

From Text to Corpus. There and Back Again

Radek Čech
Michaela Nogolová
Kateřina Pelegrinová

Acknowledgments

- Supported by grant SGS02/FF/2019 (Radek Čech)
- Supported by grant RRC/10/2019 (Kateřina Pelegrinová)
 - výstup vznikl s finanční podporou Moravskoslezského kraje v rámci řešení projektu Podpora talentovaných studentů doktorského studia na Ostravské univerzitě III (kód programu RRC/10/2019)

Background

text → corpus → text

Background

text → corpus → text

corpus → text → corpus → text

Background

text → corpus → text

corpus → text → corpus → text

corpus → text / corpus / text / corpus / text / corpus

text

corpus

text



corpus

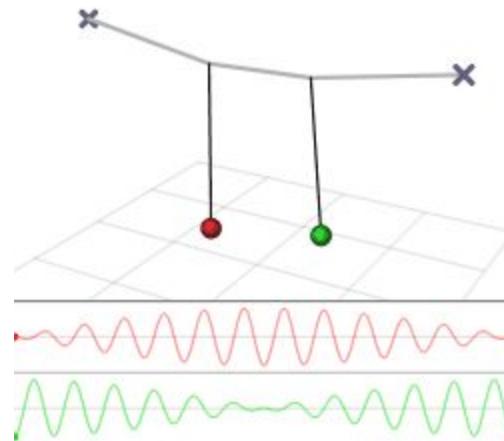
There and back again?

- no journey



There and back again?

- rather an oscillation



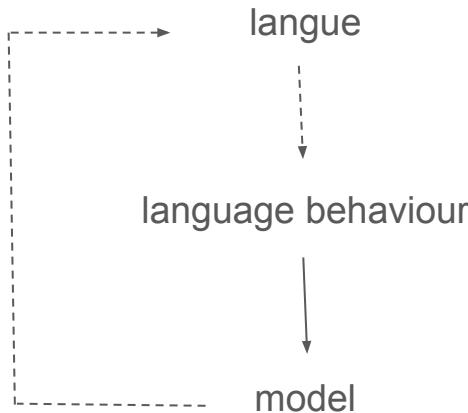
What do we model?

- langue
 - competence
-

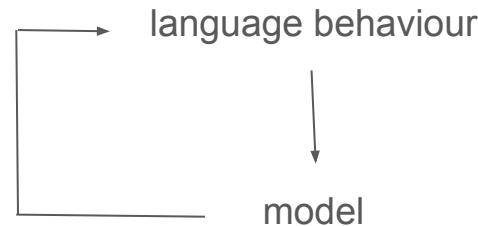
- language behaviour

What do we model?

- L-LB-M-L approach

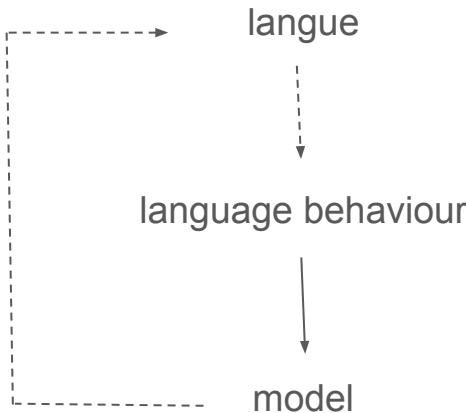


- LB-M approach

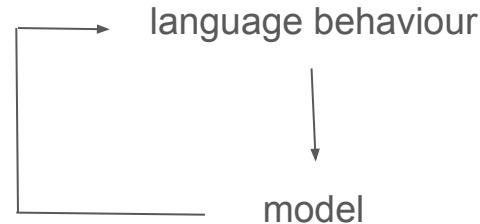


What do we model?

- L-LB-M-L approach



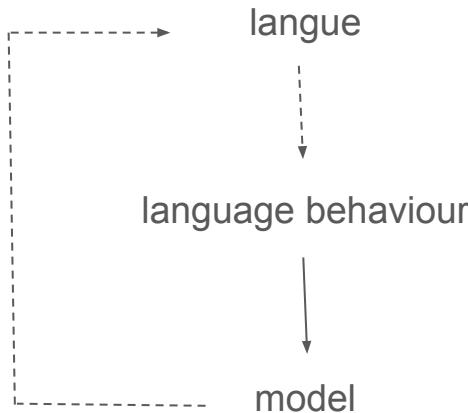
- LB-M approach



CORPUS

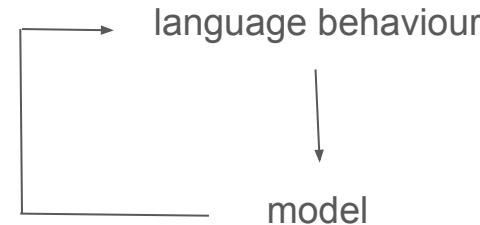
What do we model?

