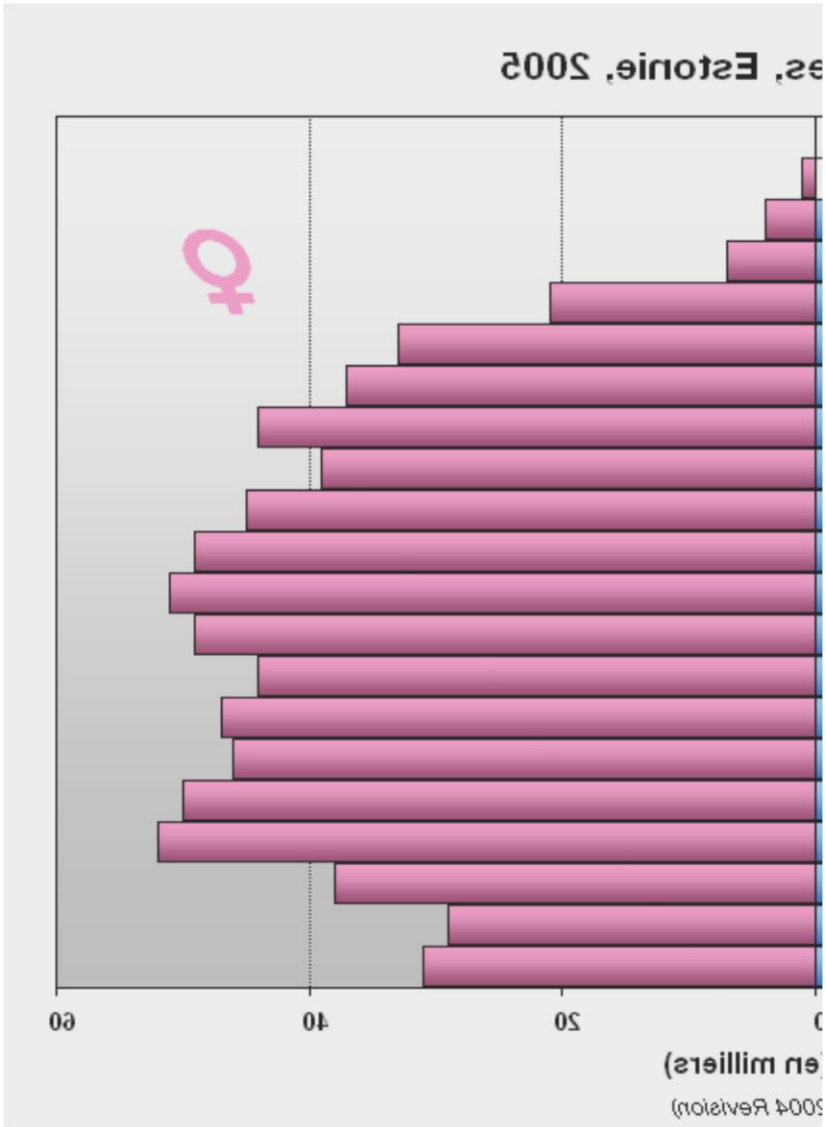
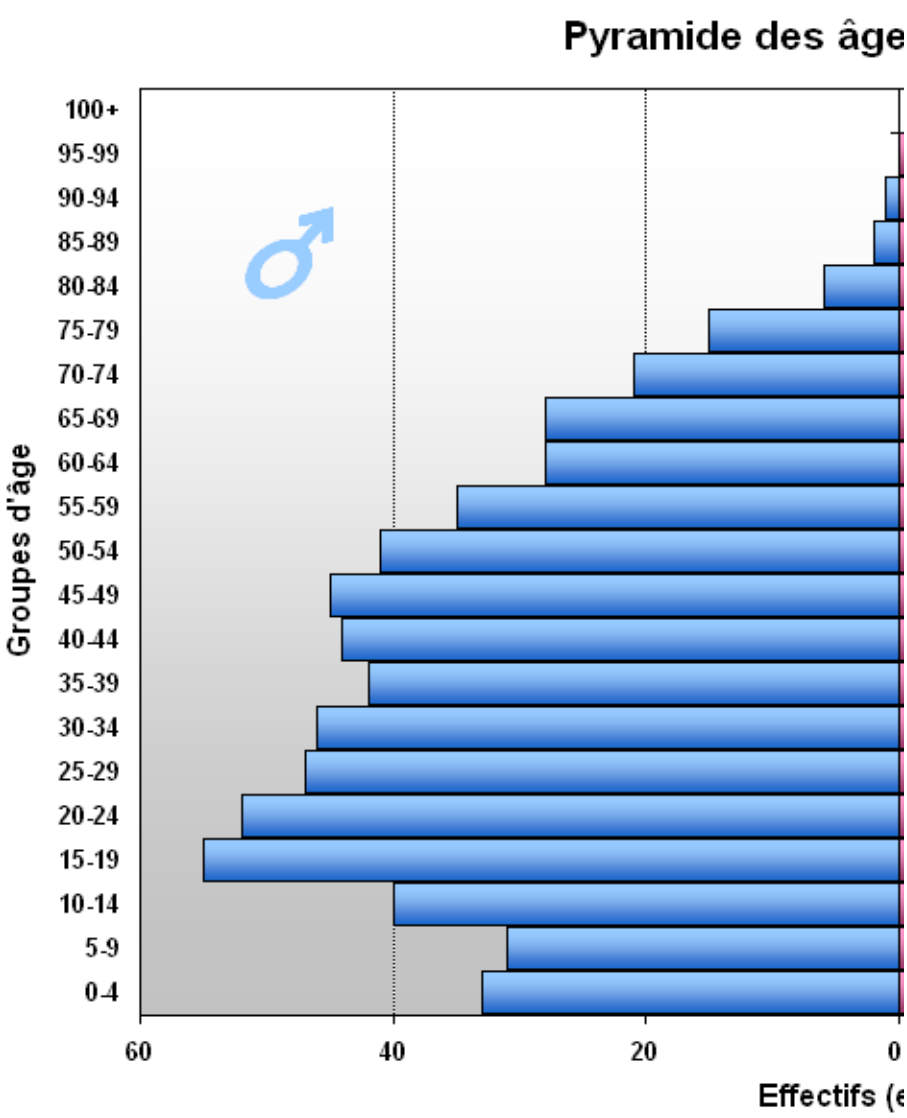


Figure 3. The distribution of adverbials expressed by adverbs and the result of the fitting of the Zipf-Alekseev function to the data.



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

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

Porovnání délek – a jeho interpretace

- očekávání
  - S budou kratší 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

