Úvod do kvantitativní lingvistiky

ZS 2022

Opakování 1

- vyhodnoťte vztah mezi frekvencí slova a jeho délkou
  - použijte korelační koeficient
  - zaznamenejte postup

- dají se předpokládat rozdíly mezi typem textu a tímto vztahem?

- Statskingdom

Opakování 2

- vyhodnoťte vztah mezi perfektivitou a mono/ditranzitivitou slovesa
- hypotéza: perfektivní slovesa by se měla častěji realizovat jako ditranzitivní než monotranzitivní

PDT		ditrnas.	monotrans.	% ditrans
doporučit	perf.	31	23	
doporučovat	imperf.	18	38	
poskytnout	perf.	28	23	
poskytovat	imperf.	21	37	

- https://www.socscistatistics.com/tests/chisquare/

Kvantitativní analýza textů

- proč má smysl analyzovat texty?

- jak má smysl analyzovat texy?
  - po textech jako jednotkách?
  - agregace textů → korpusy?

Kvantitativní analýza textů

- proč má smysl analyzovat texty?

- jak má smysl analyzovat texy?
  - po textech jako jednotkách?
  - agregace textů → korpusy?

- záleží na cíli analýzy!

Kvantitativní analýza textů

- porovnávání jednotlivých textů / skupin textů
- analýza jazykových zákonů / mechanismů

Porovnávání jednotlivých textů / skupin textů

- individuální texty
- autorství
- žánrová analýza
- stylistika / stylometrie

Analýza jazykových zákonů / mechanismů

- Law of Brevity
- Menzerath-Altmann Law
- Heaps Law

Text vs. korpus

- 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.

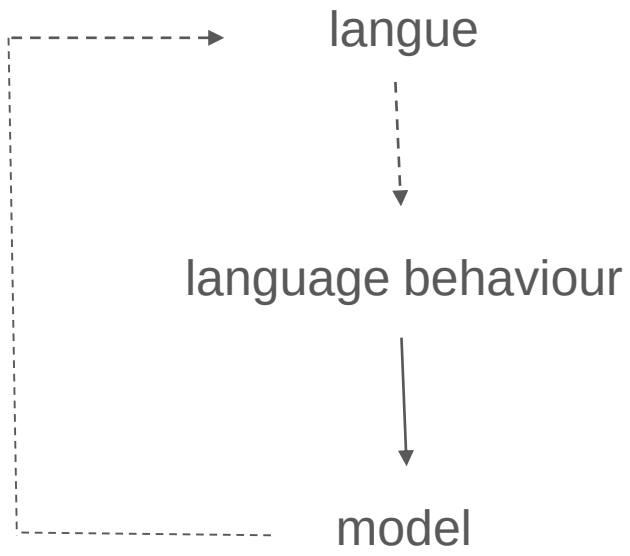
What do we model?

- langue
- competence

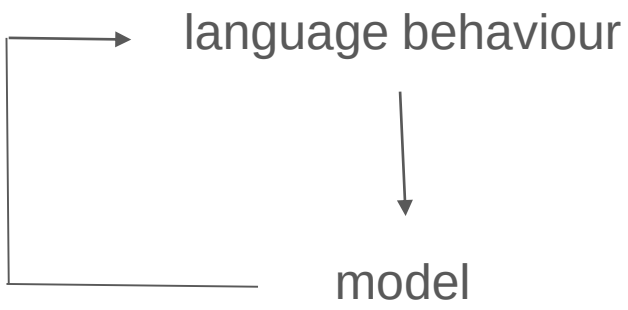
- language behaviour

What do we model?

- L-LB-M-L approach

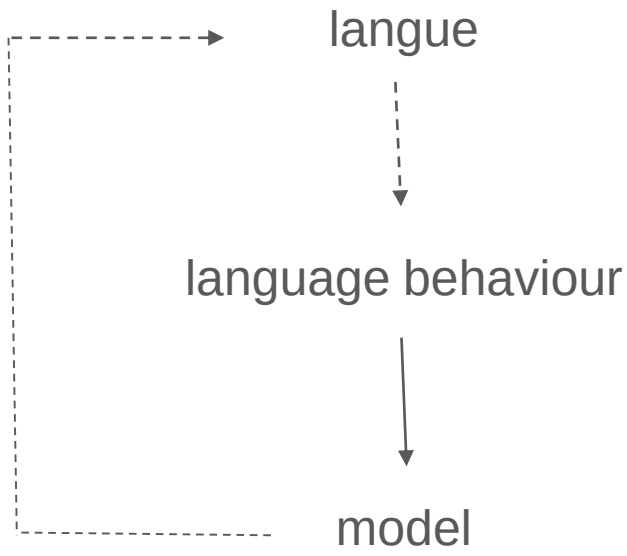


- LB-M approach

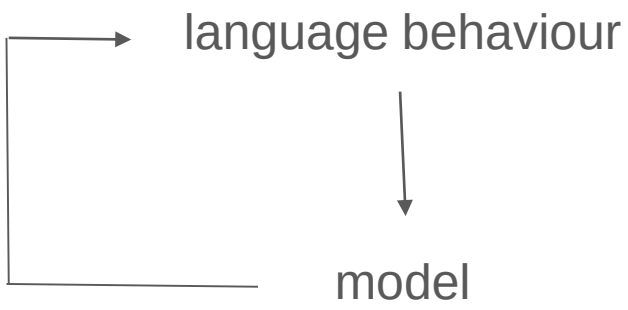


What do we model?

- L-LB-M-L approach



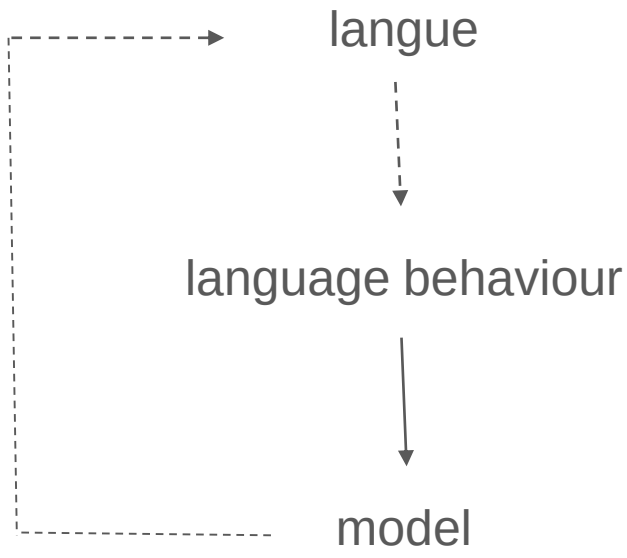
- LB-M approach



CORPUS

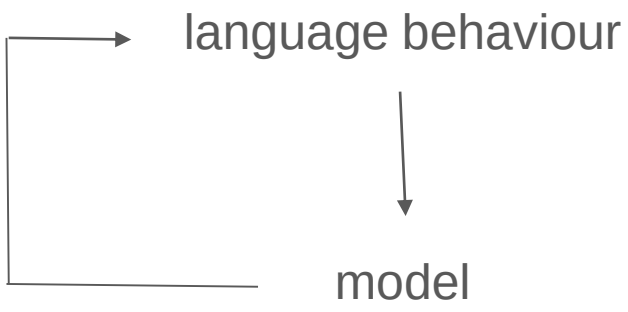
What do we model?

- L-LB-M-L approach



CORPUS

- LB-M approach



CORPUS or/and TEXT

Language behavior → model

- inductive approach
  - regularities, correlations etc.
  - generalisations
  - description
    - corpus driven

Language behavior → model

- inductive approach
  - regularities, correlations etc.
  - generalisations
  - description
    - corpus driven

- deductive approach
  - theoretical assumptions
  - theory
    - Zipf (least effort principle)
    - synergetic linguistics
  - model of mechanism
    - language behaviour

Language behavior → model

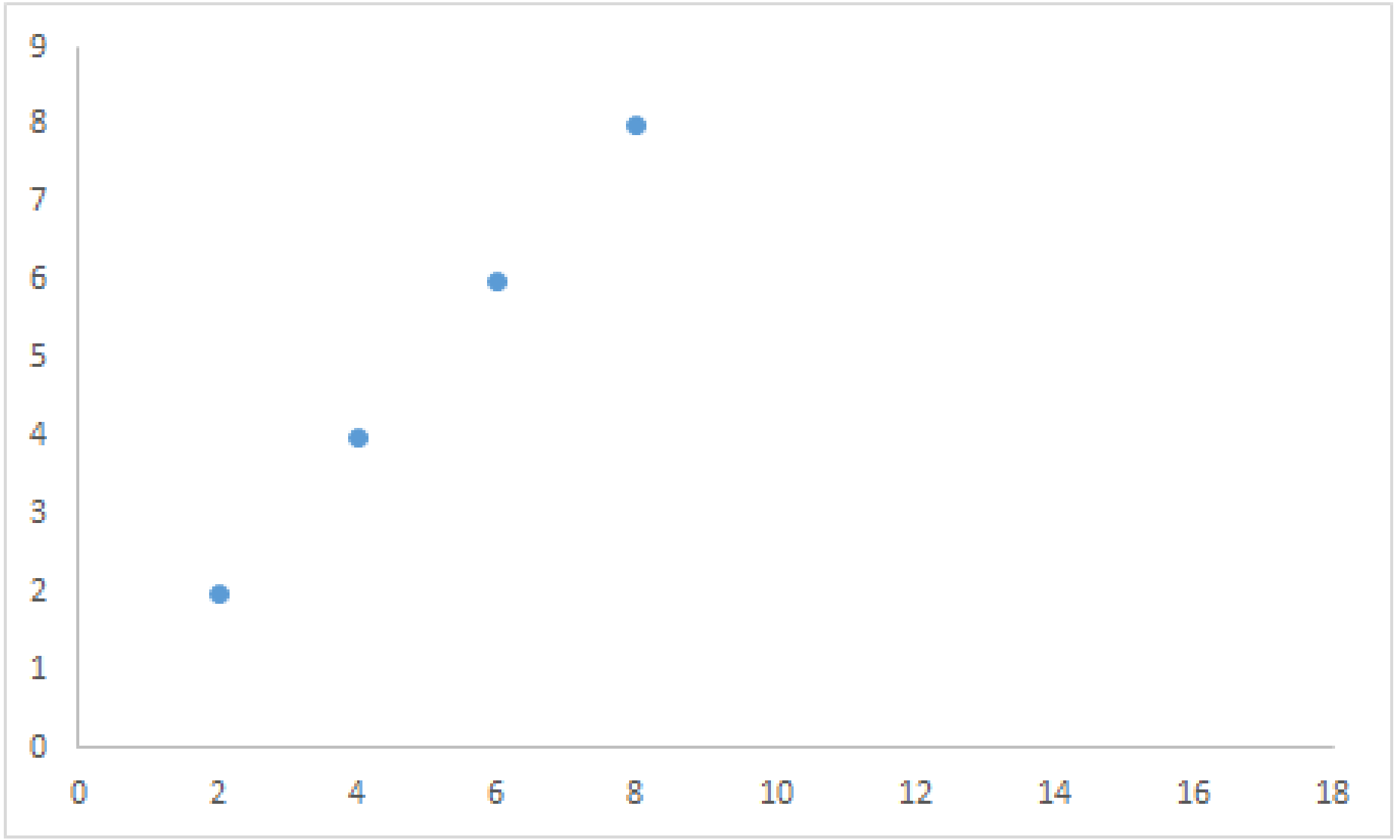
- inductive approach
  - regularities, correlations etc.
  - generalisations
  - description
    - corpus driven