- 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

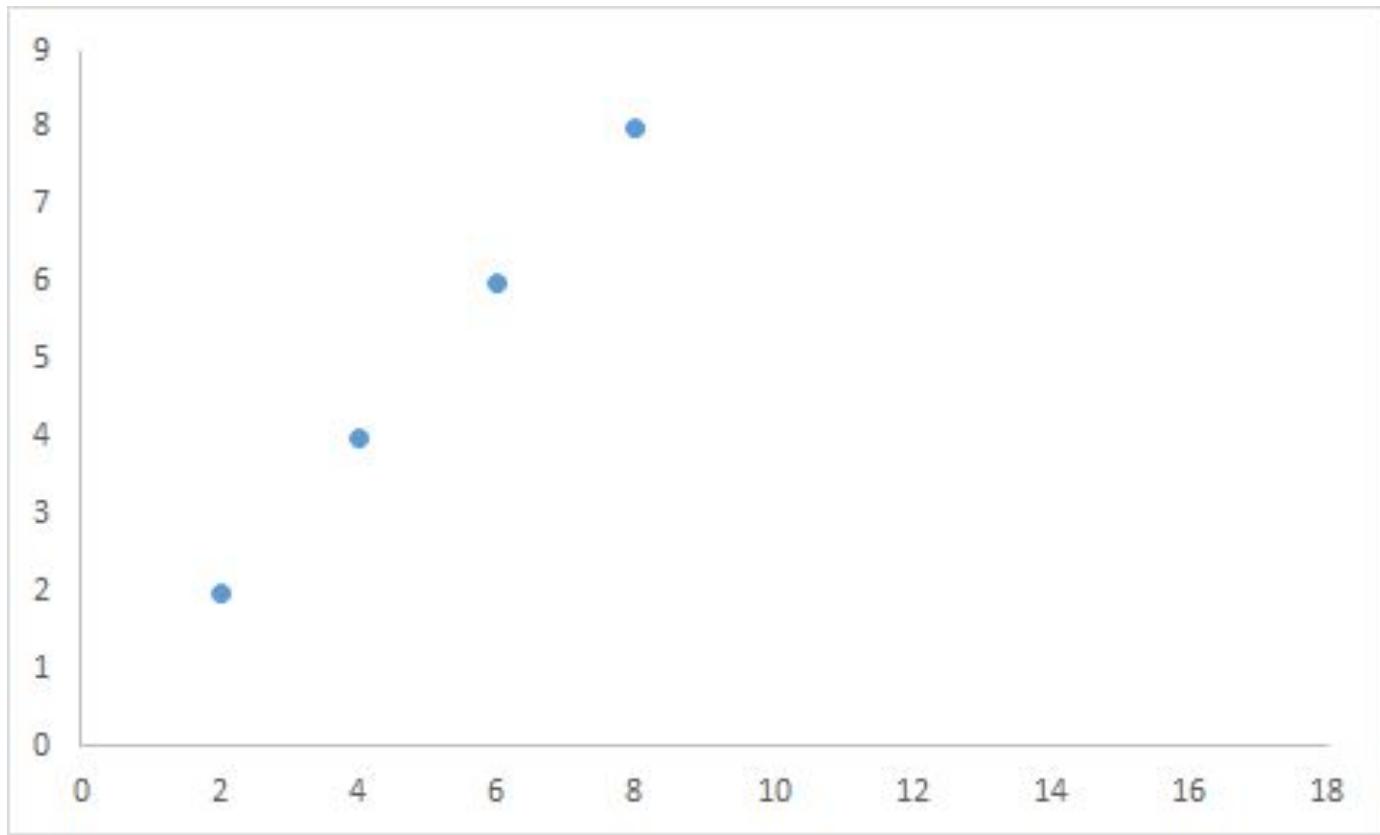
Language behavior → model

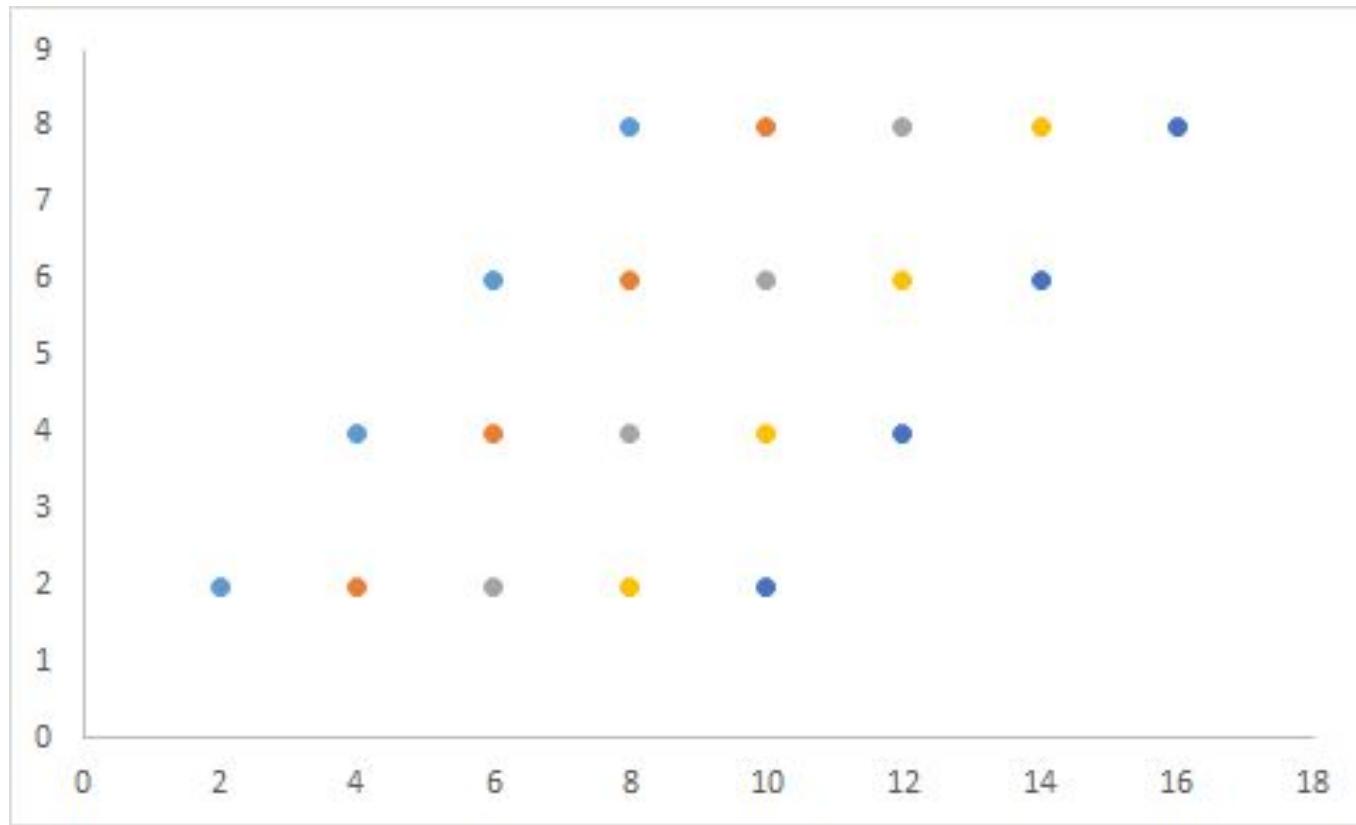
- inductive approach
 - regularities, correlations etc.
 - generalisations
 - description
 - corpus driven
- deductive approach
 - theoretical assumptions
 - theory
 - Zipf (least effort principle)
 - synergetic linguistics
 - model of mechanism
 - language behaviour

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)

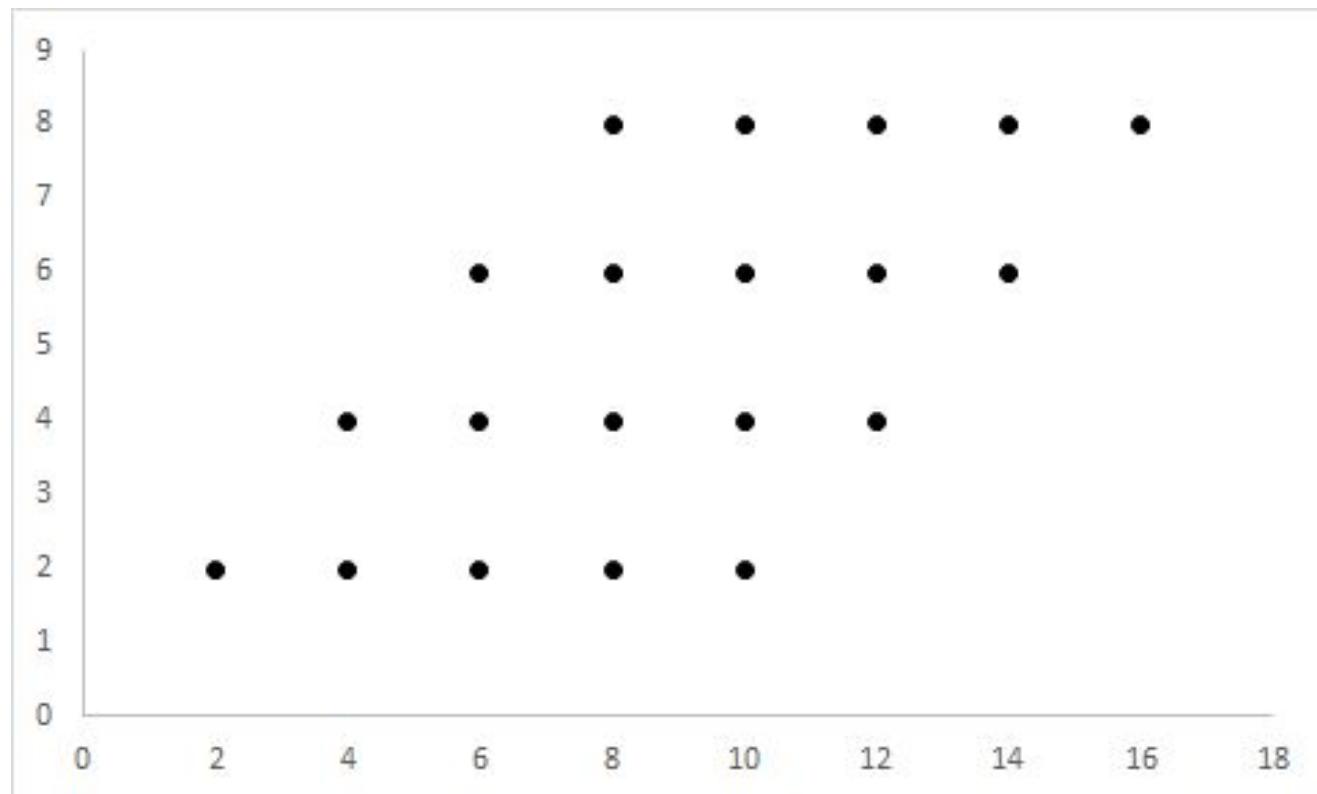
a) Ak chceme overovať nejaký zákon, tak nesmieme miešať texty, lebo v každom texte sú iné tzv. počiatočné podmienky. Dokonca je niekedy potrebné analyzovať oddelene aj jednotlivé kapitoly románu alebo vety symfónie. Dá sa napríklad ukázať, že zákon pre rozdelenie dĺžky slova je ľahko overiteľný napr. 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.