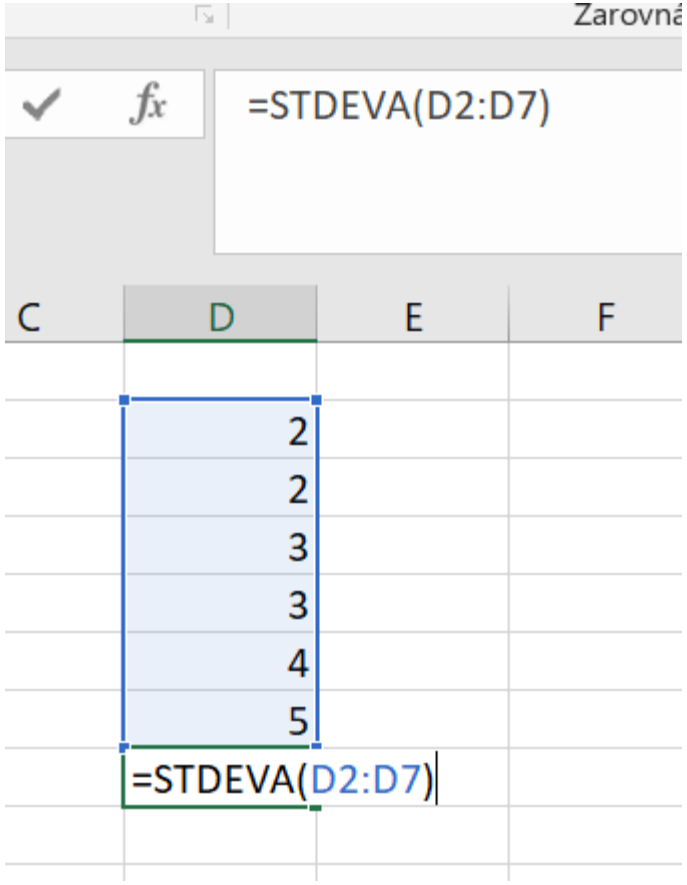
Variabilita dat – směrodatná odchylka

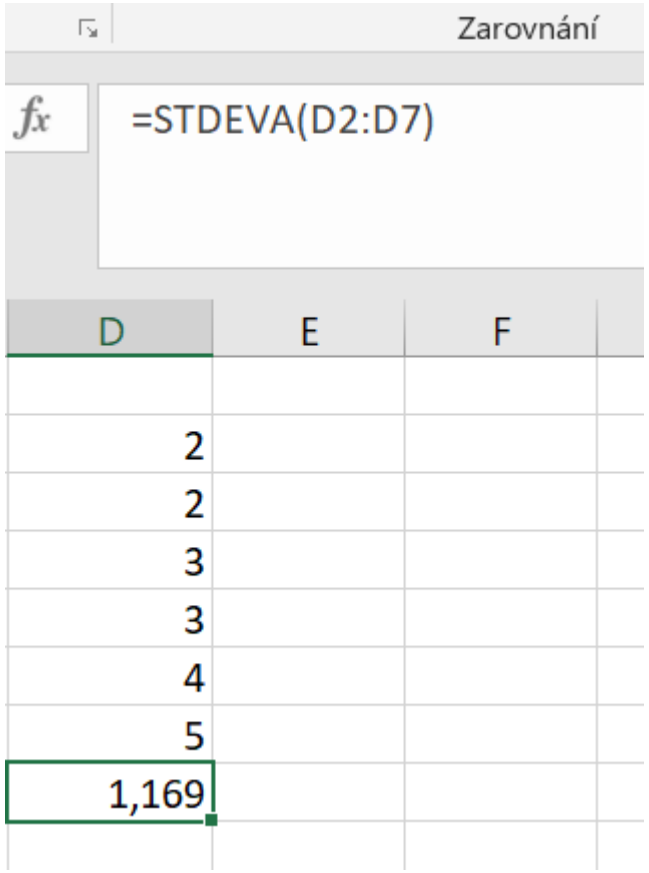
{2,2,3,3,4,5} průměr = 3,17 SD = 1,17

{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



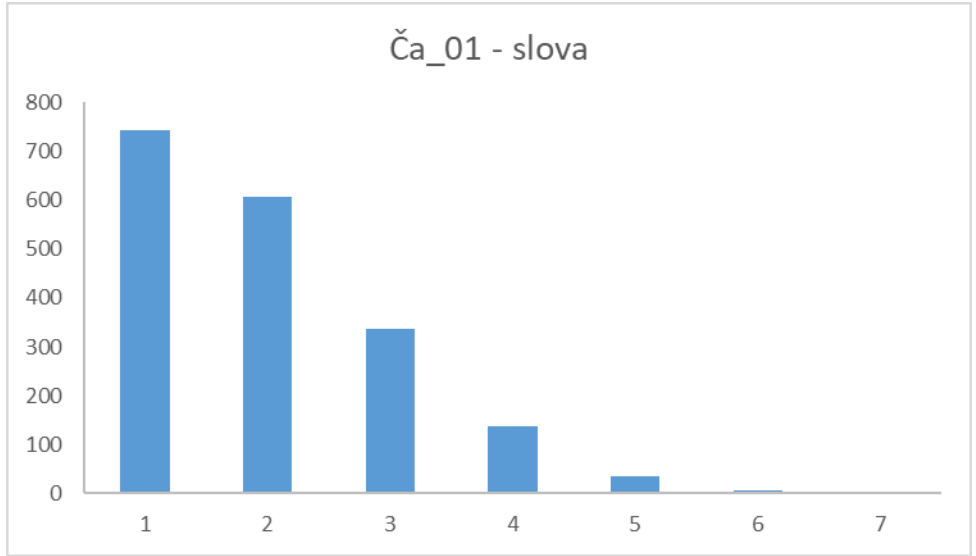


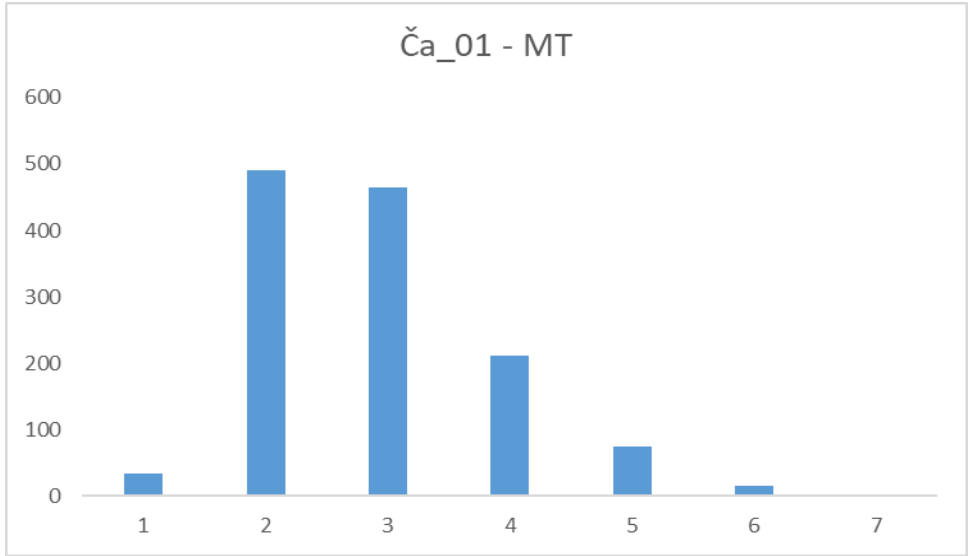
více viz https://support.office.com/cs-cz/article/stdeva-funkce-5ff38888-7ea5-48de-9a6d-11ed73b29e9d