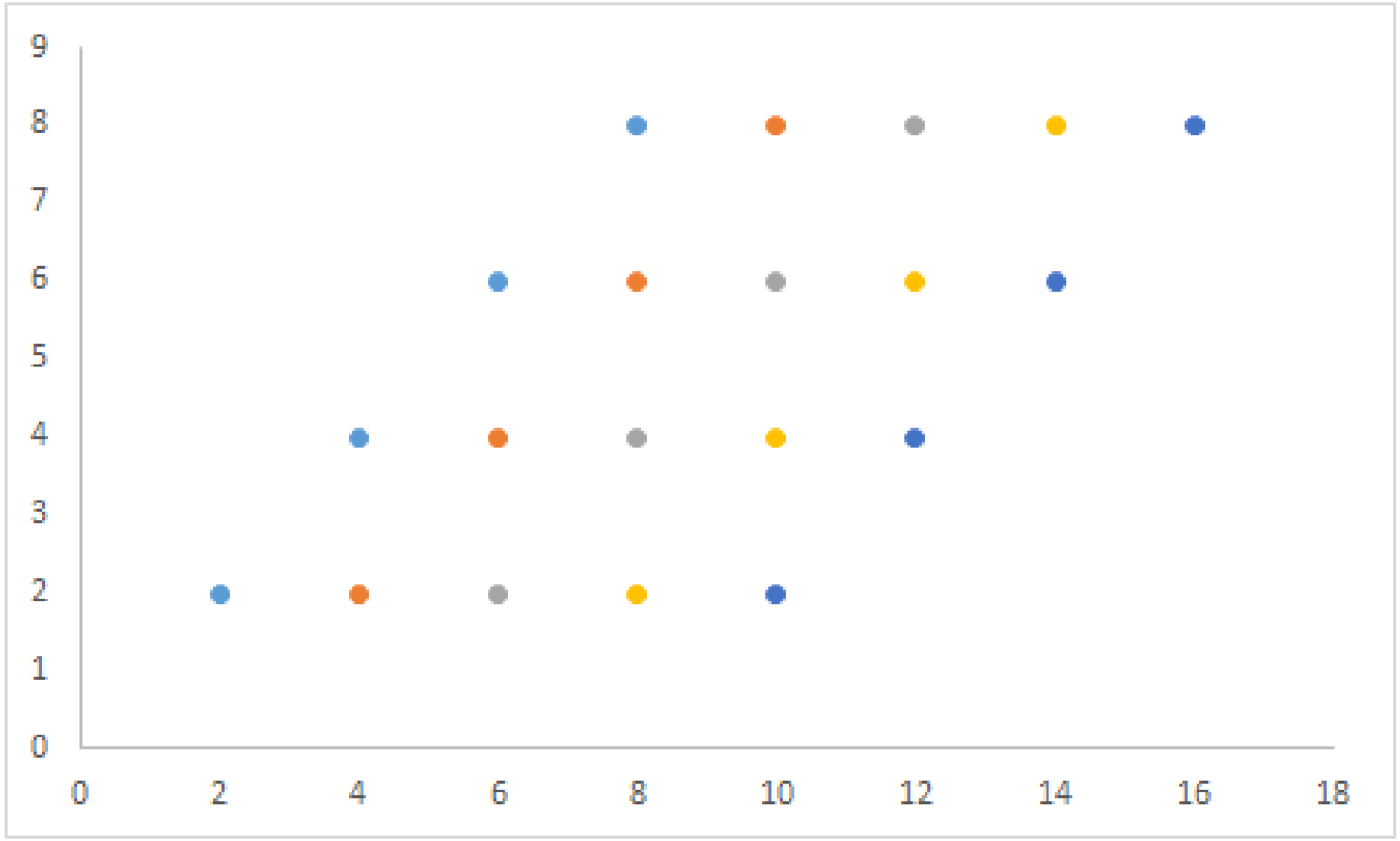
- corpus

- deductive approach
  - theoretical assumptions
  - theory
    - Zipf (least effort principle)
    - synergetic linguistics
  - model of mechanism
    - language behaviour

- text (corpus)

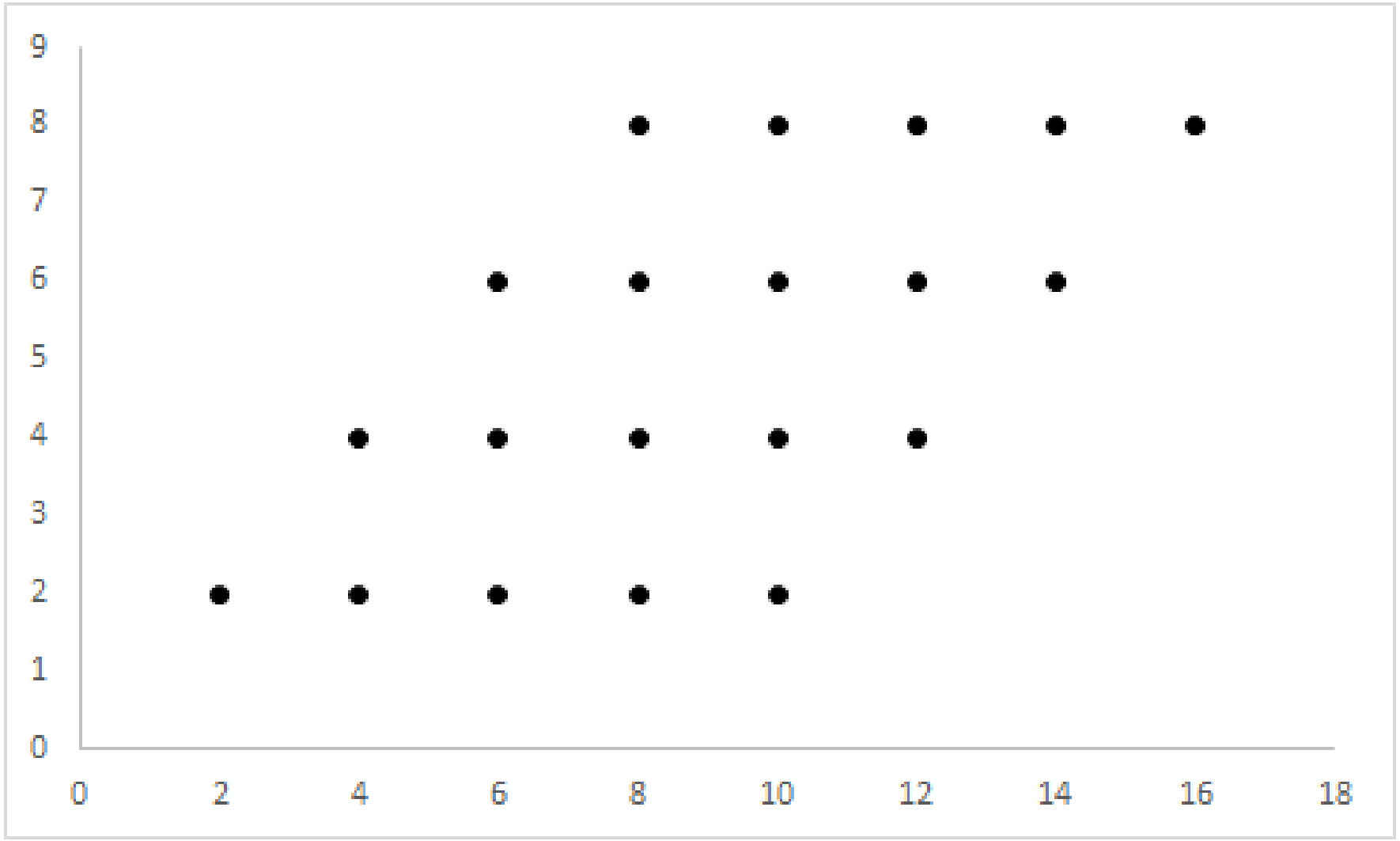
a) Ak chceme overovať nejaký zákon, tak nesmieme miešať texty, lebo v každom texte sú iné tzv. počiatočné podmienky. Dokonca je niekedy potrebné analyzovať oddelene aj jednotlivé kapitoly románu alebo vety symfónie. Dá sa napríklad ukázať, že zákon pre rozdelenie dĺžky slova je ľahko overiteľný napr. na jednotlivých Goetheho listoch. Čím viac listov však zlúčime do spoločného výberu, tým menej vhodným sa stáva daný model, lebo v každom liste má zákon iné parametre. V takých prípadoch sa odporúča prinajmenšom miešanie, kombinovanie modelov (porov. Altmann 1992a), ale ešte vhodnejšia je separácia homogénnych častí. Tak napríklad tzv. frekvenčný slovník jedného celého jazyka (žánru a pod.) je vhodný len na veľmi obmedzené teoretické účely, lebo je konštruovaný zo zmesi textov.