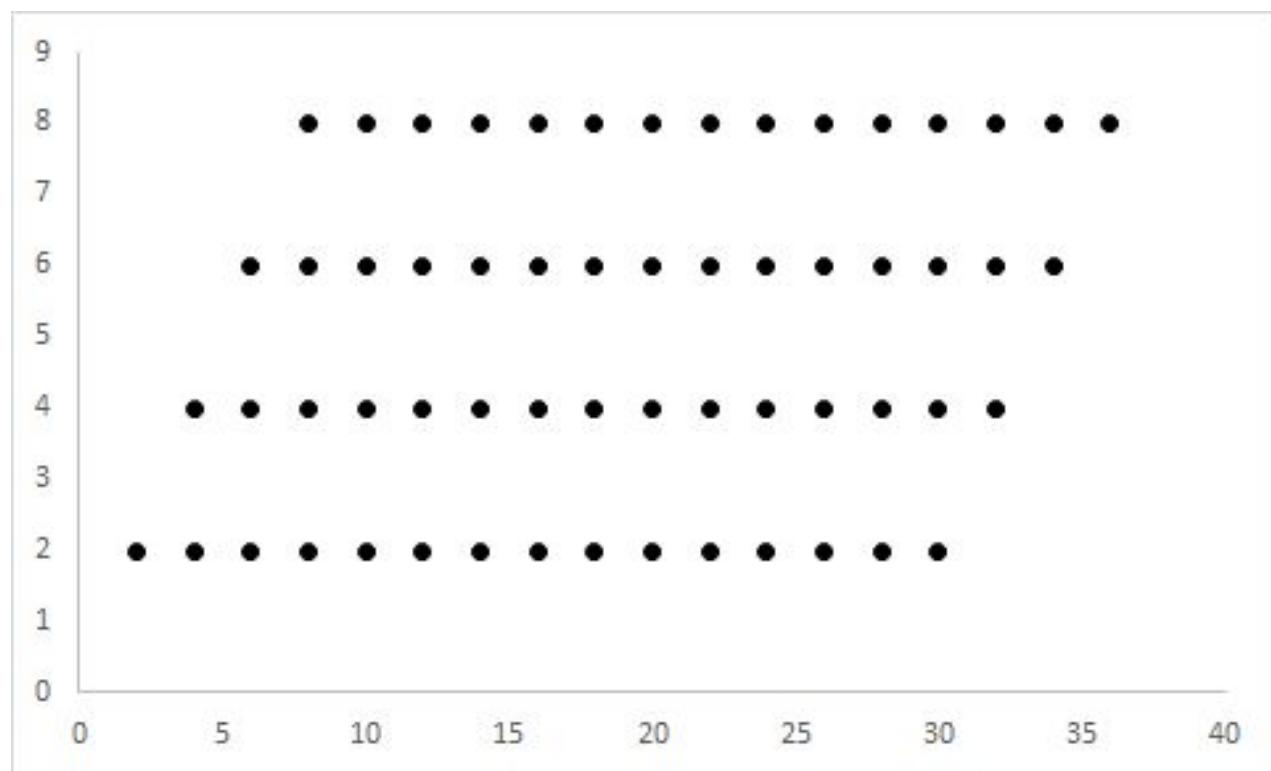
$\tau = 0.50$

p-value = 0.006



$\tau = 0.19$

p-value = 0.06

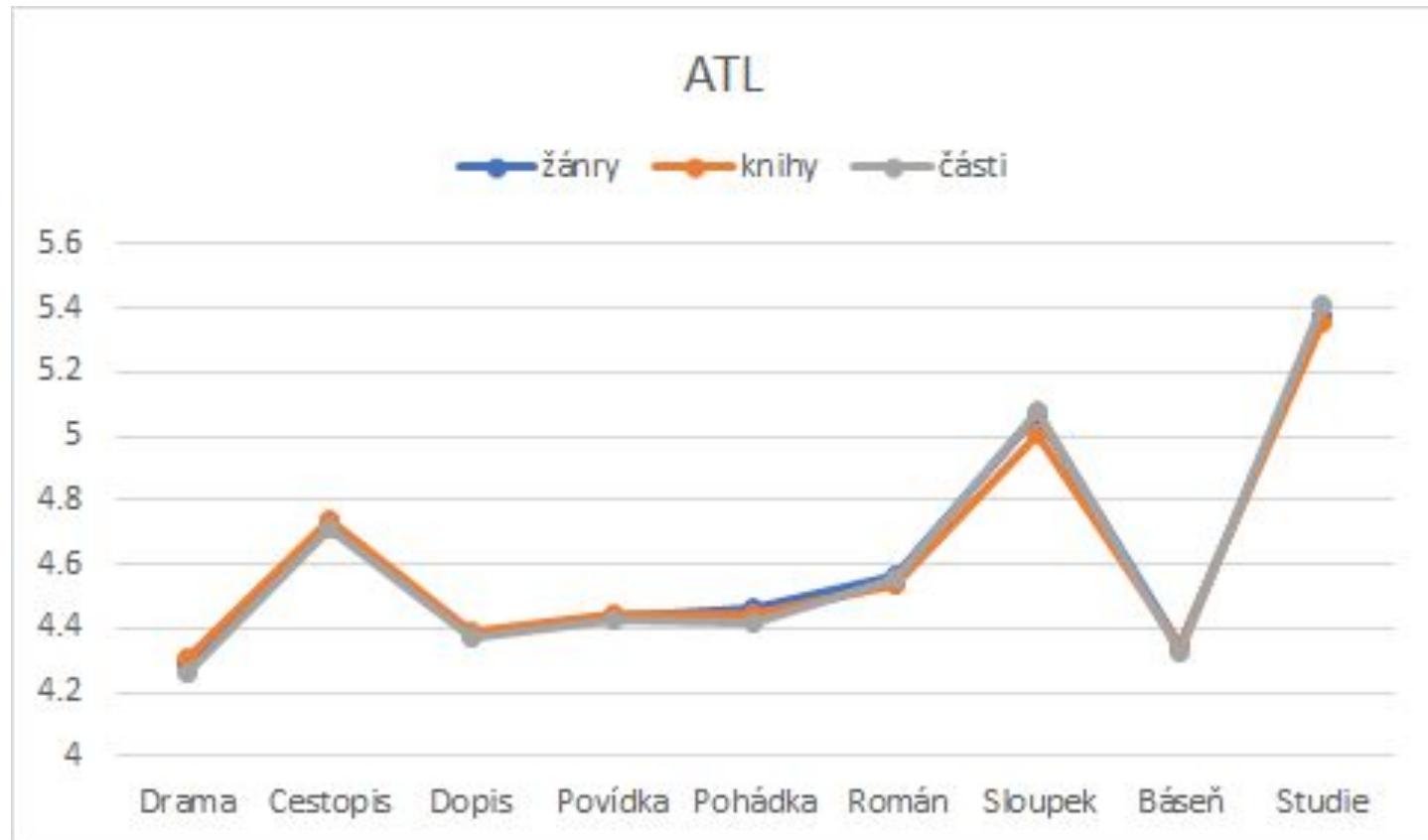


Case study no. 1 - genre analysis

- M. Kubát
 - K. Čapek
 - analysis of particular methods
 - text as a basic unit
- D. Lukeš's review
 - genre analysis = text should be aggregated