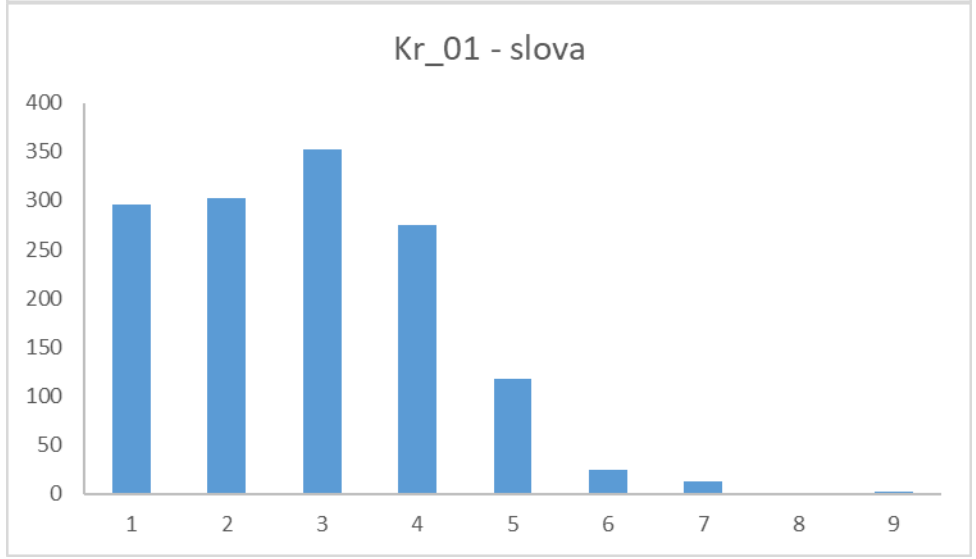
Výsledky – průměrné délky a SD

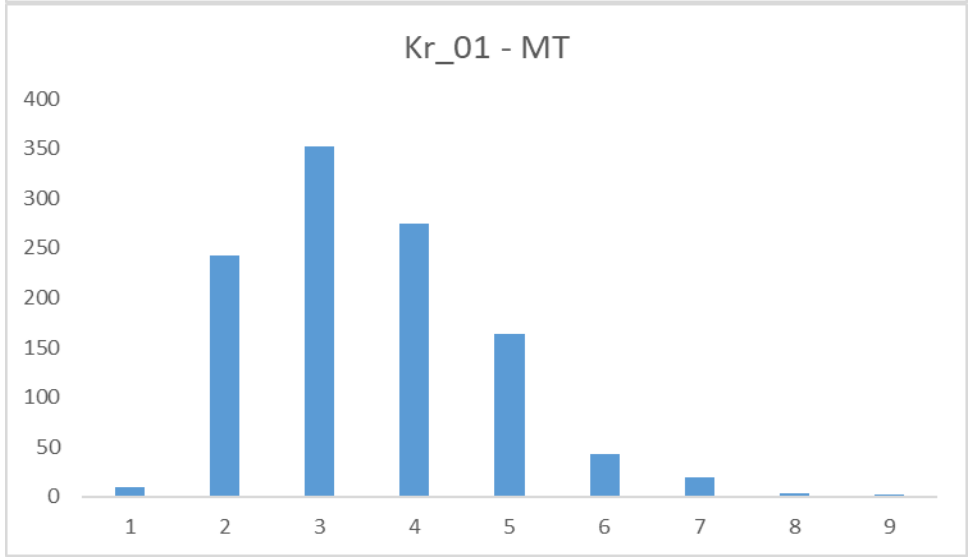
	L_S	SD_S	L_{MT}	SD_{MT}
bel	2	1,05	2,89	1
odb	2,83	1,4	3,51	1,23