9
8
7
6
5
4
3
2
1
0
0
2
4
6
8
10
12
14
16
18


9
8
7
6
5
4
3
2
1
0
0
2
4
6
8
10
12
14
16
18

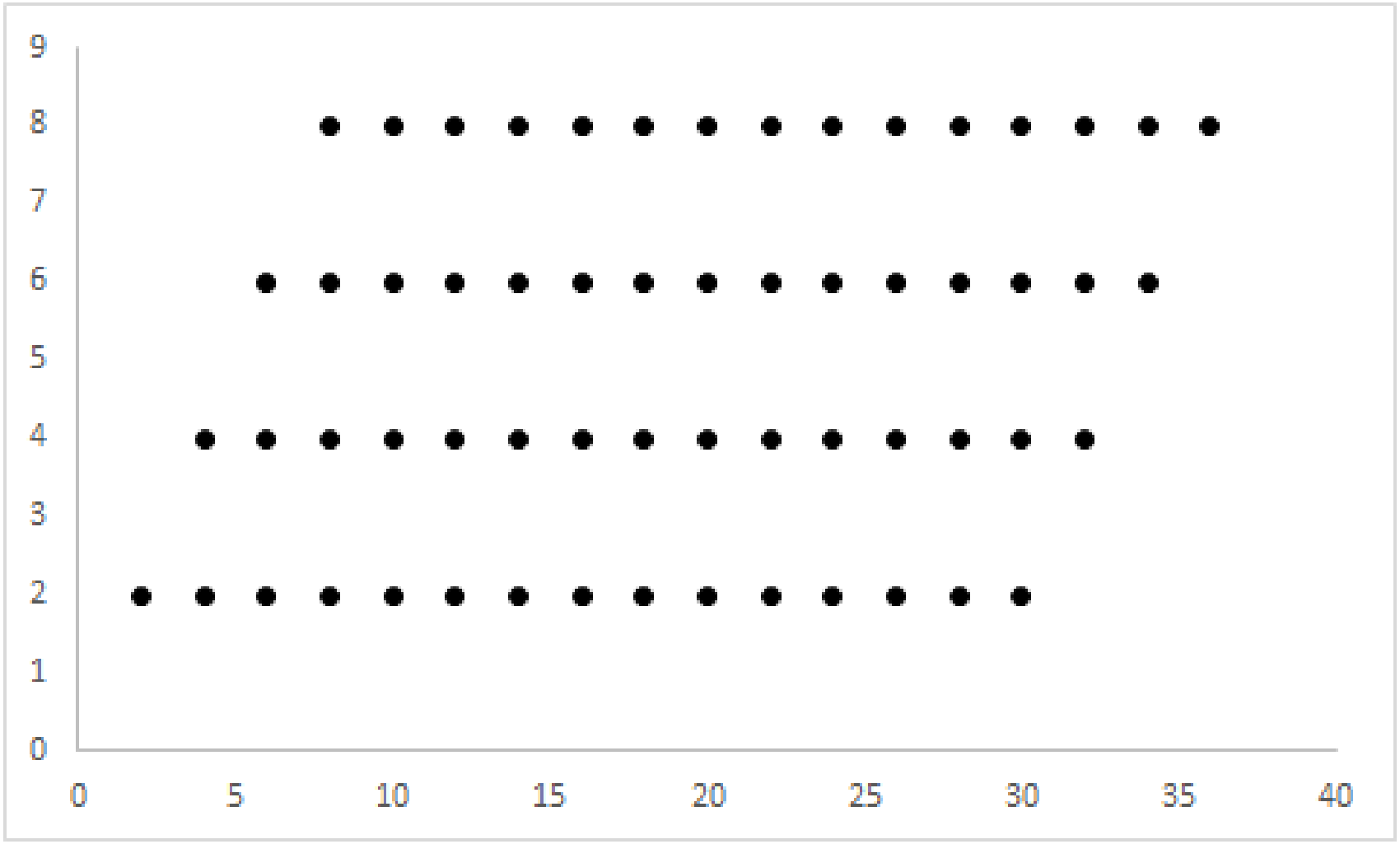
$\tau = 0.50$

p-value = 0.006



$\tau = 0.19$

p-value = 0.06



Case study no. 1 - genre analysis

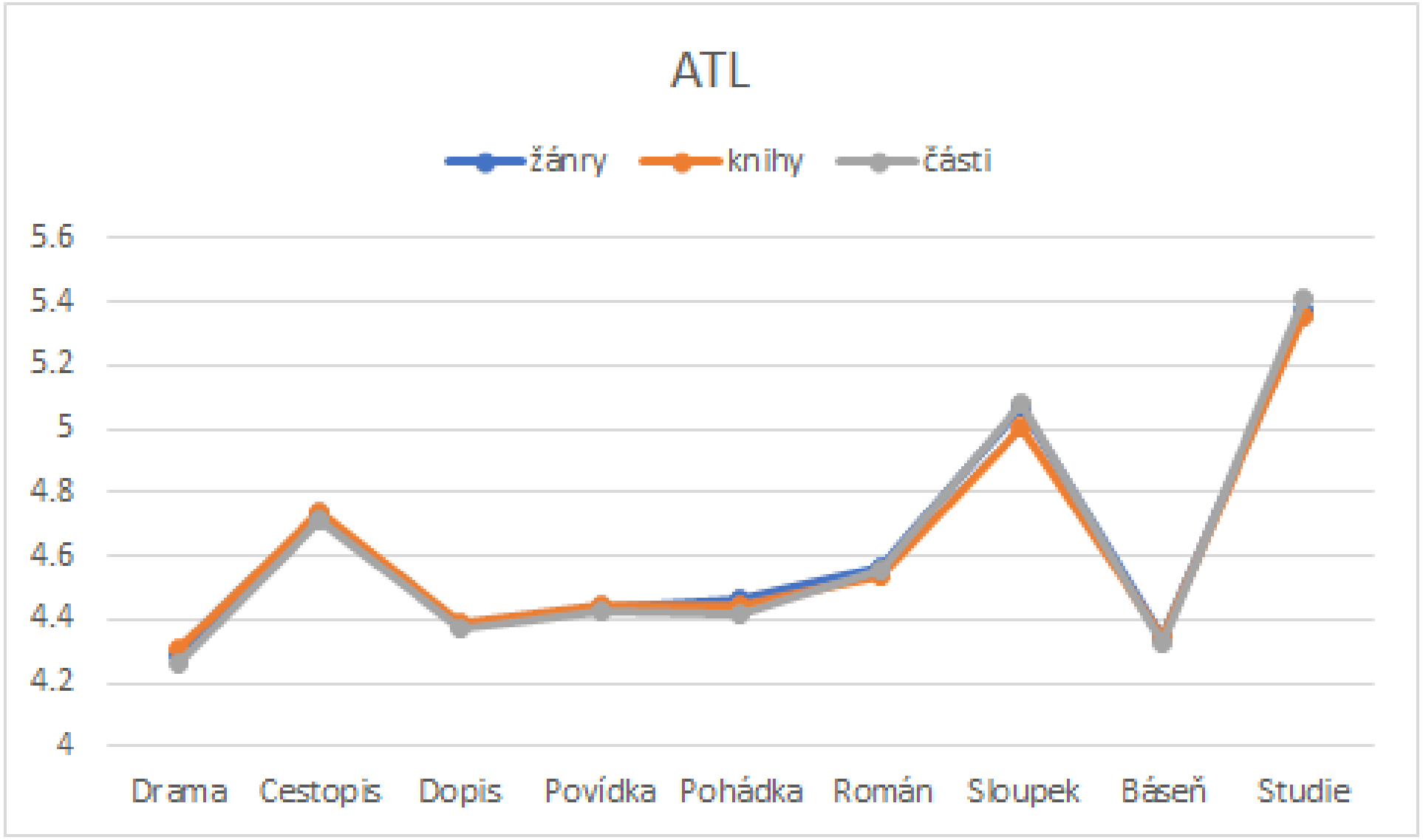
- M. Kubát
  - K. Čapek
  - analysis of particular methods
  - text as a basic unit

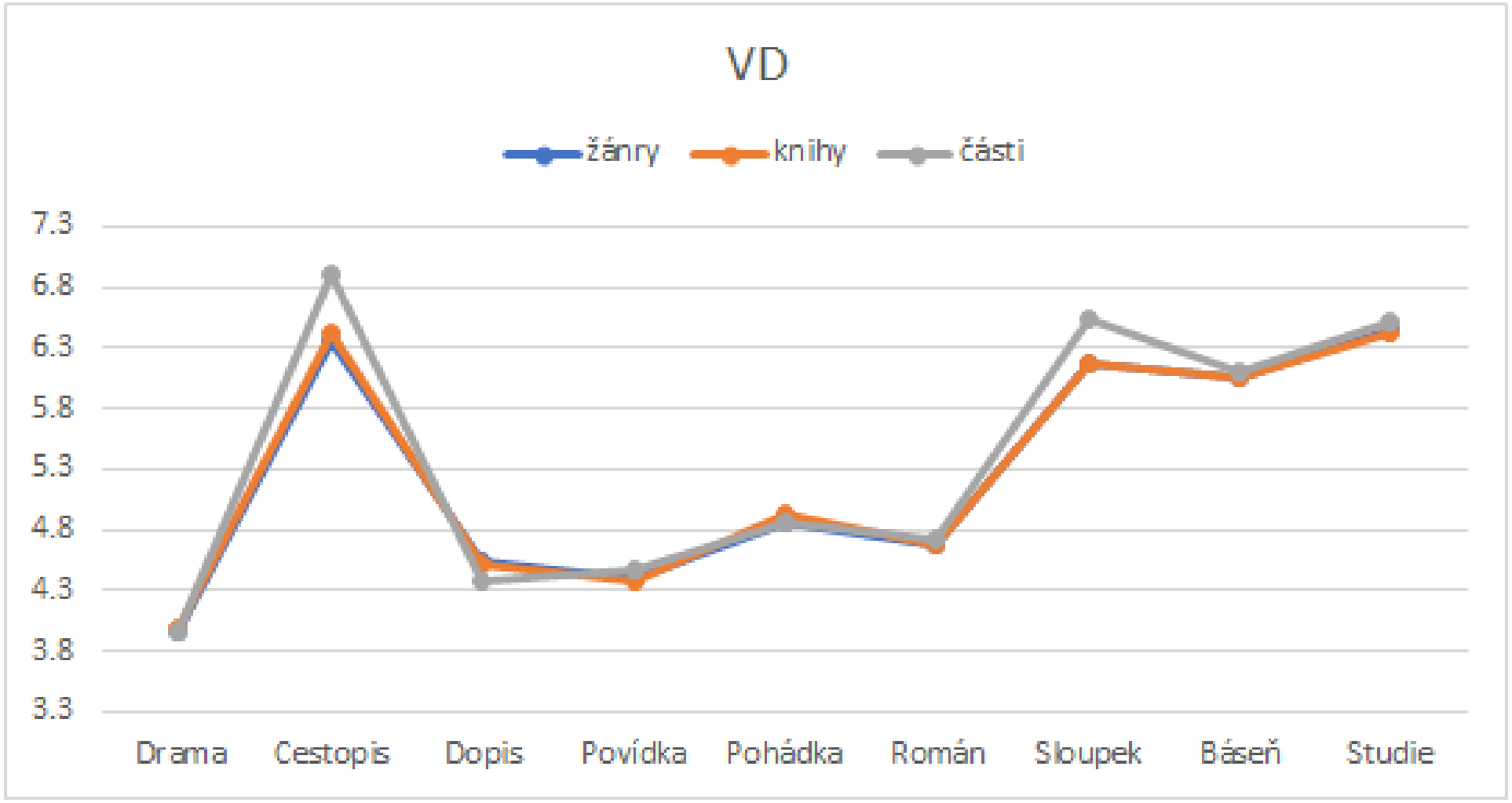
- D. Lukeš's review
  - genre analysis = text should be aggregated