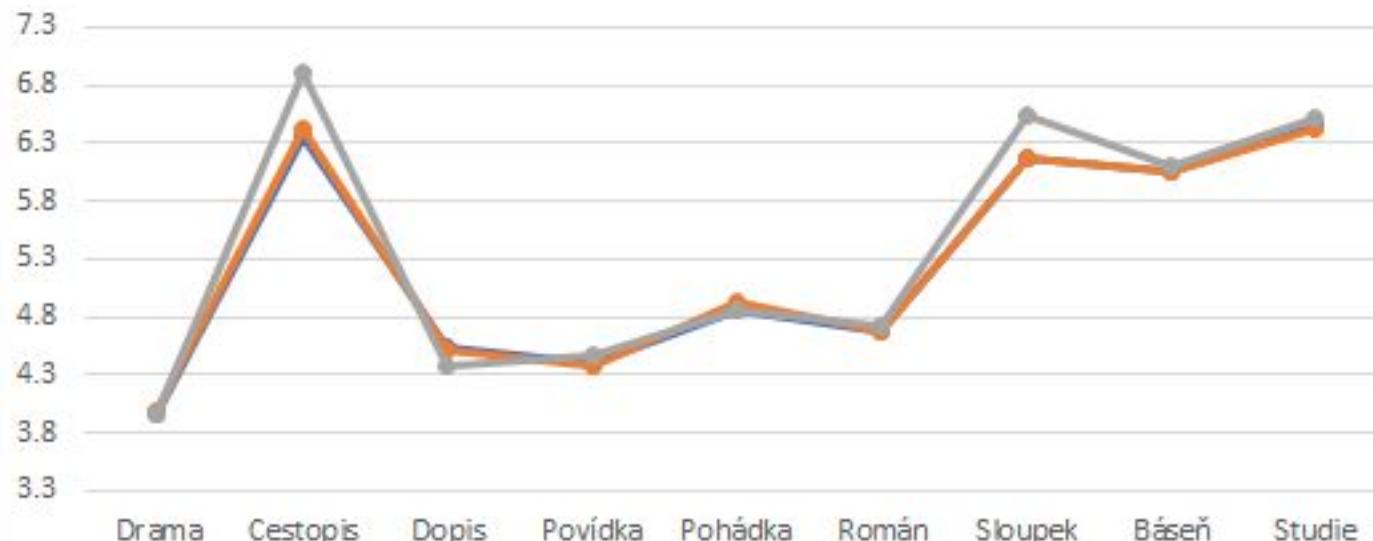
Case study no. 1 - genre analysis

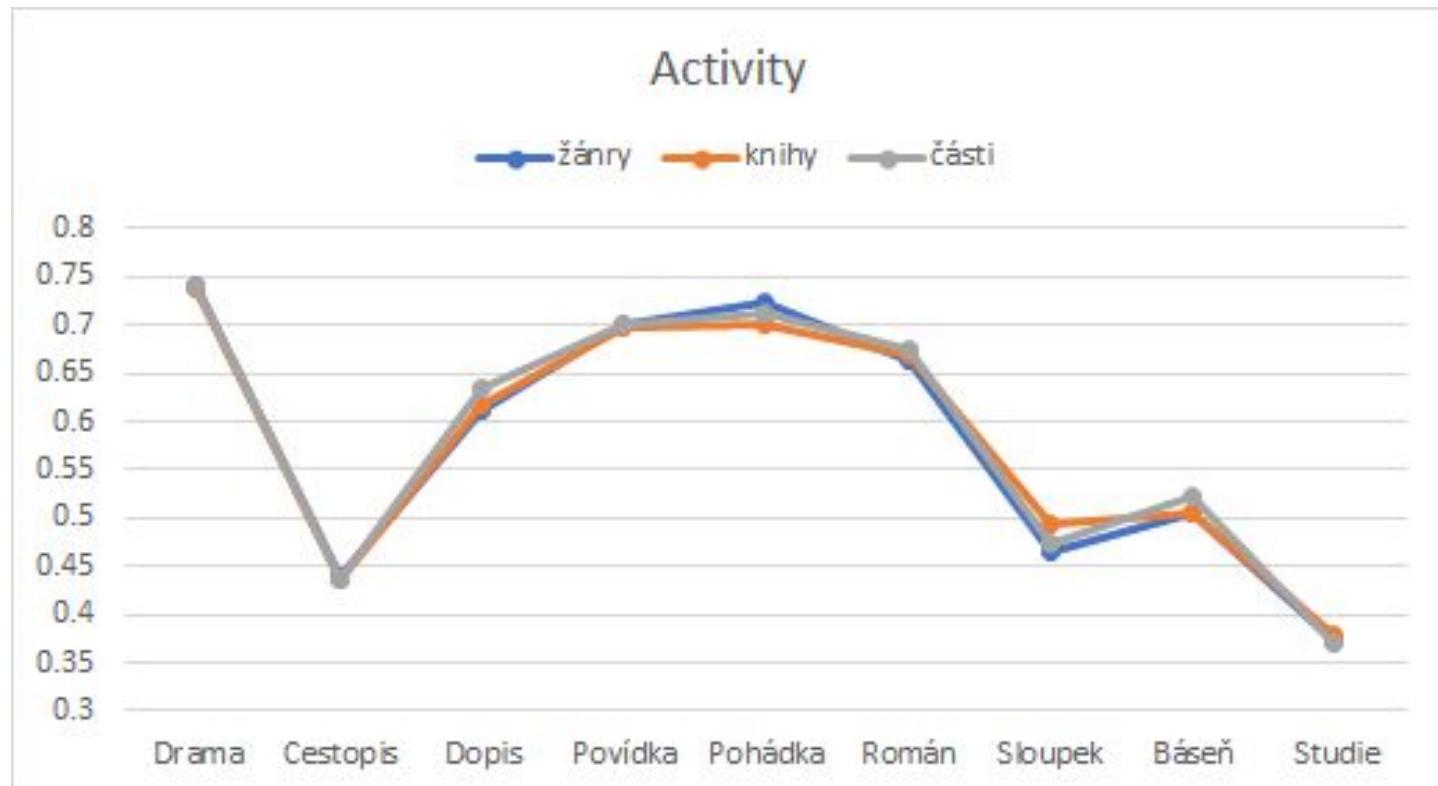
- M. Nogolová
- the goal
 - Are there differences between “text” and “corpus” approaches?
 - How (if so) do particular methods differ with regard to both approaches?
- analyzed units
 - chapters
 - books
 - corpus of particular genre