Porovnání délek – jeho interpretace & grafické znázornění



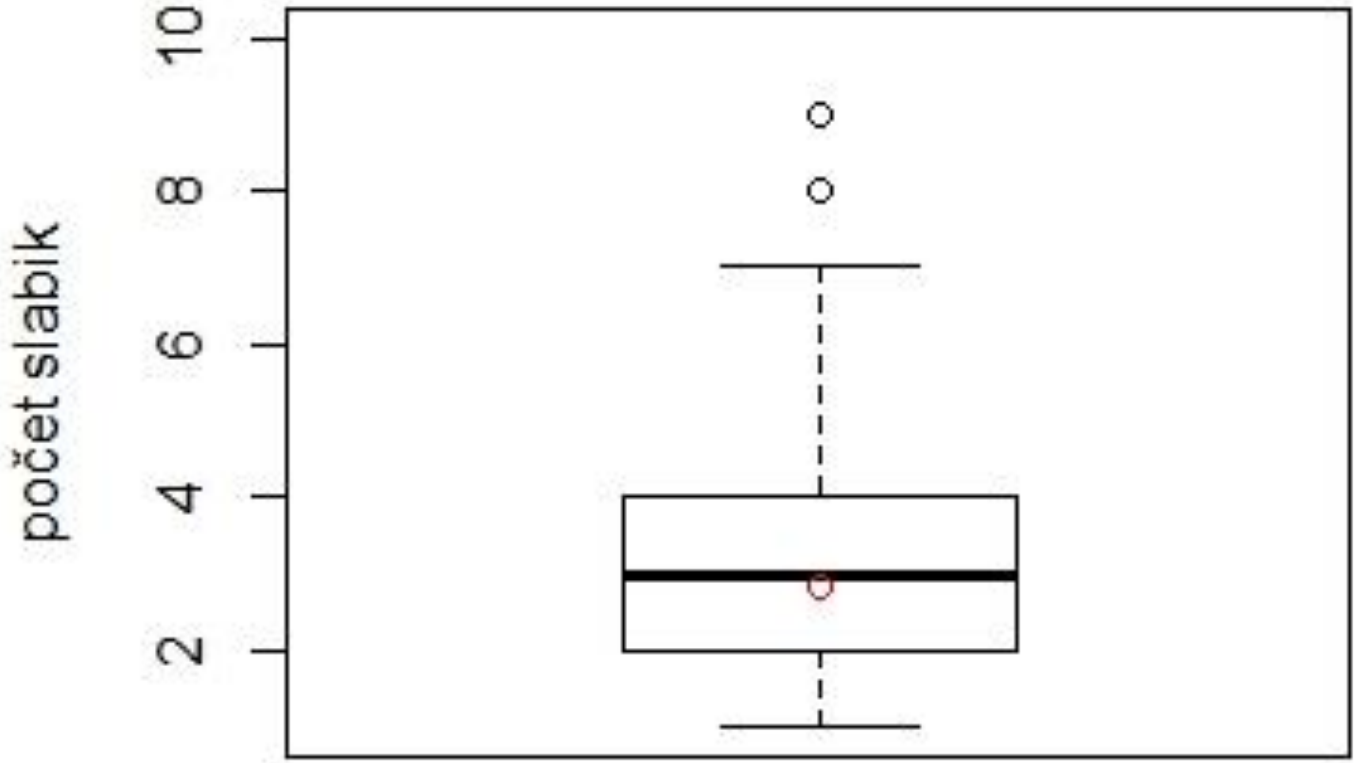




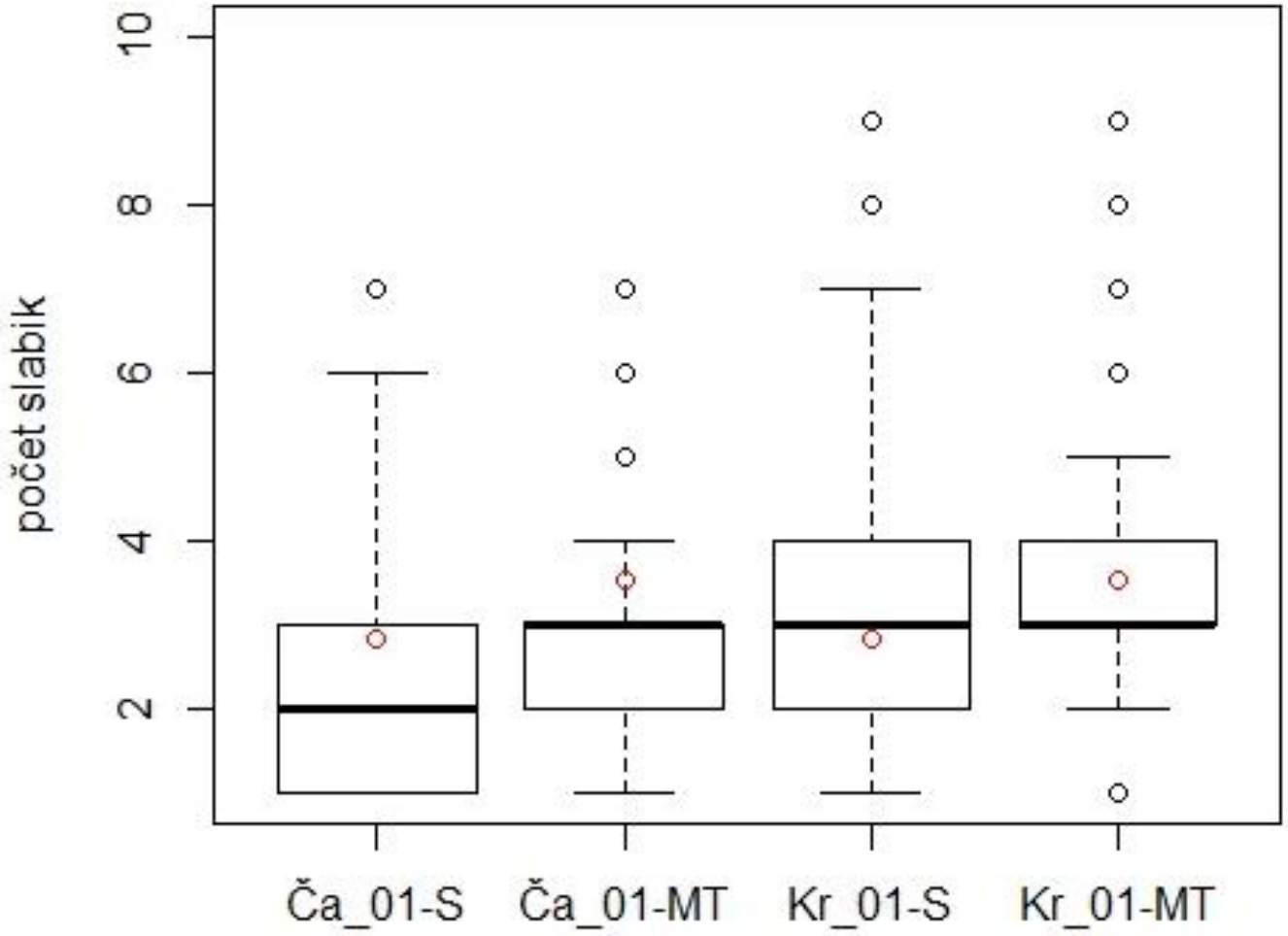


Porovnání délek – jeho interpretace & grafické znázornění

- 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

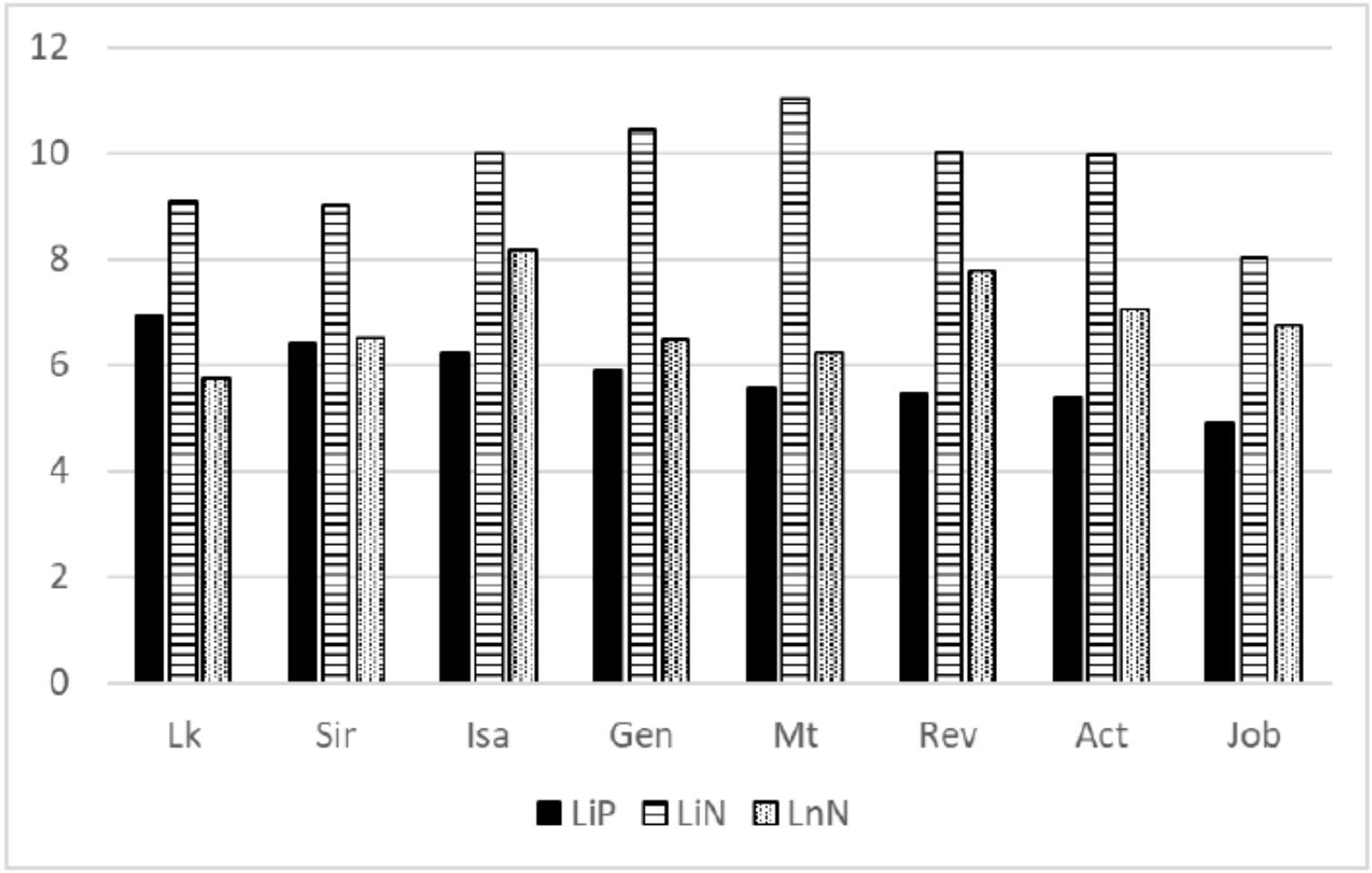
Výsledky – průměrné délky, SD, medián

	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*

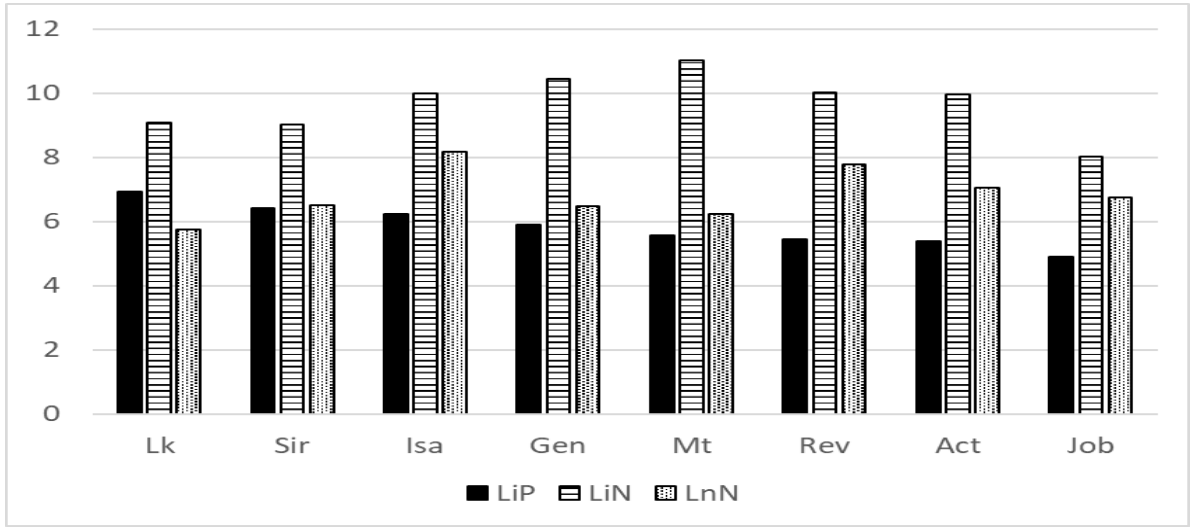


Figure 2 Average length of phrases of *sě* presented in Table 4.