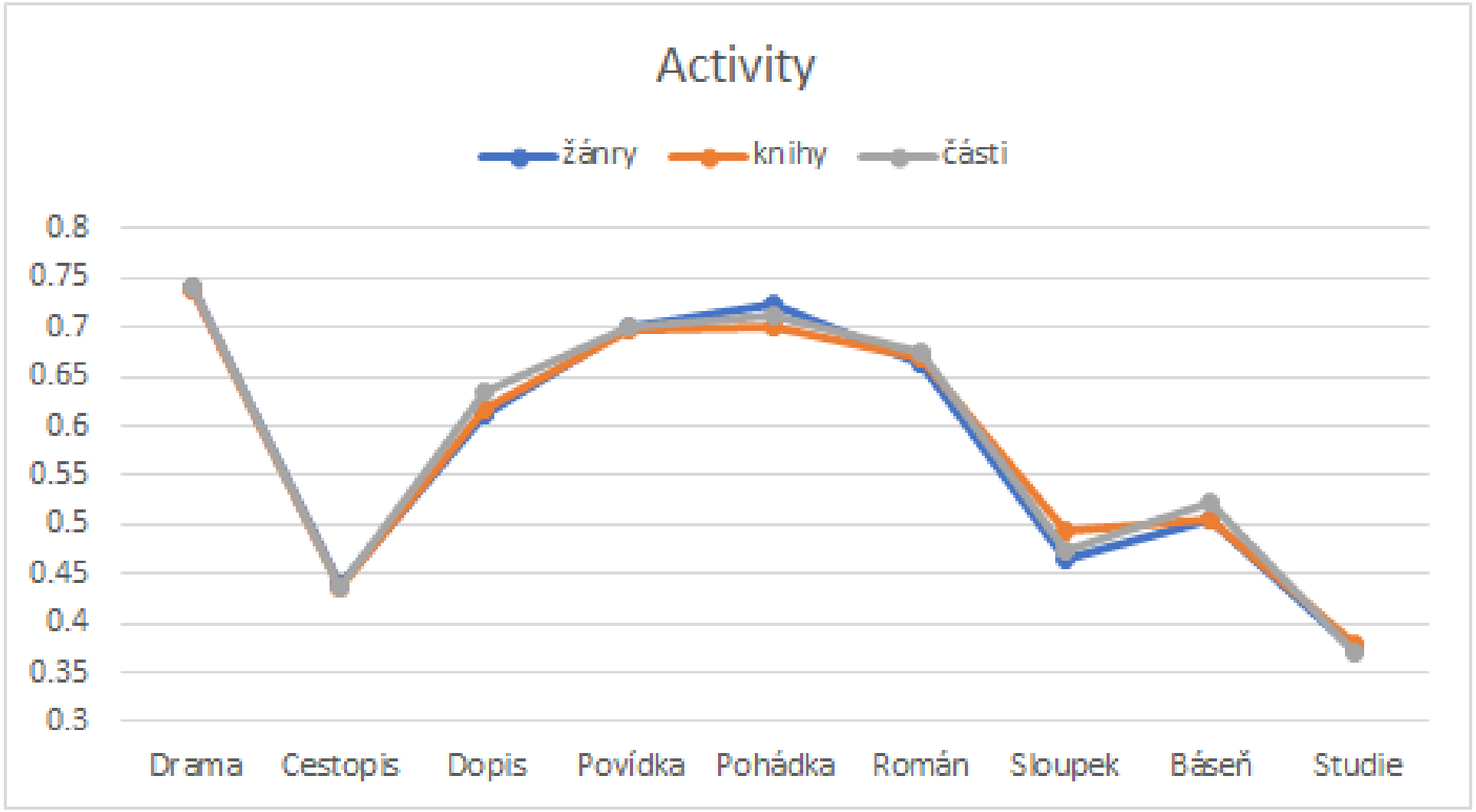
Case study no. 1 - genre analysis

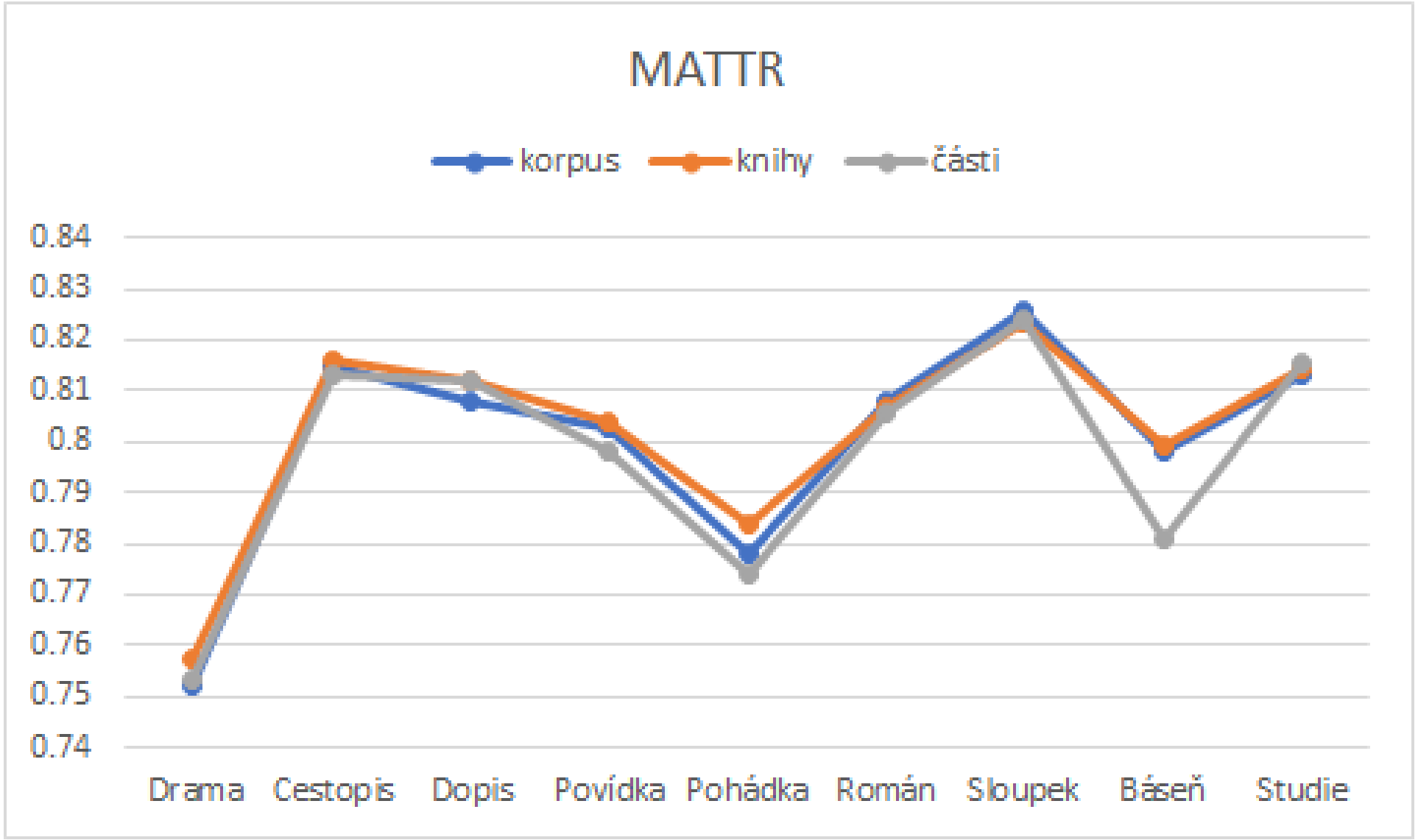
- M. Nogolová
- the goal
  - Are there differences between “text” and “corpus” approaches?
  - How (if so) do particular methods differ with regard to both approaches?

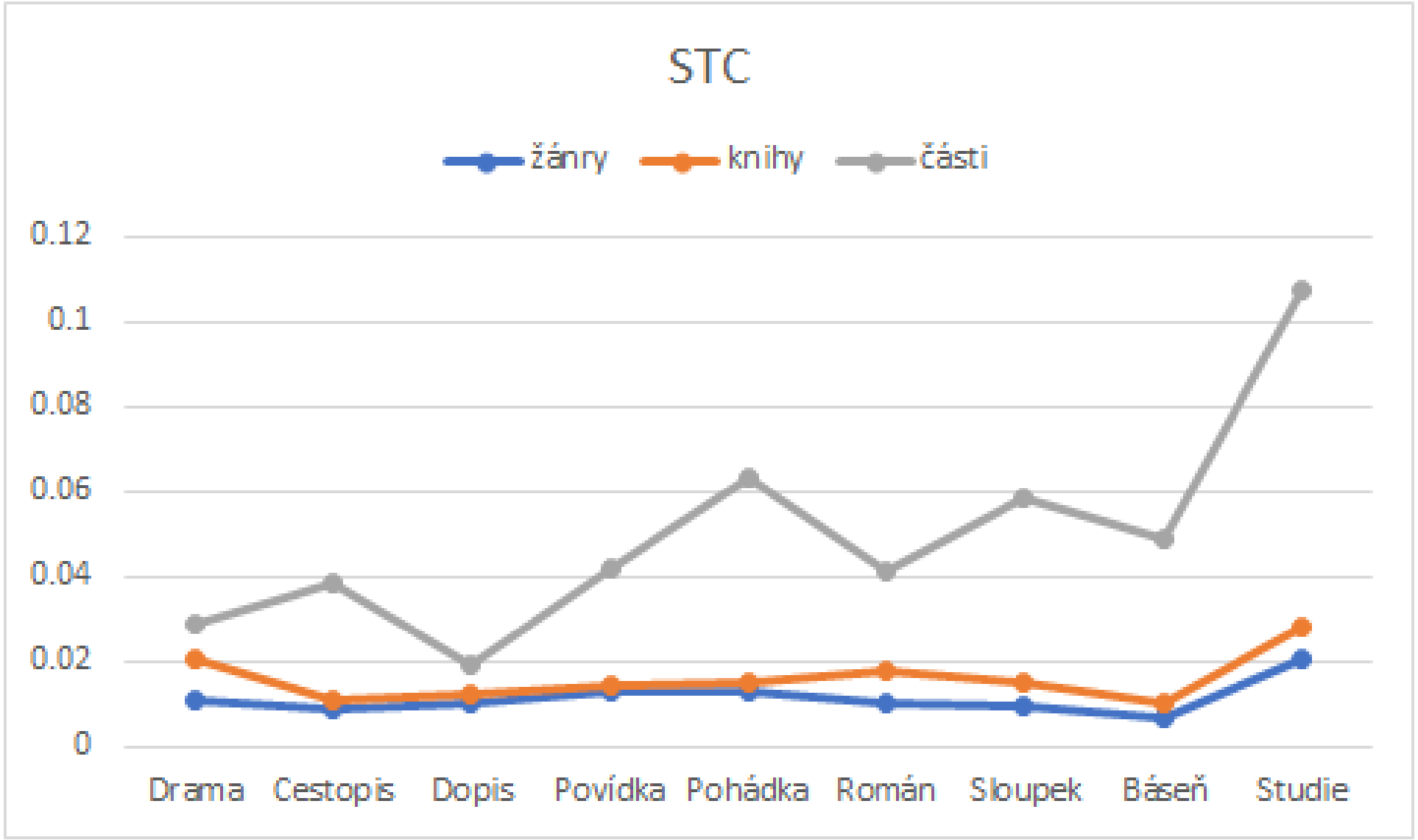
- analyzed units
  - chapters
  - books
  - corpus ot particular genre


ATL
žánry
knihy
části
5.6
5.4
5.2
5
4.8
4.6
4.4
4.2
4
Drama
Cestopis
Dopis
Povídka
Pohádka
Román
Sloupek
Báseň
Studie


VD
žánry
knihy
části
7.3
6.8
6.3
5.8
5.3
4.8
4.3
3.8
3.3
Drama
Cestopis
Dopis
Povídka
Pohádka
Román
Sloupek
Báseň
Studie


Activity
žánry
knihy
části
0.8
0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
Drama
Cestopis
Dopis
Povídka
Pohádka
Román
Sloupek
Báseň
Studie