VD

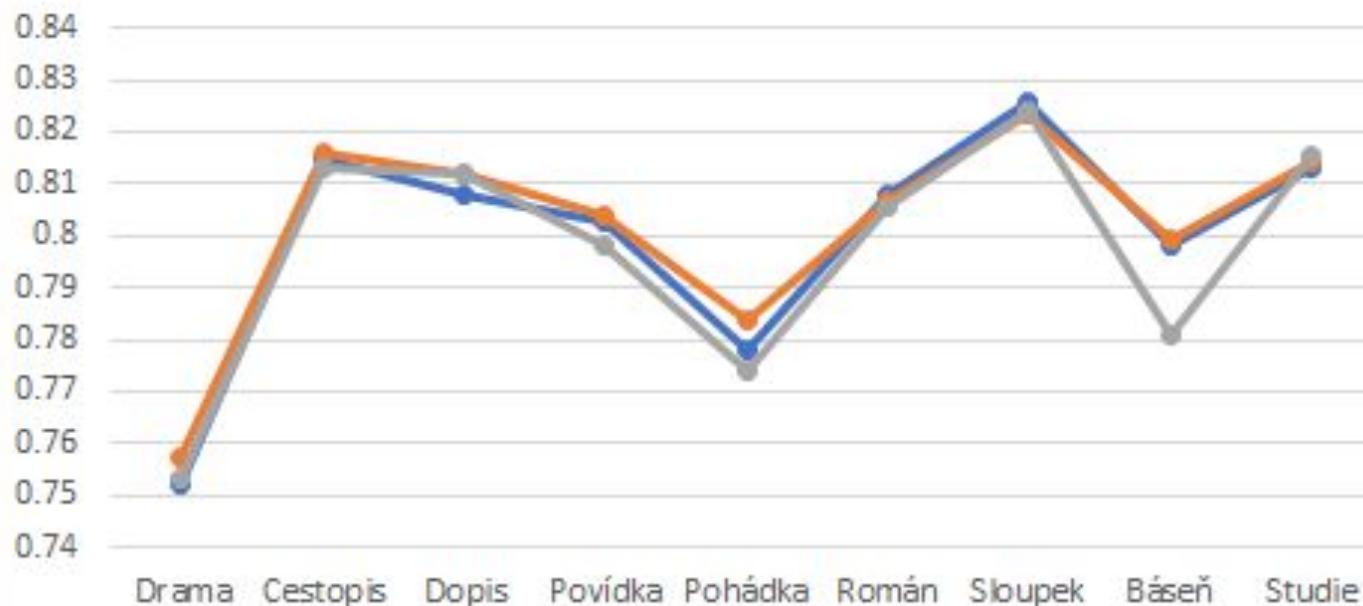
žánry knihy části





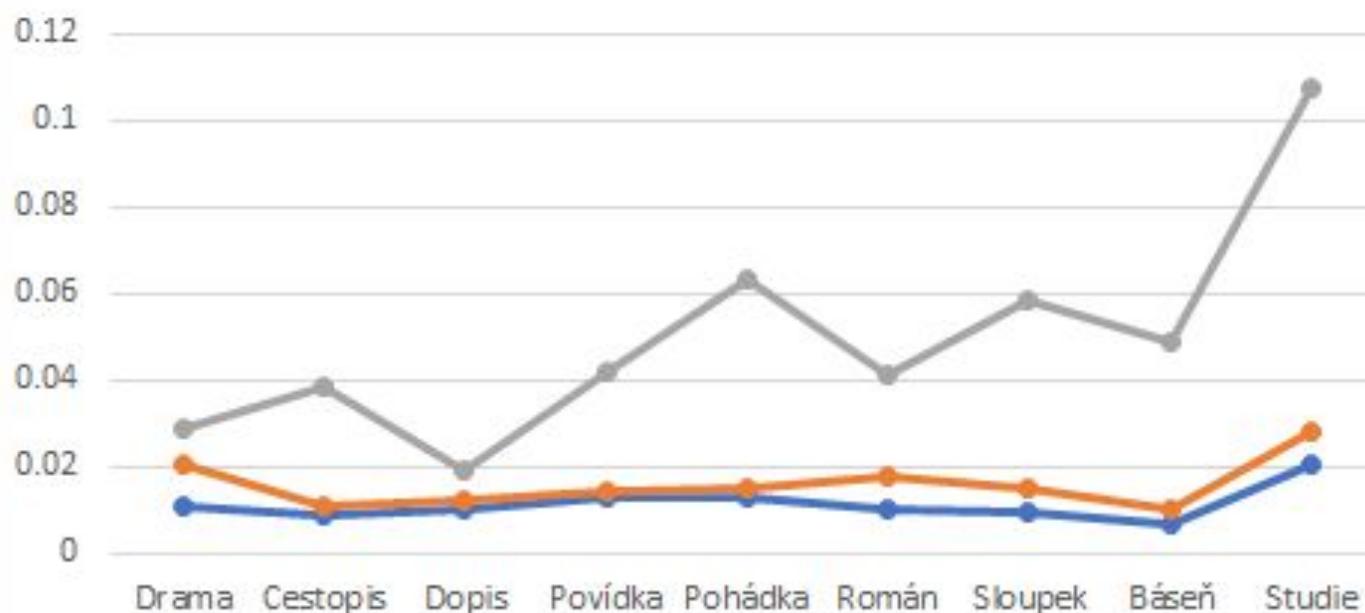
MATTR

korpus knihy části



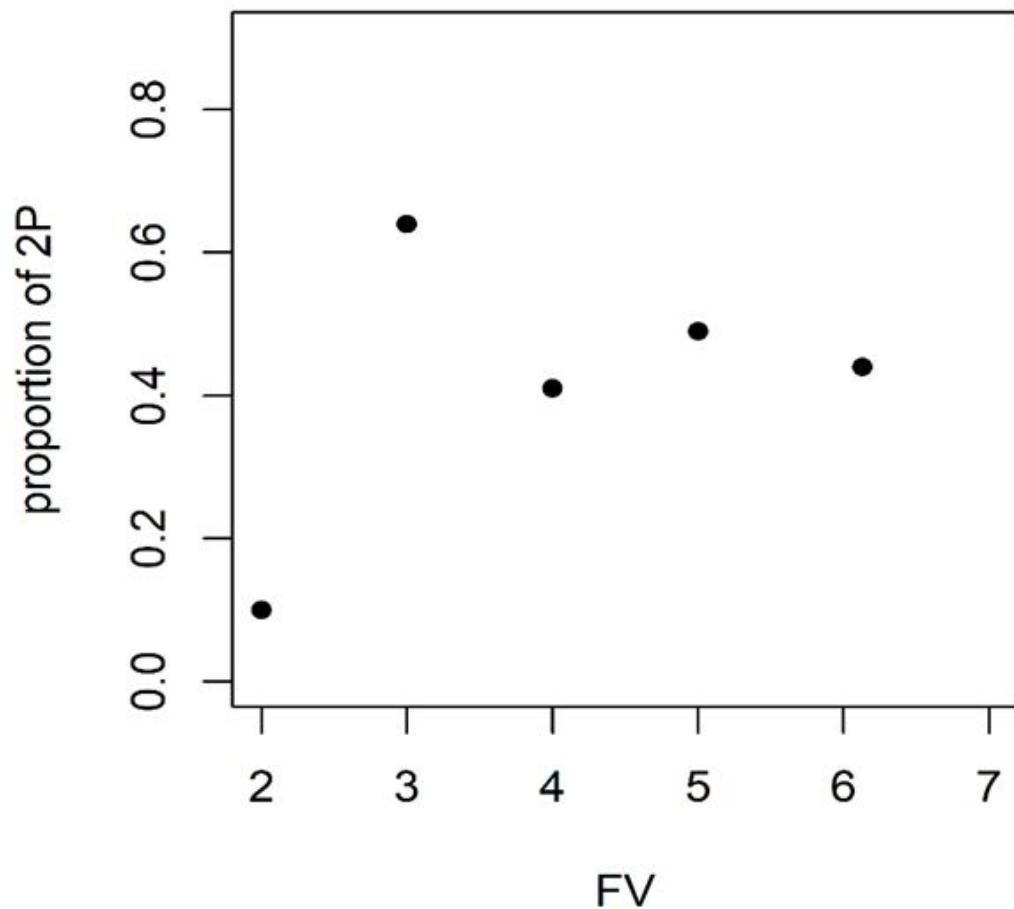
STC

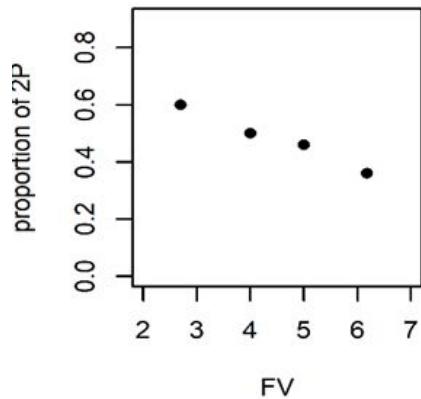
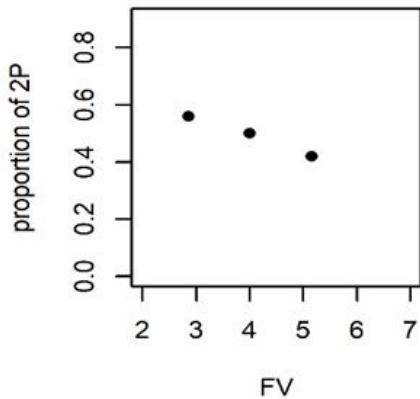
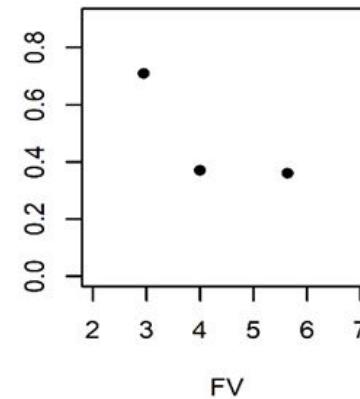
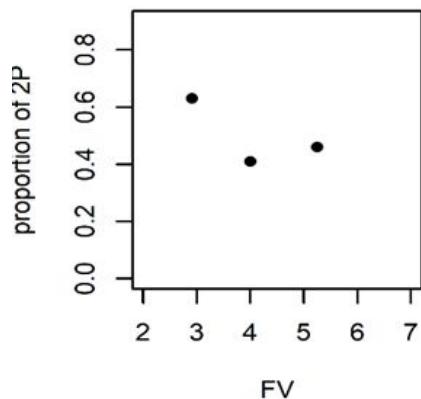
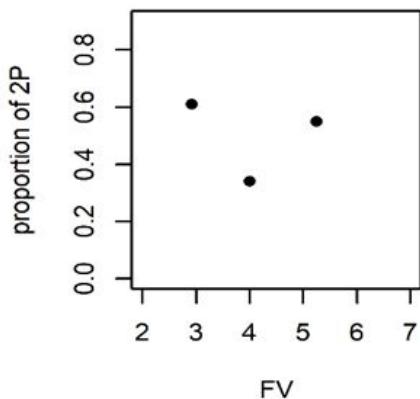
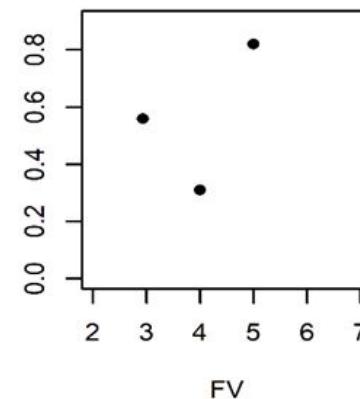
žánry knihy části



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)