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*

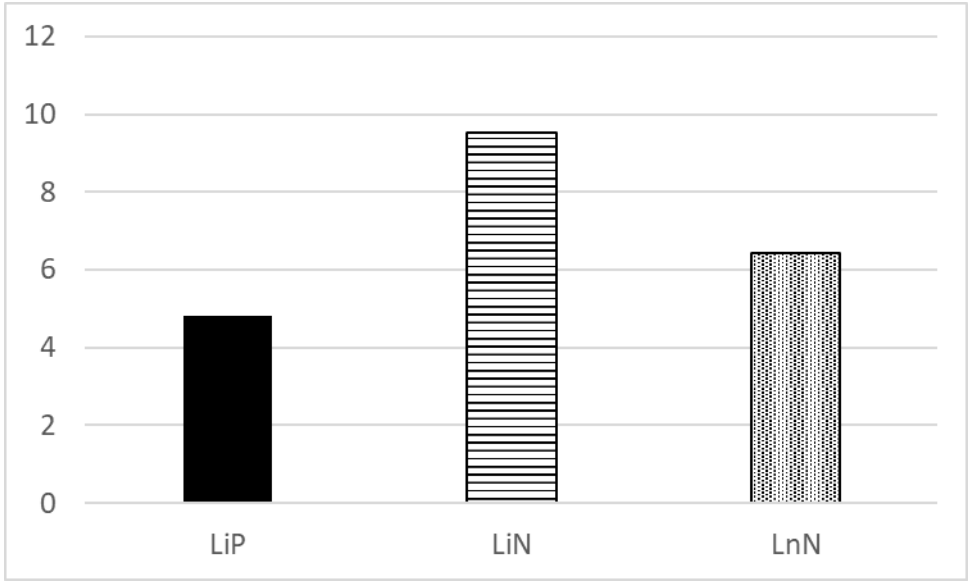
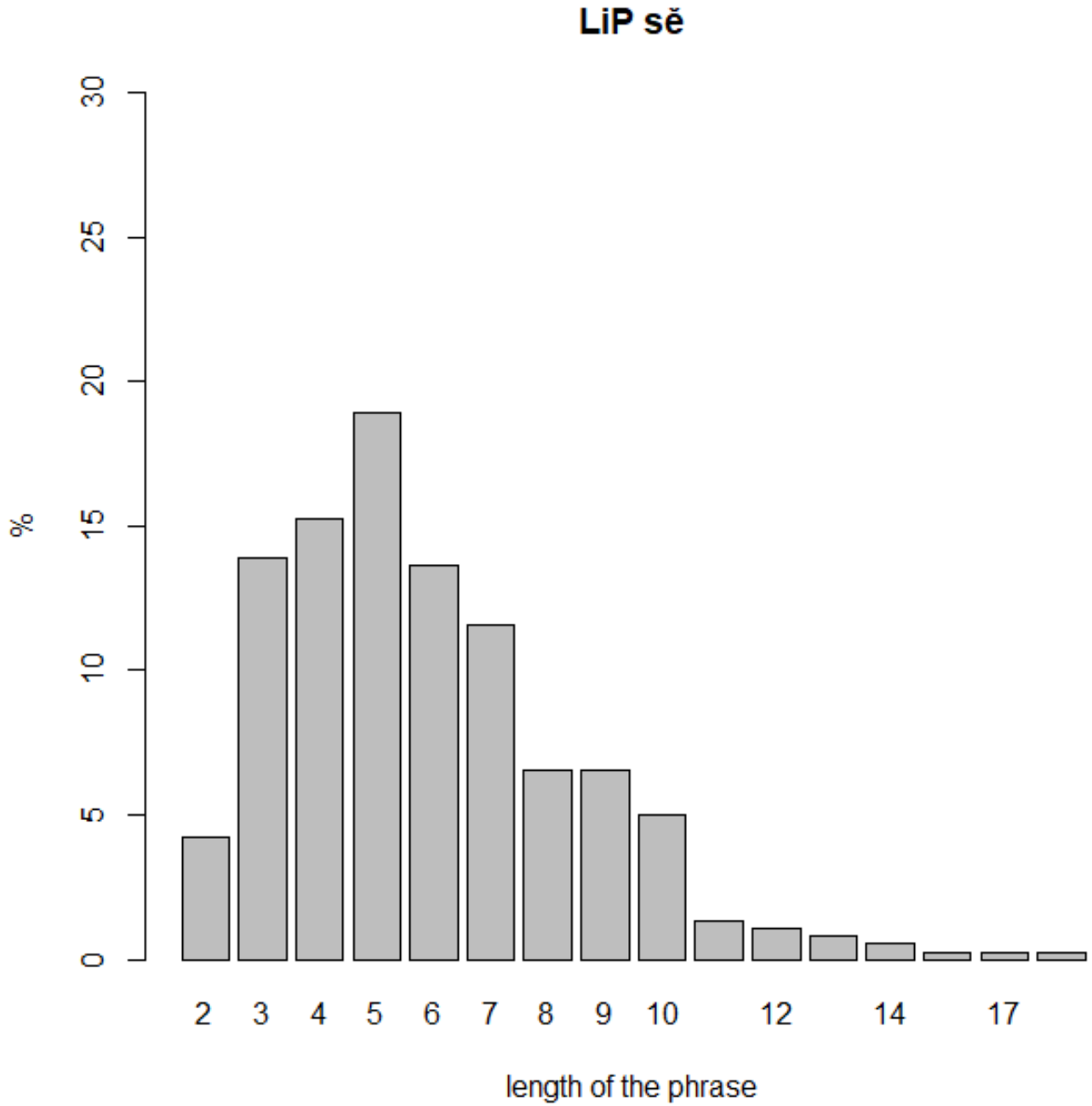