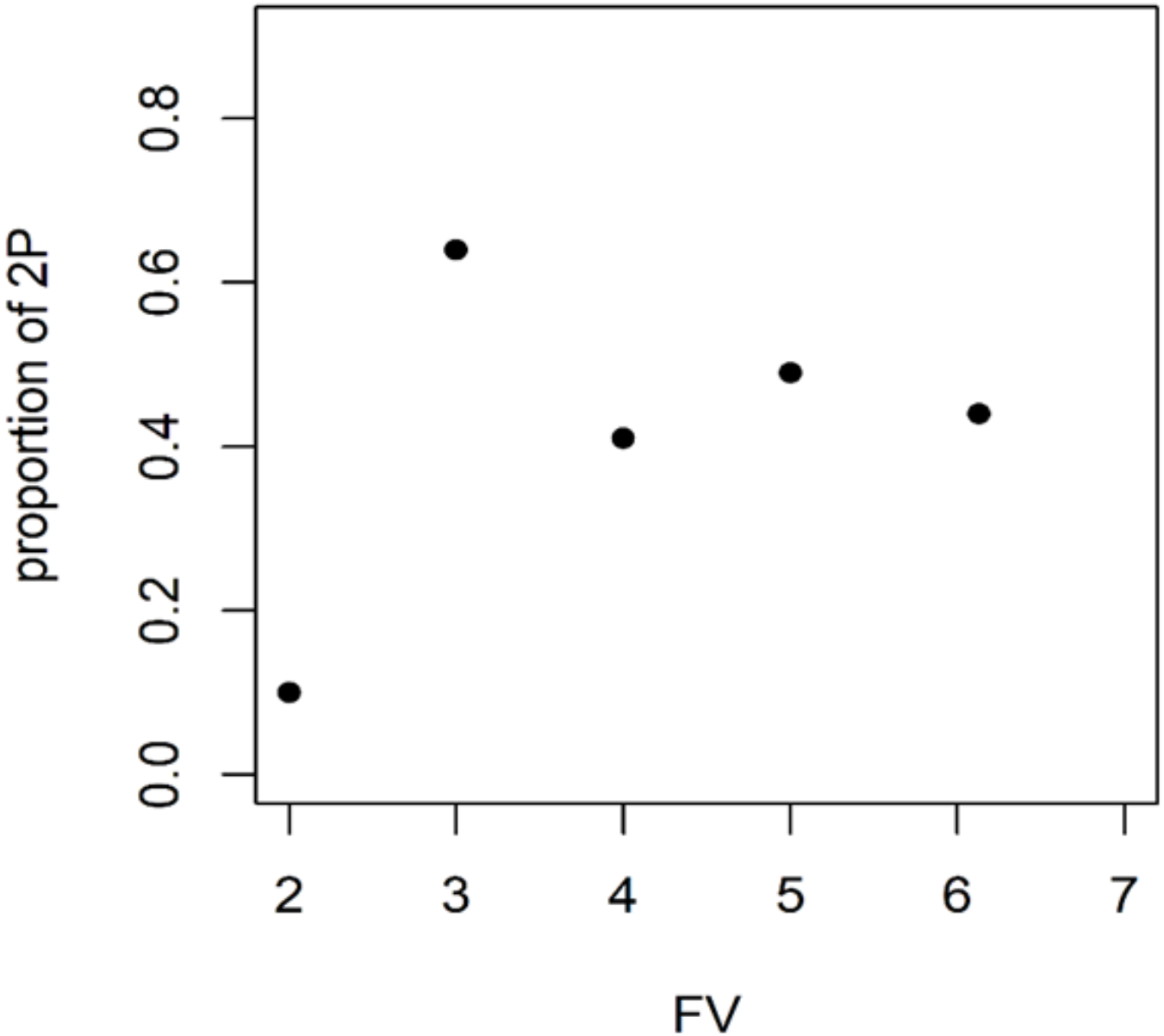

MATTR
korpus
knihy
části
0.84
0.83
0.82
0.81
0.8
0.79
0.78
0.77
0.76
0.75
0.74
Drama
Cestopis
Dopis
Povídka
Pohádka
Román
Sloupek
Báseň
Studie

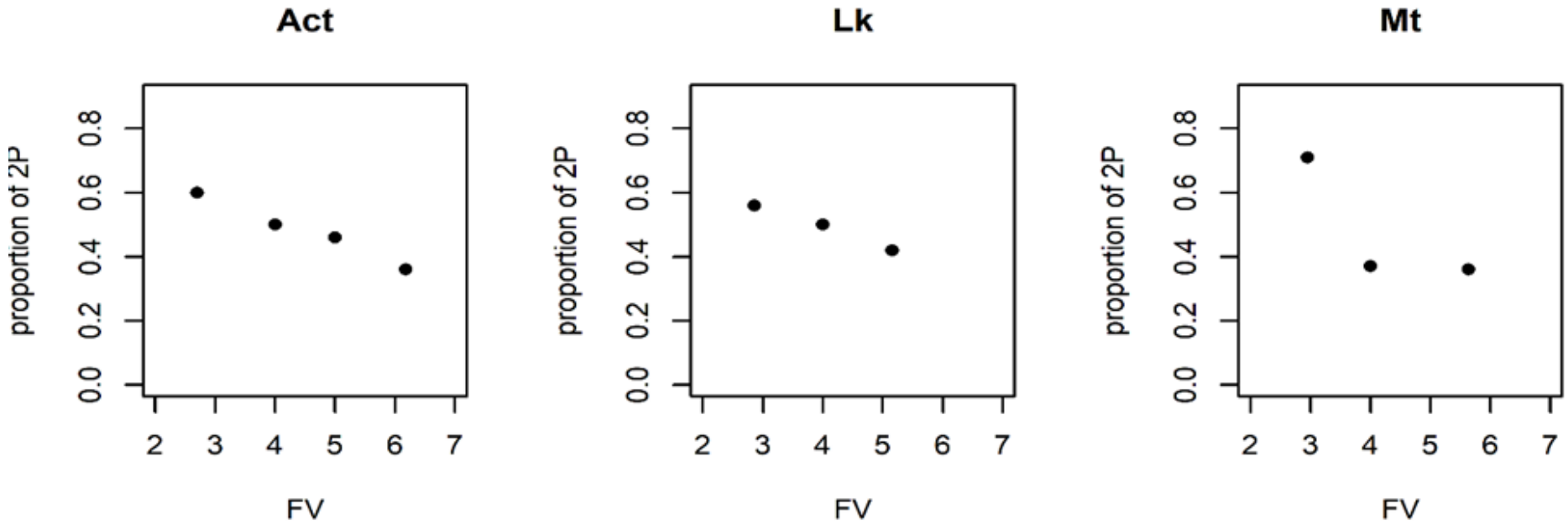

STC
žánry
knihy
části
0.12
0.1
0.08
0.06
0.04
0.02
0
Drama
Cestopis
Dopis
Povídka
Pohádka
Román
Sloupek
Báseň
Studie

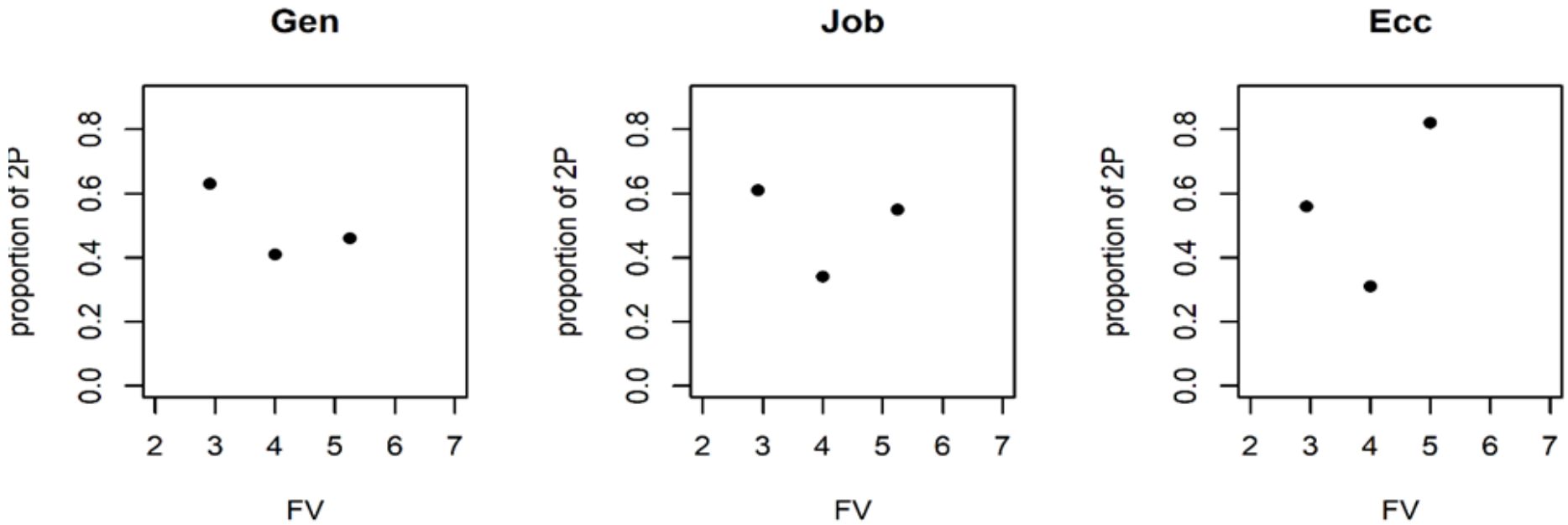
Case study no. 2 - word order of enclitics vs. full valency

- hypothesis
  - The higher the full valency of the predicate, the lower the probability of the occurrence of the enclitic after the initial phrase of the clause.

- Old Testament
  - Genesis (Gen), Isaiah (Is), Job (Job), Ecclesiastes (Ecc)
- New Testament
  - Gospel of St. Matthew (Mt), Gospel of St. Luke (Lk), Acts (Act), and Revelation (Rev)