Act**Lk****Mt****Gen****Job****Ecc**

Case study no. 3 - the Menzerath-Altmann law

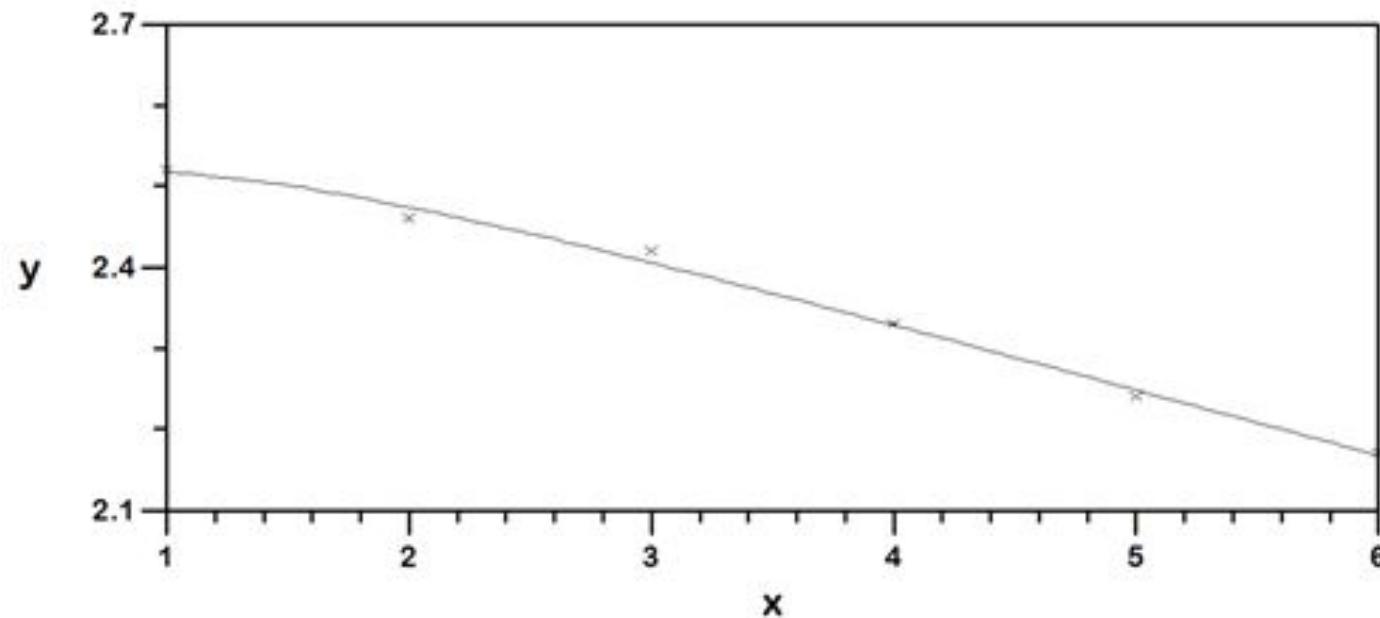
- word - syllable - sound

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

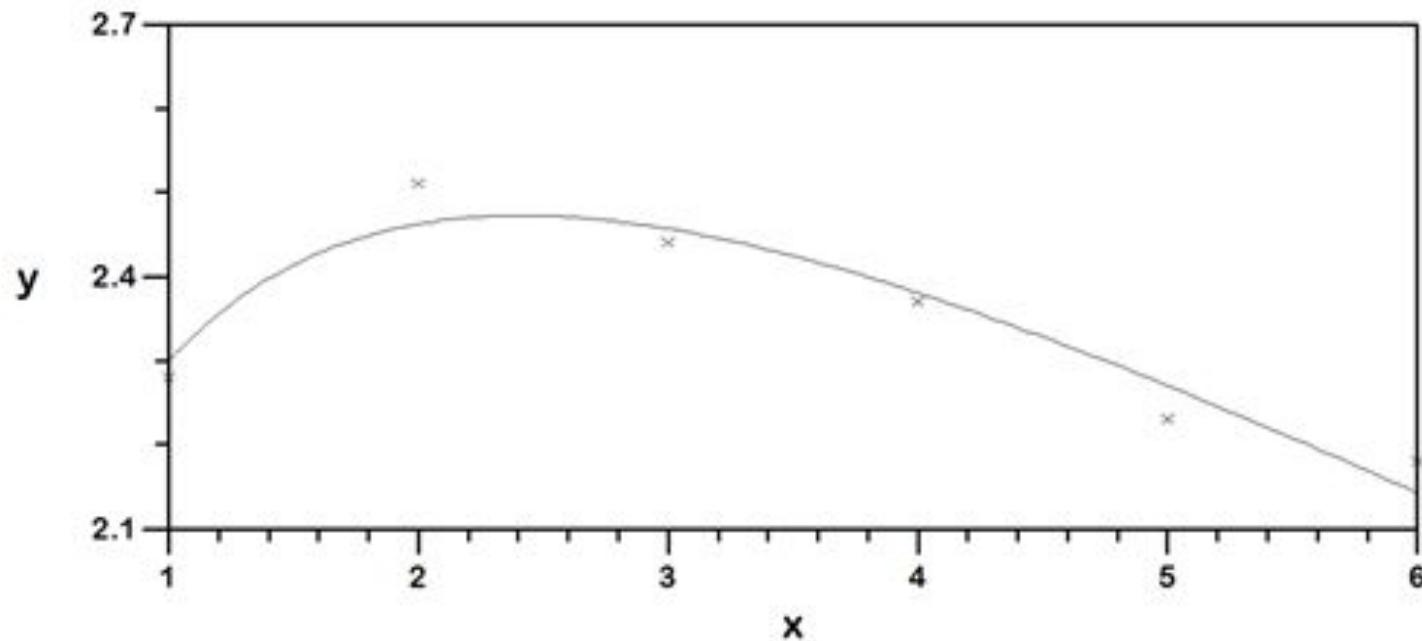
- 5 texts
 - Gottwald 1949
 - Havel 1999
 - V. Cvrček: Za ještě tvrdší kodifikační diktát? (Naše řeč, 2006, s. 26-29)
 - K. J. Erben: Zlatý kolovrat
 - 10 years child: Já a první žárovka

| | 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 |
| R² | 0.99 | 0.95 | 0.91 | 0.93 | 0.94 | 0.95 |