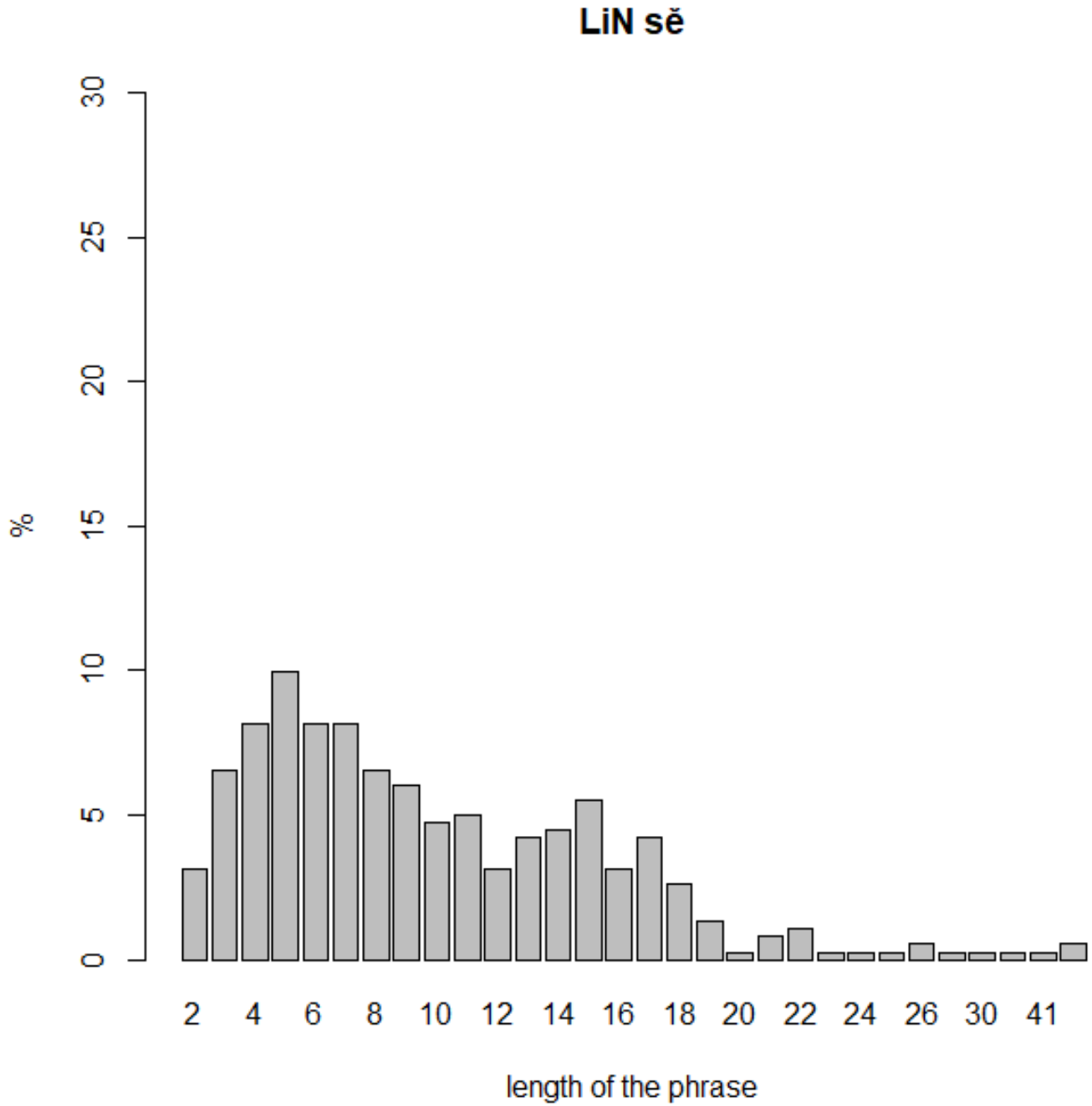
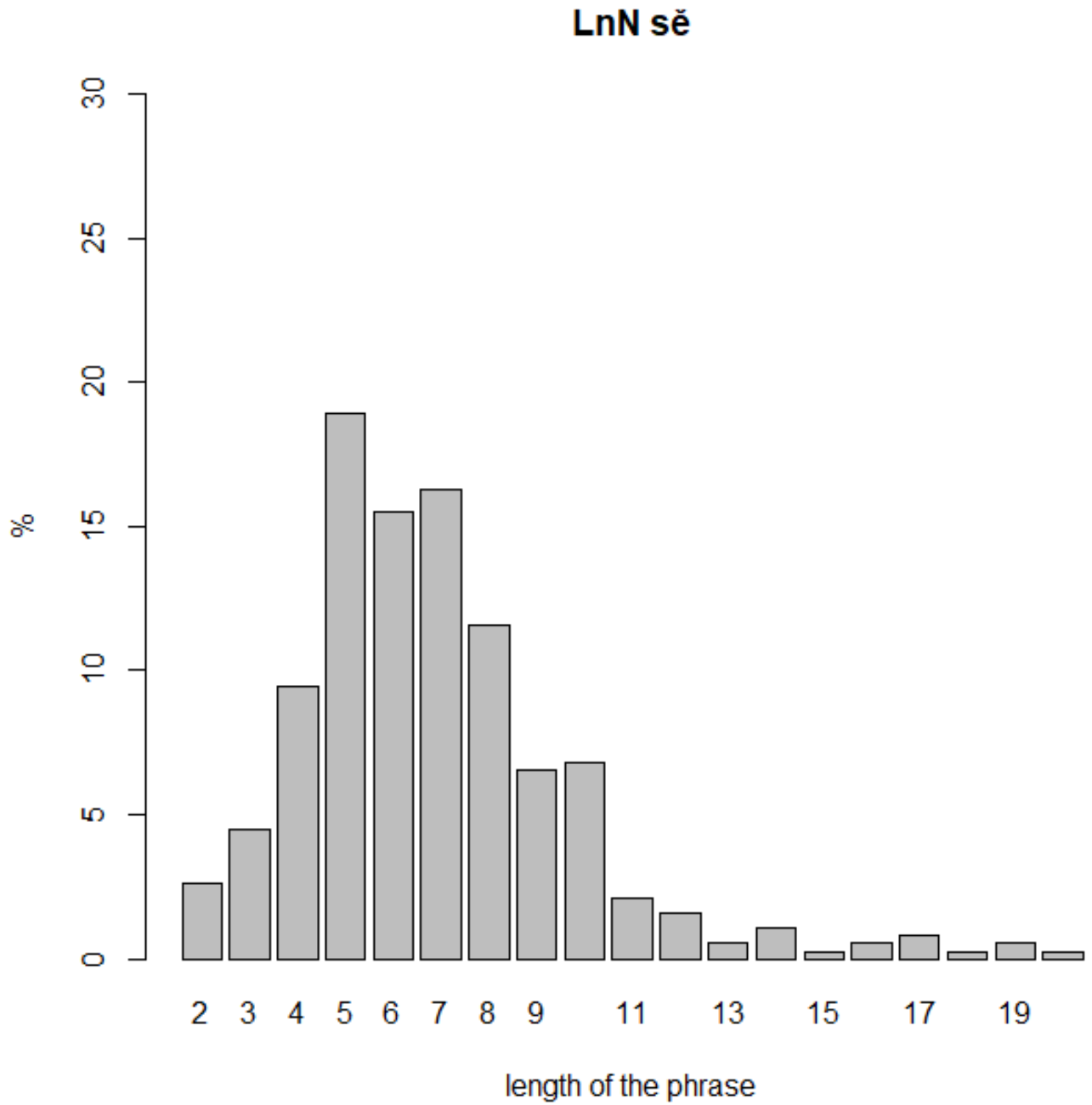
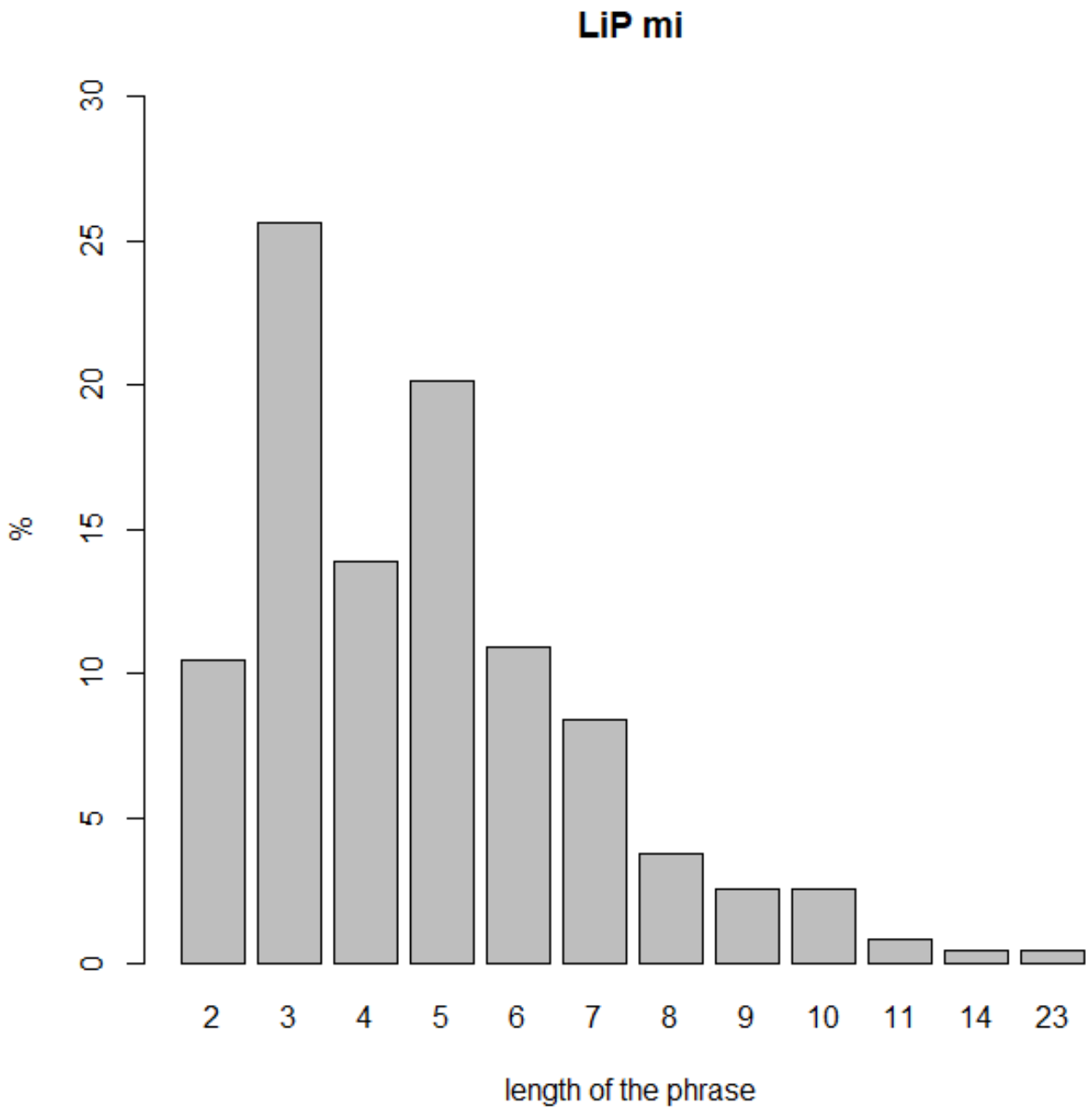