proportion of 2P
0.0
0.2
0.4
0.6
0.8
2
3
4
5
6
7
FV


Act
Lk
Mt
proportion of 2P
0.0
0.2
0.4
0.6
0.8
2
3
4
5
6
7
FV


Gen
Job
Ecc
proportion of 2P
0.0
0.2
0.4
0.6
0.8
2
3
4
5
6
7
FV

Case study no. 3 - the Menzerath-Altmann law

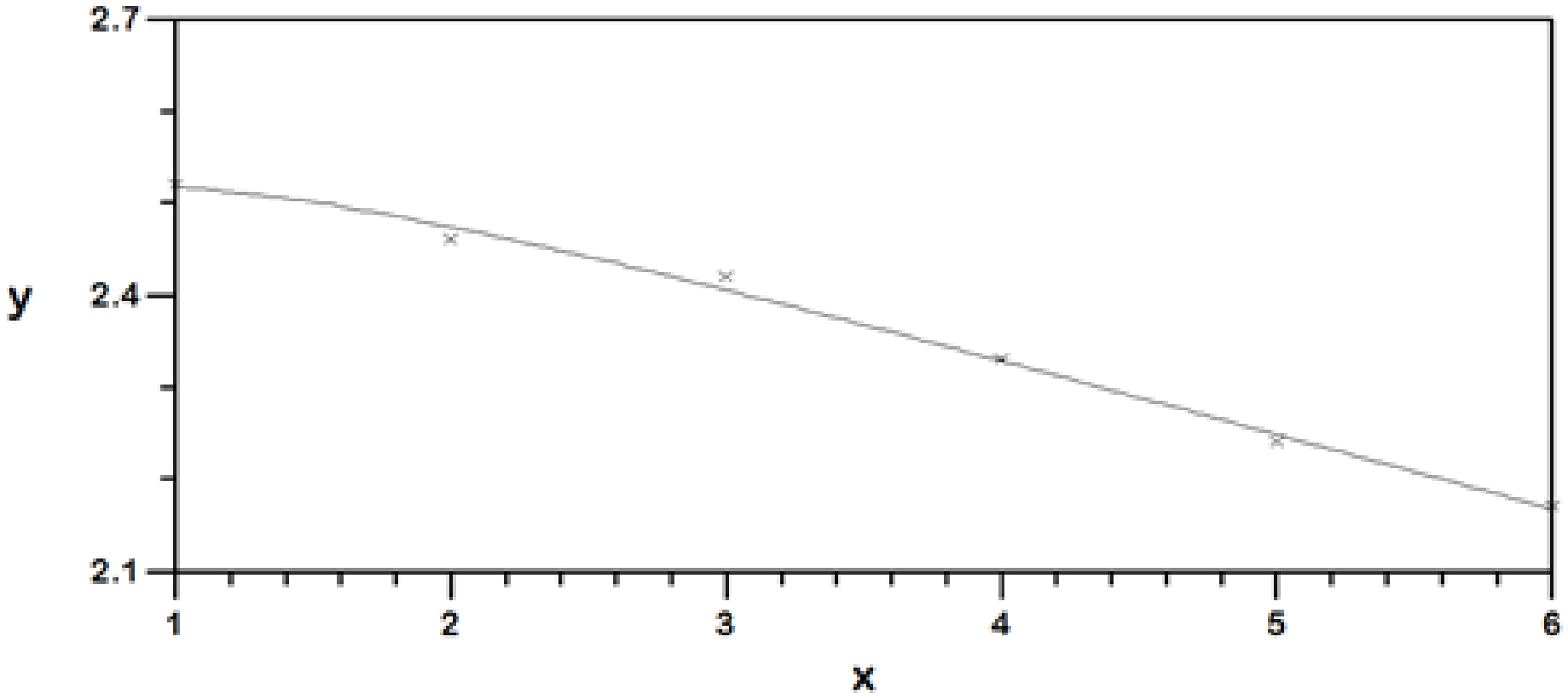
- word - syllable - sound

$$y(x) = ax^b e^{-cx}$$

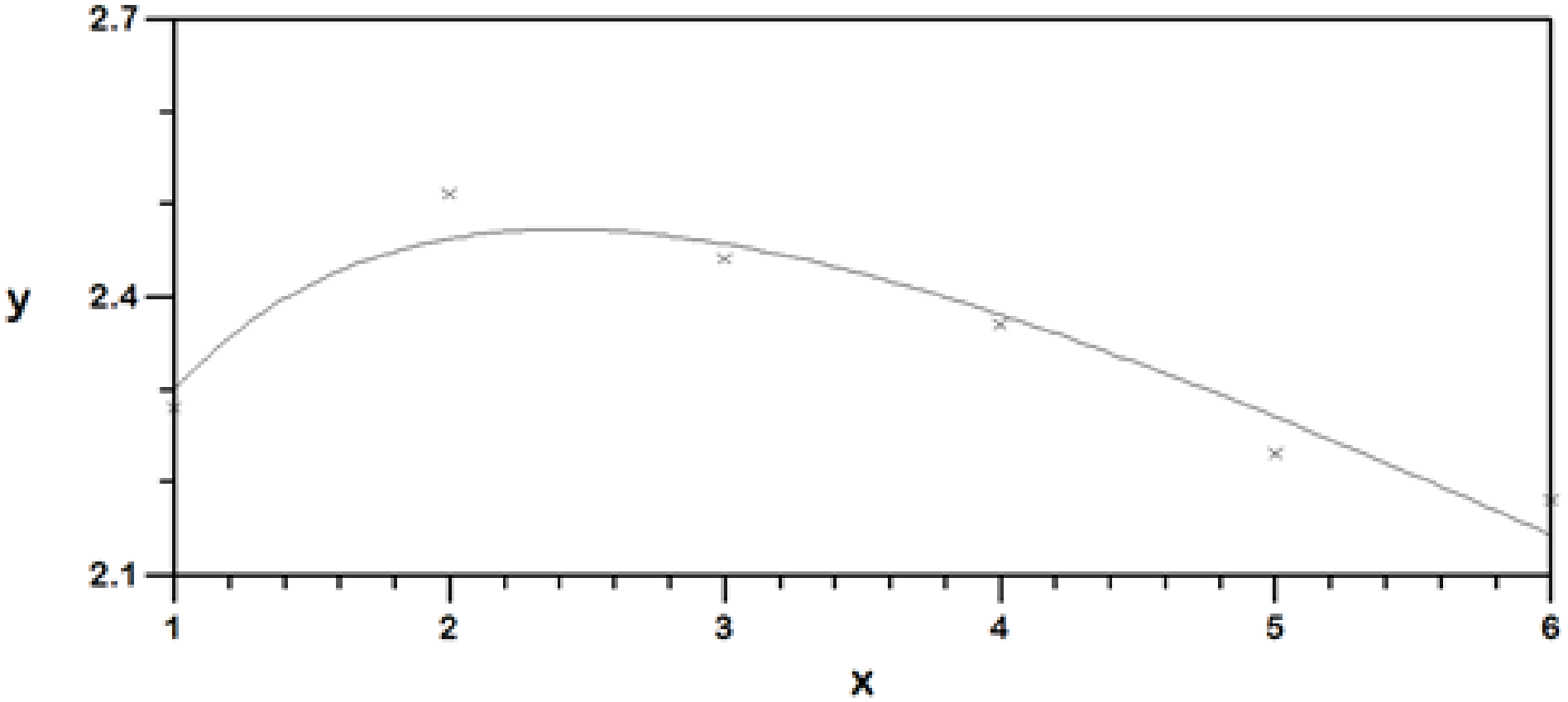
- 5 texts
  - Gottwald 1949
  - Havel 1999
  - V. Cvrček: Za ještě tvrdší kodifikační diktát? (Naše řeč, 2006, s. 26-29)
  - K. J. Erben: Zlatý kolovrat
  - 10 years child: Já a první žárovka

	vše dohromady	Gottwald	Havel	Cvrček	Erben	žákovský text
a	2.63	2.62	2.59	2.55	2.64	2.57
b	0.04	0.12	0.13	0.25	-0.15	0.05
c	-0.04	-0.07	-0.07	-0.1	0.01	-0.04
R^2	0.99	0.95	0.91	0.93	0.94	0.95