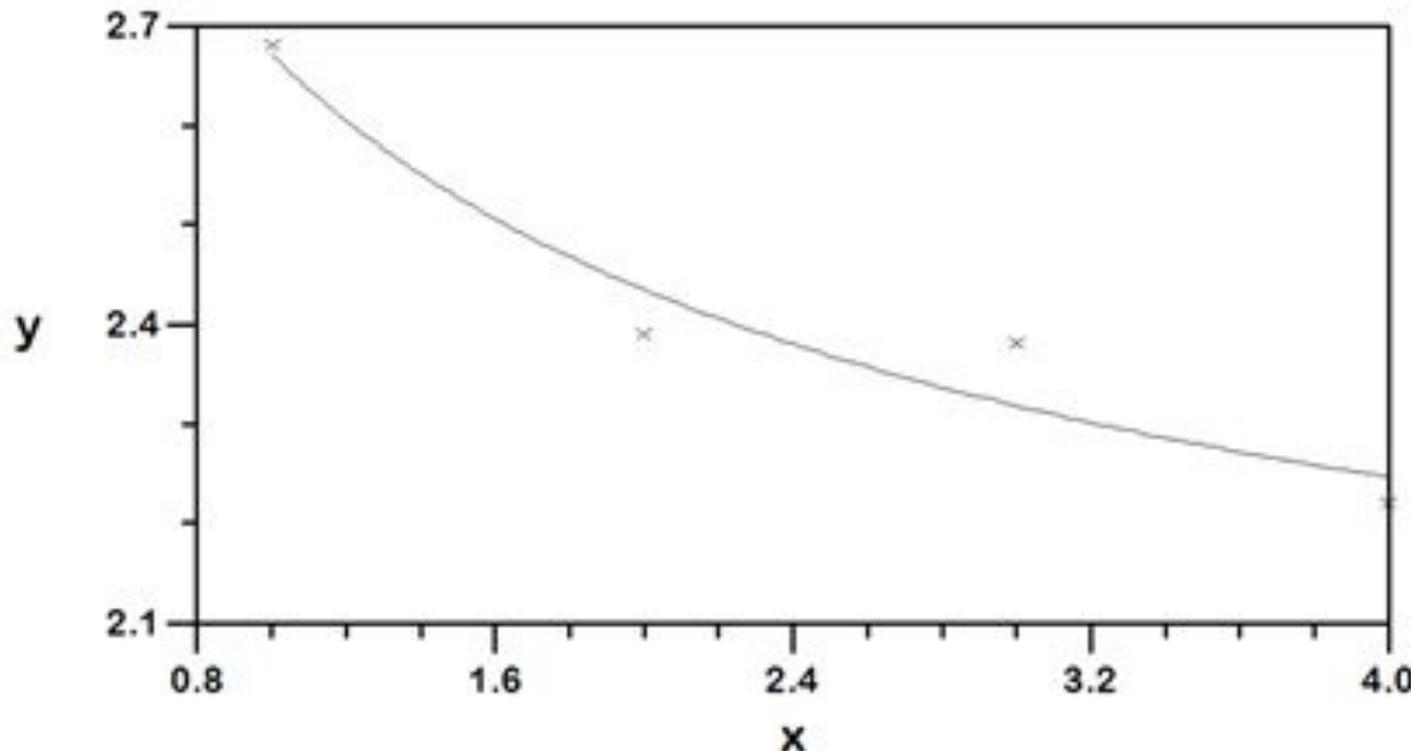
All texts together



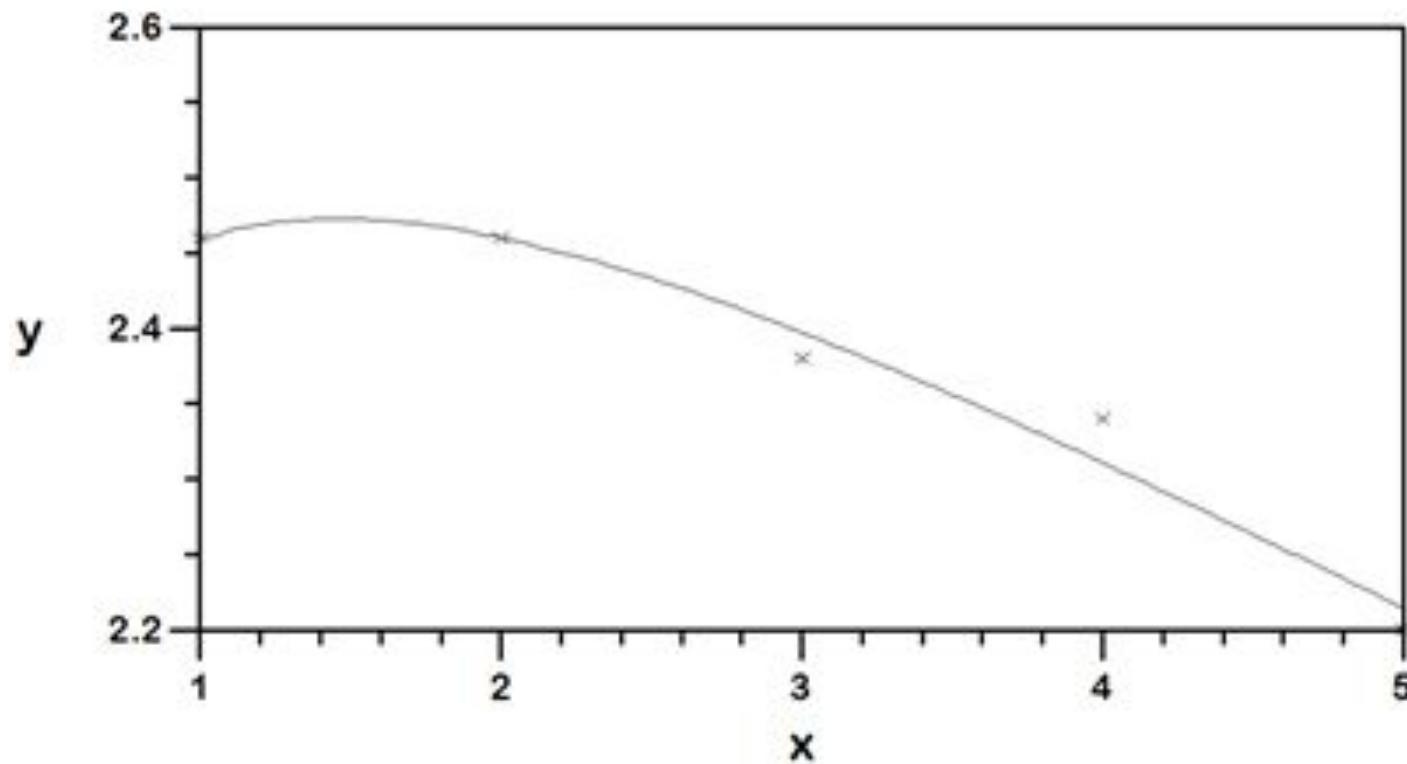
V. Cvrček: Za ještě tvrdší kodifikační diktát?



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