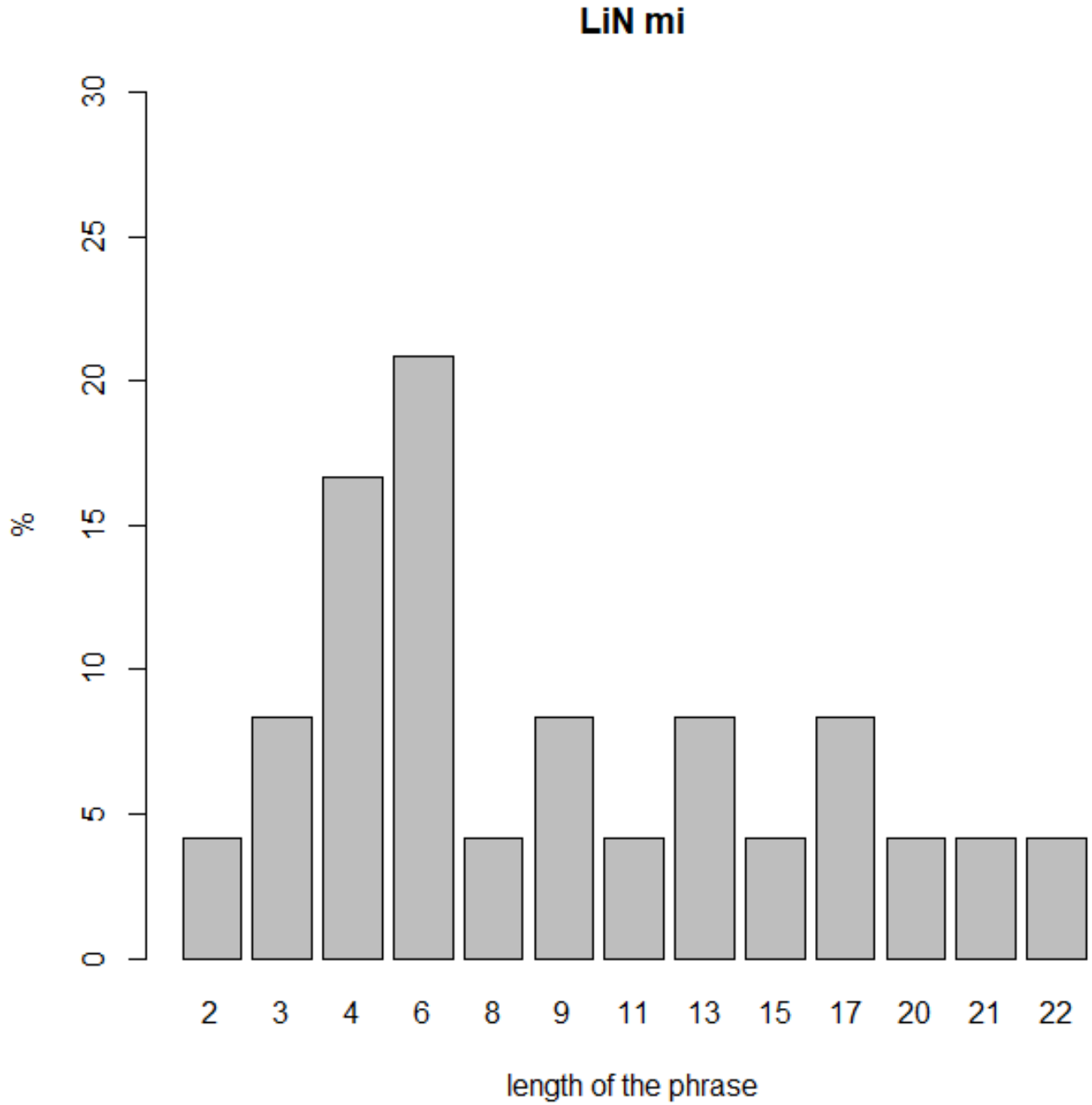
Figure 3 Average length of phrases of *mi* presented in Table 11