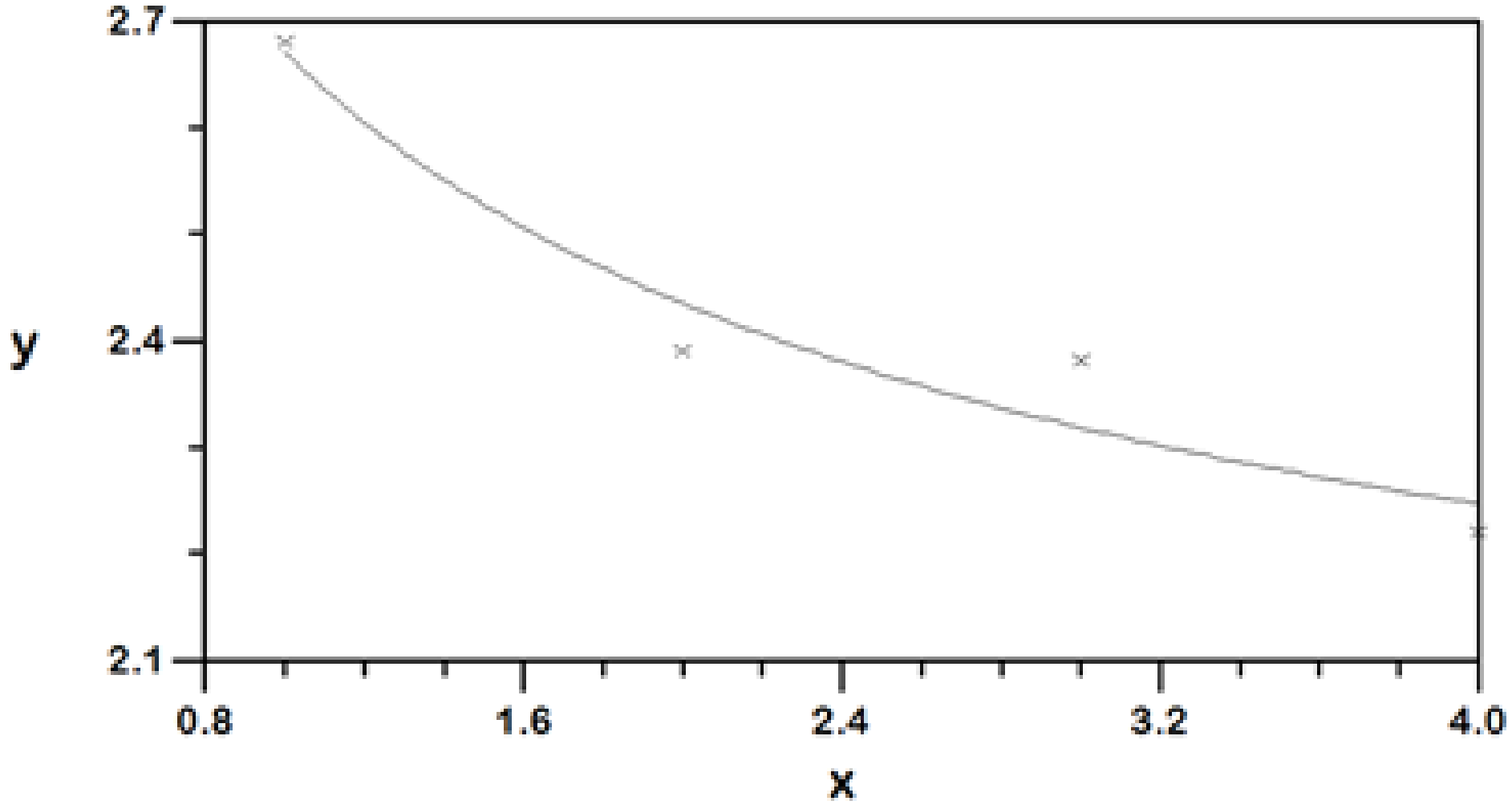
All texts together



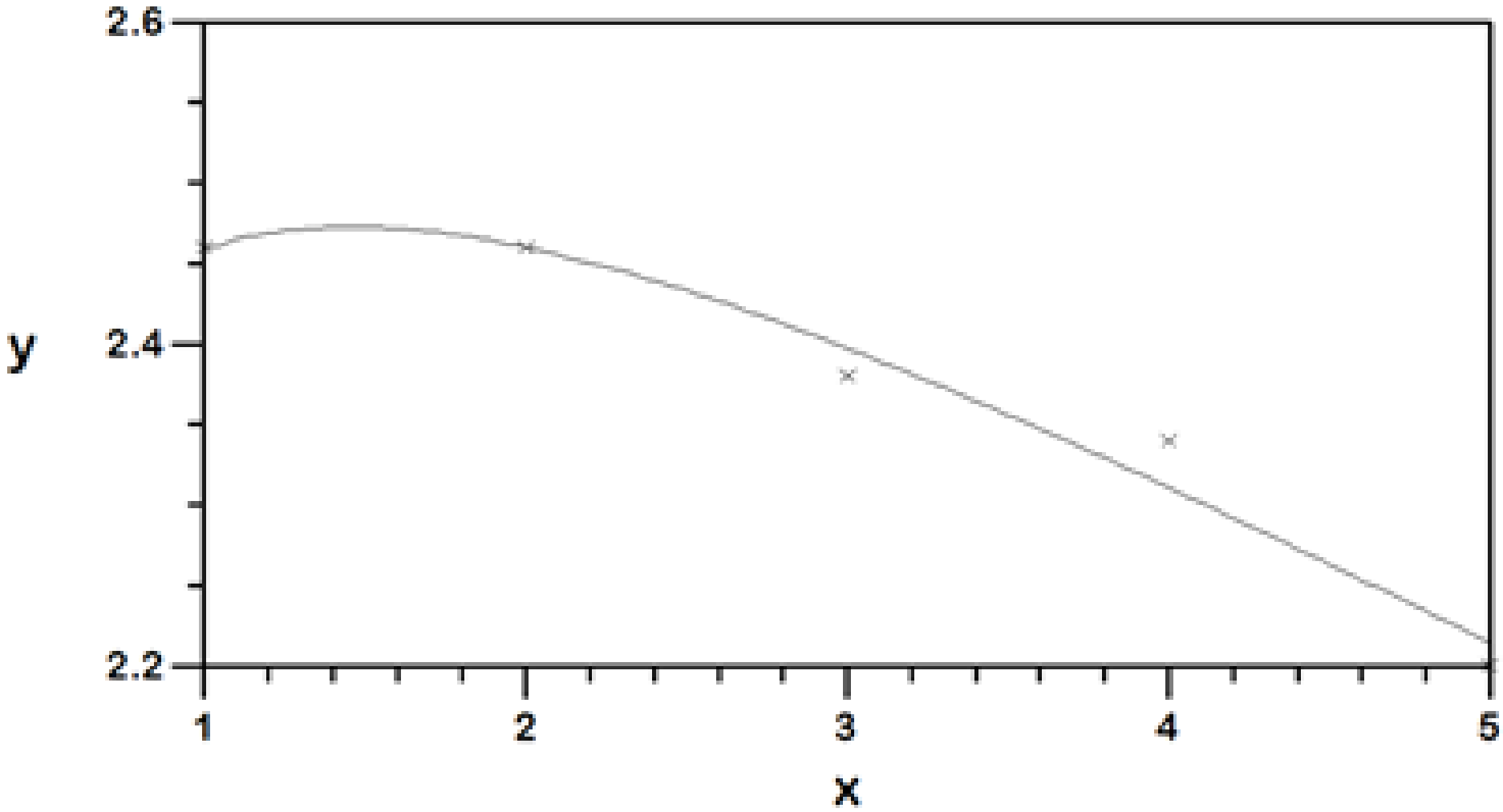
V. Cvrček: Za ještě tvrdší kodifikační diktát?



K. J. Erbena: Zlatý kolovrat



Já a první žárovka



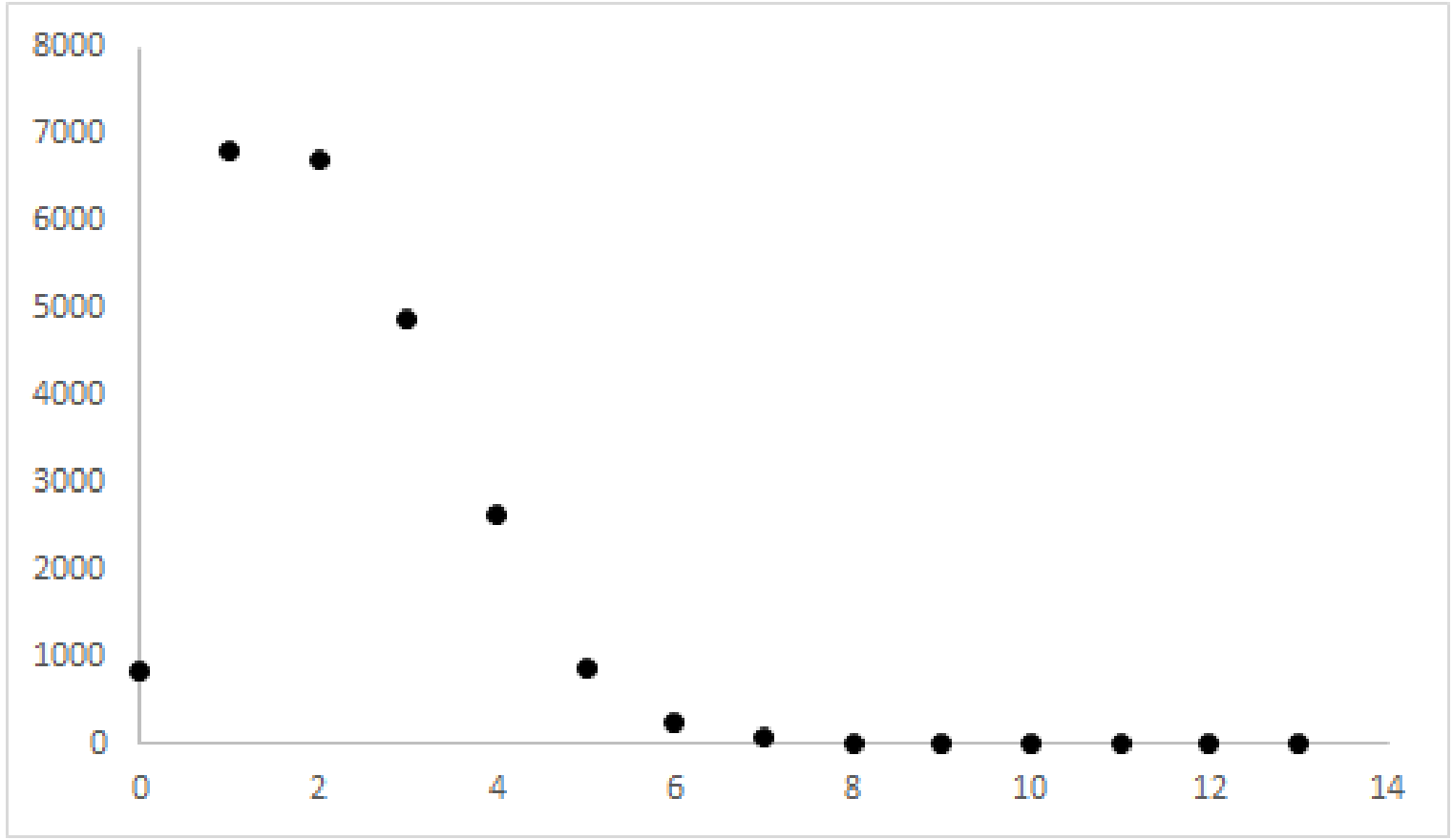
Case study no. 4 - distribution of word lengths

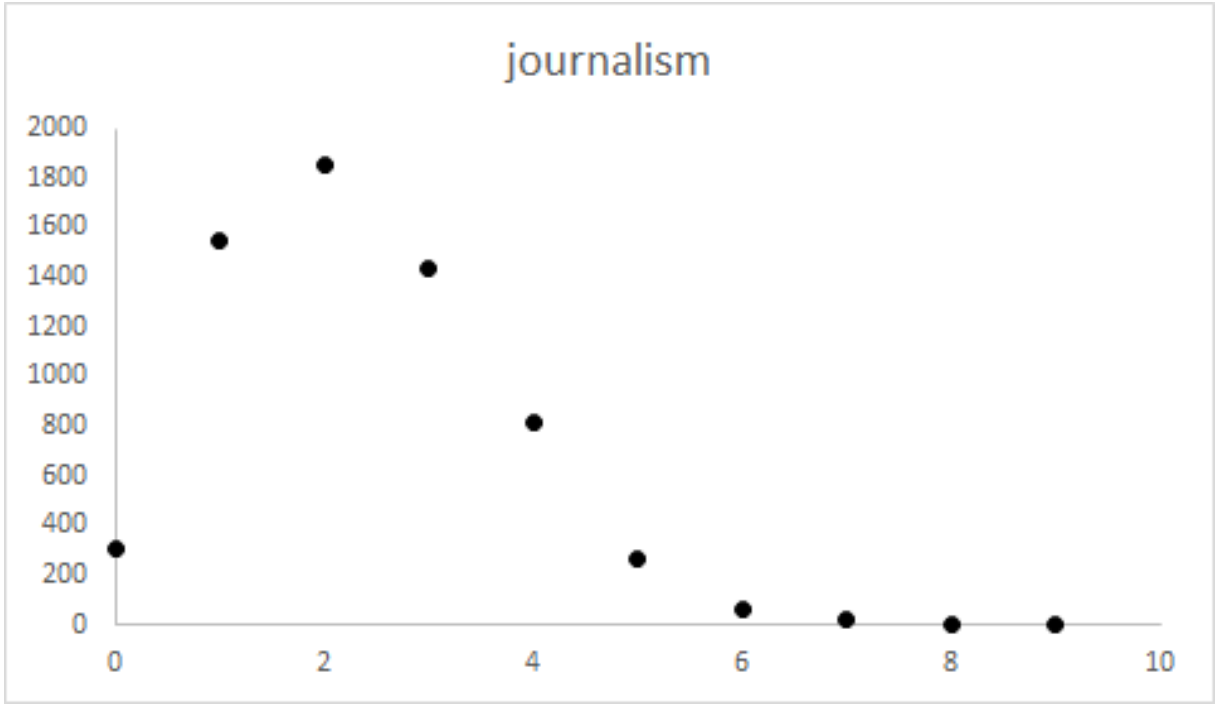
- K. Pelegrinová
- word length measured in syllables
- 15 texts of 3 genres

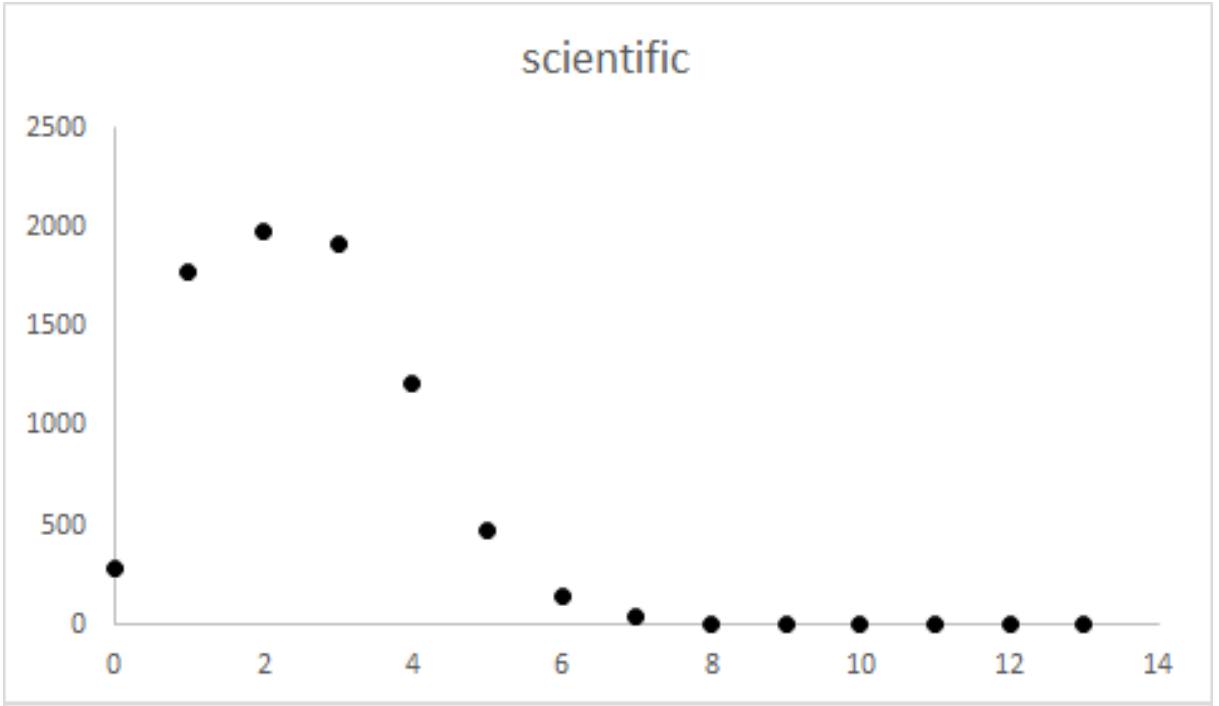
All texts together

hyper-Poisson distribution

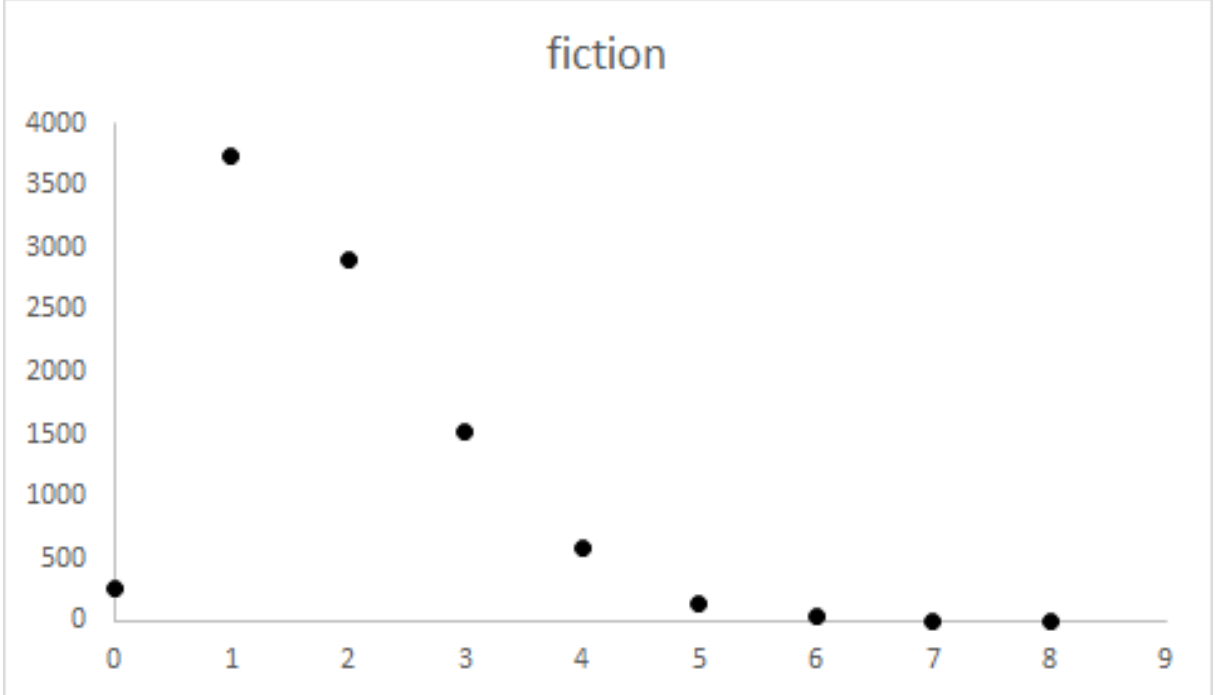
a = 1.46
b = 0.18
C = 0.012



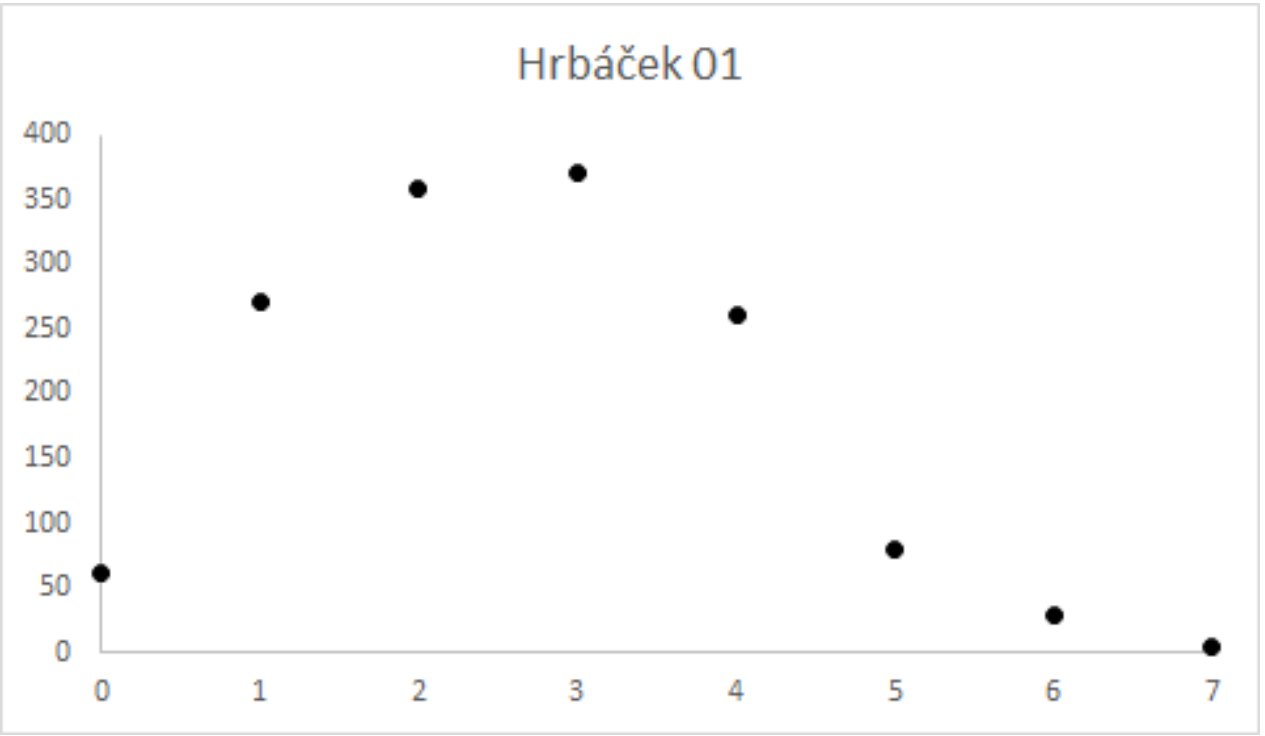


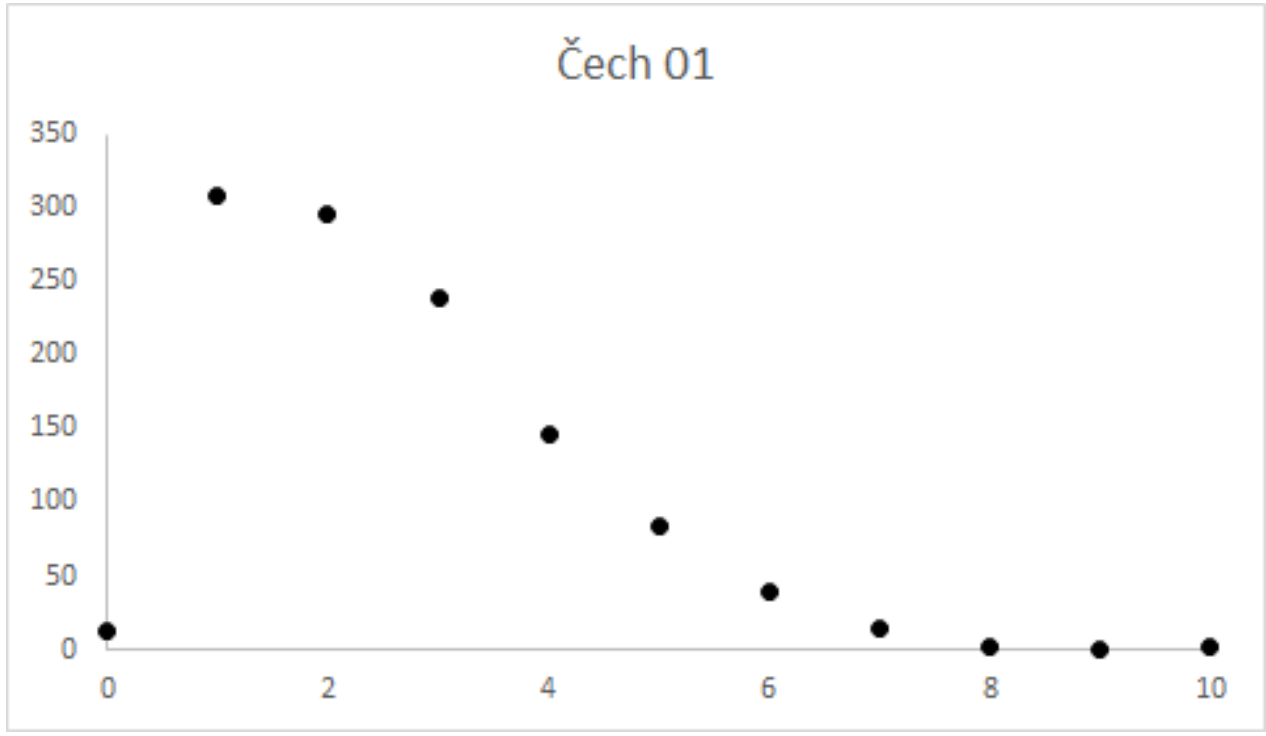


	a	b	C
journalism	1.703	0.339	0.005
scientific	1.904	0.335	0.012
fiction	0.982	0.068	0.008



Scientific


Hrbáček 01
400
350
300
250
200
150
100
50
0
0
1
2
3
4
5
6
7


Čech 01
350
300
250
200
150
100
50
0
0
2
4
6
8
10

Opakování 3

- intuice → budou delší slova a větší rozptyl u Ortenovy básně, nebo u Škvoreckého povídky?

- spočítejte

Tematická analýza textu

- jak na ni?

Tematická analýza textu

- frekvence?

Tematická analýza textu

- analýza klíčových slov

- analýza tematických slov