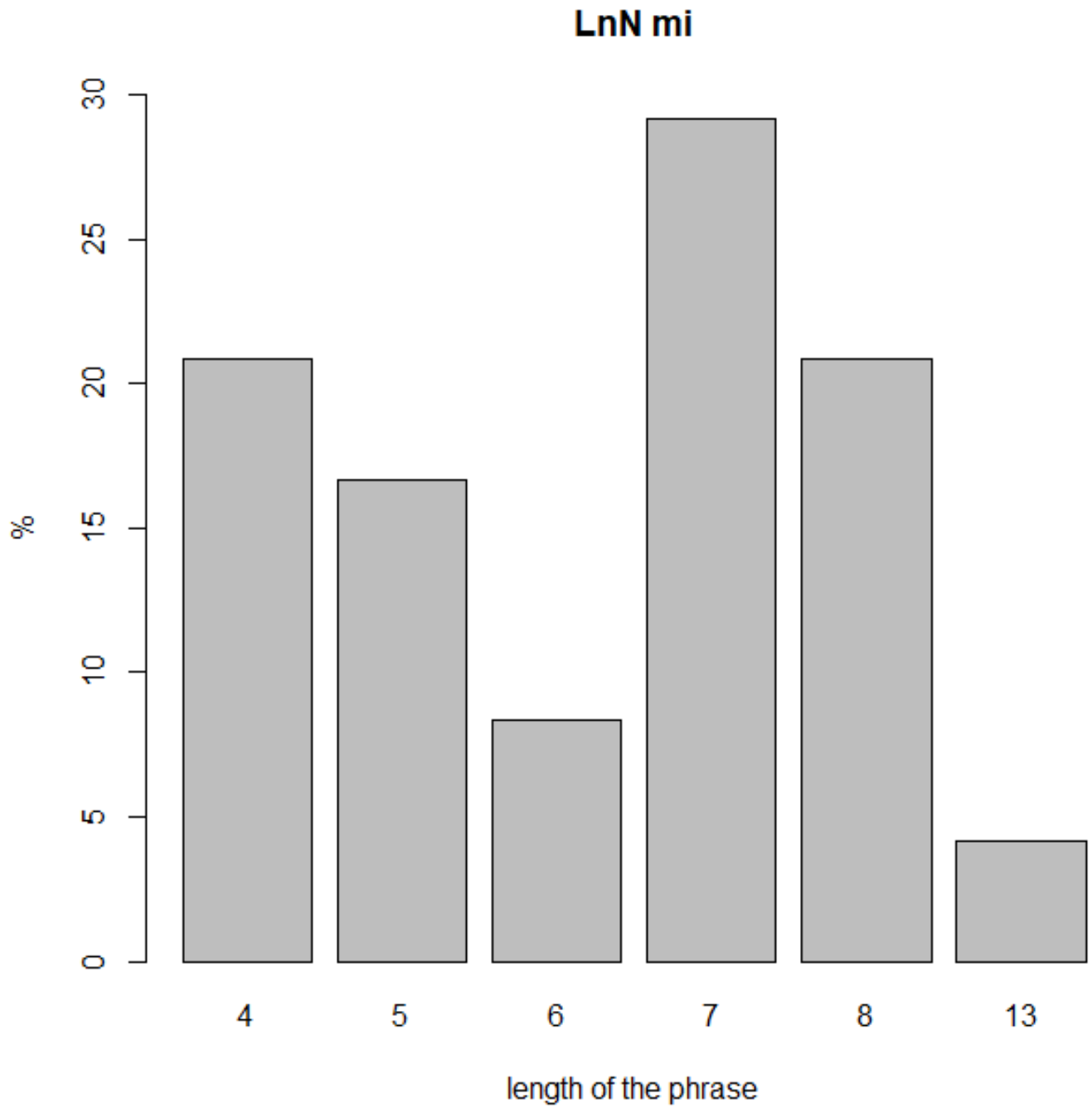

LiP sě
%
0
5
10
15
20
25
30
2
3
4
5
6
7
8
9
10
12
14
17
length of the phrase


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

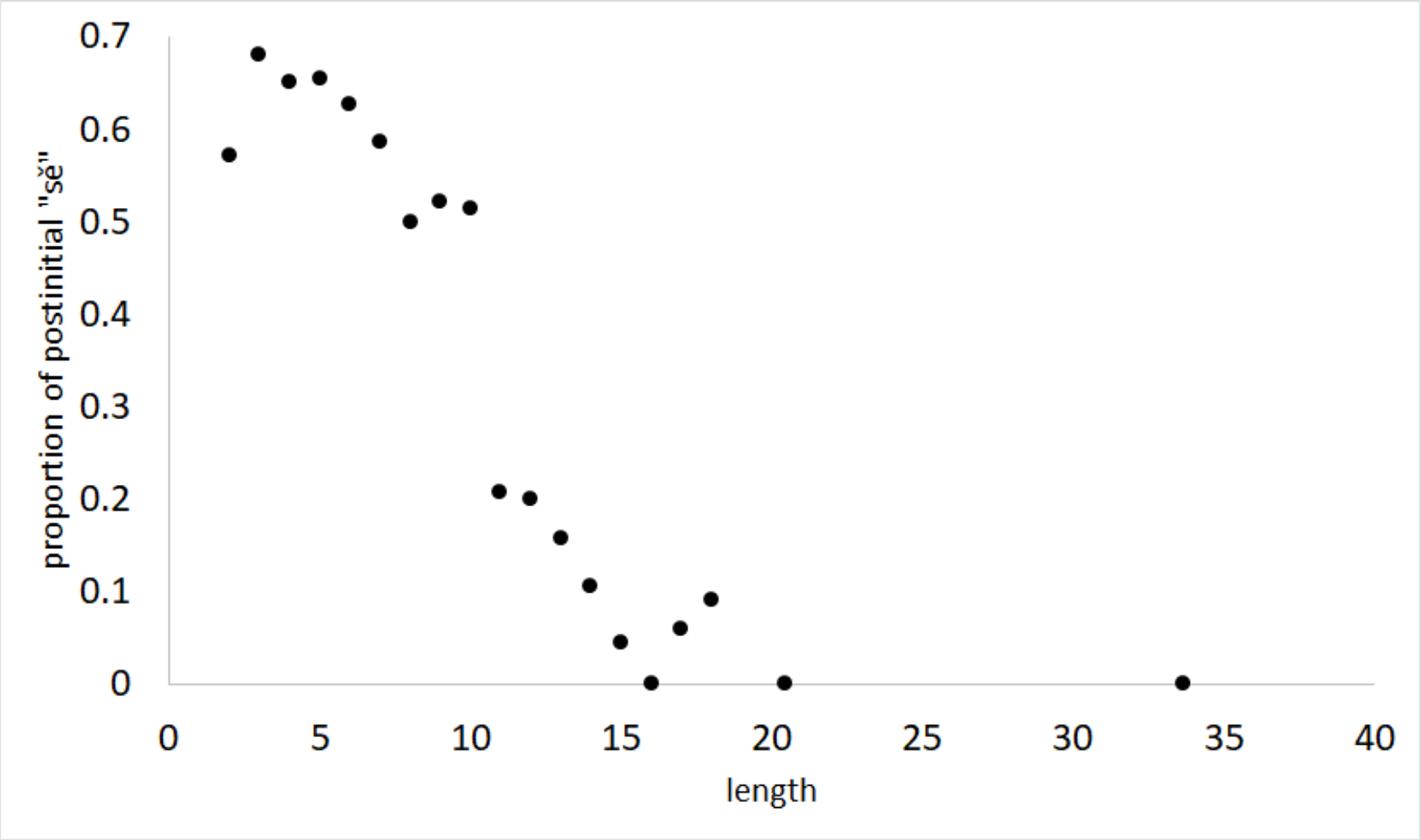
Za hranice popisu…

Za hranice popisu… k testování hypotéz

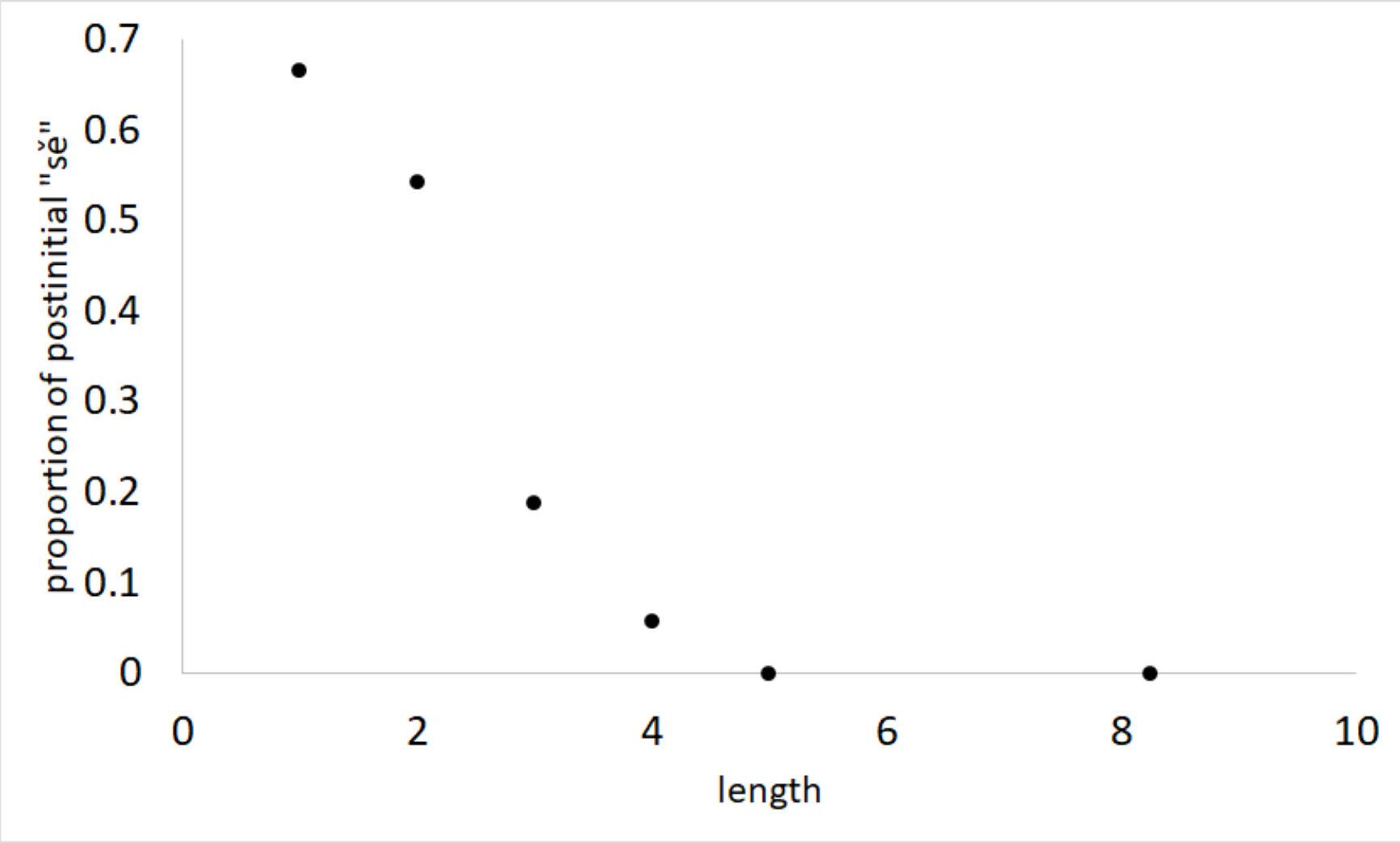
- teoretická zdůvodnění

- 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…