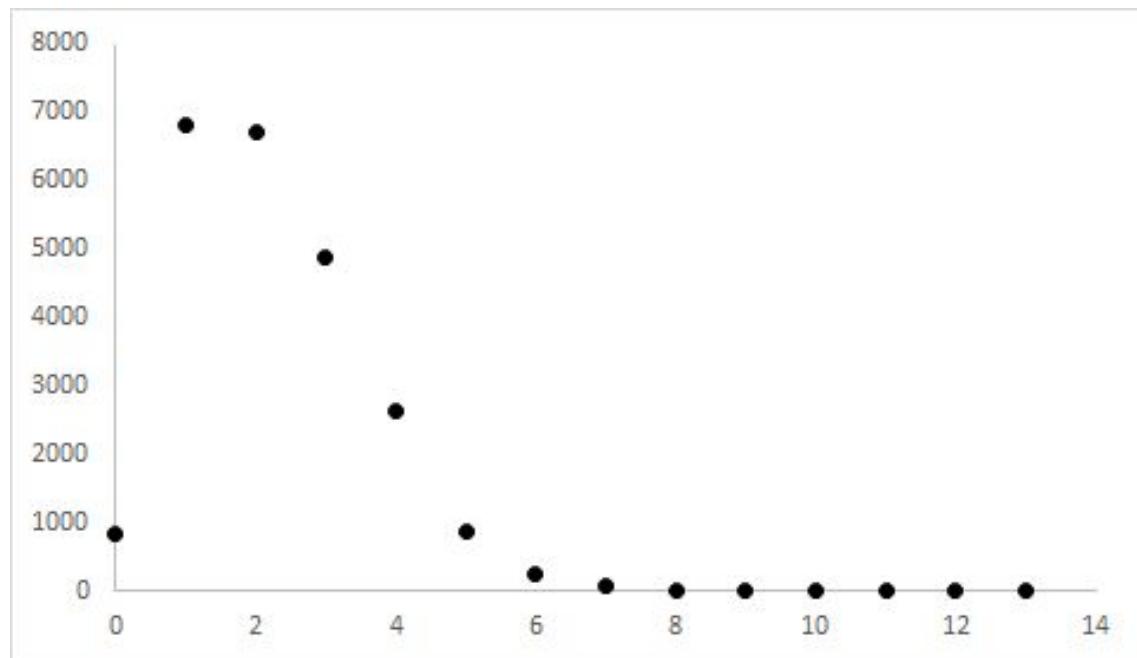
All texts together

hyper-Poisson distribution

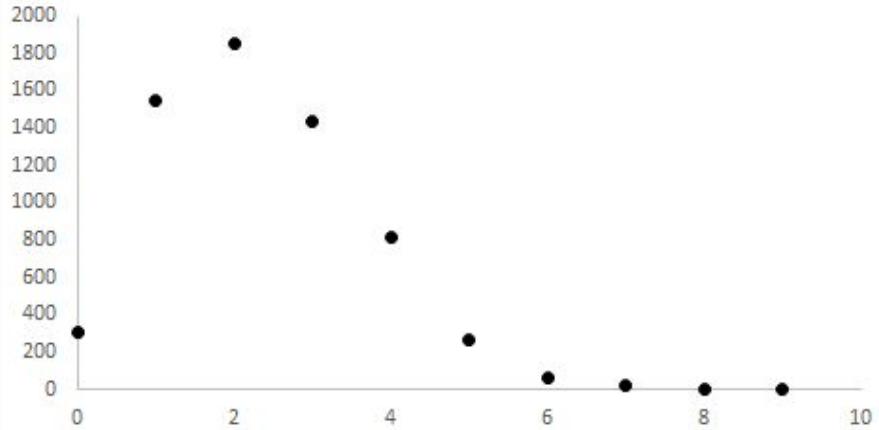
$$a = 1.46$$

$$b = 0.18$$

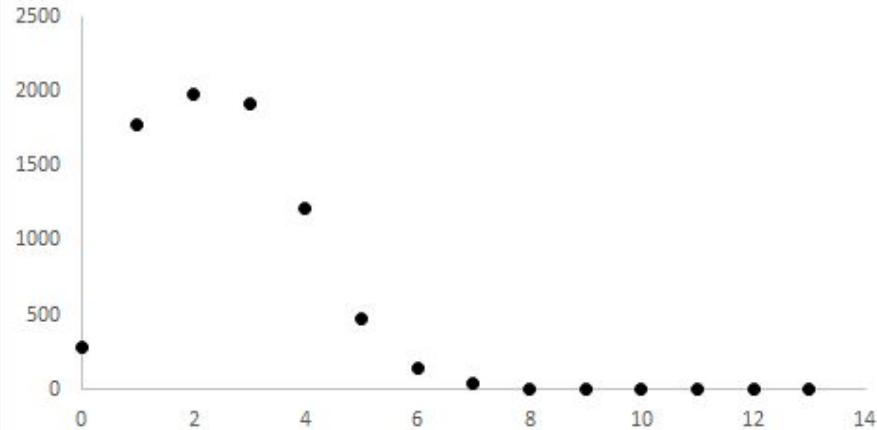
$$C = 0.012$$



journalism



scientific



a b

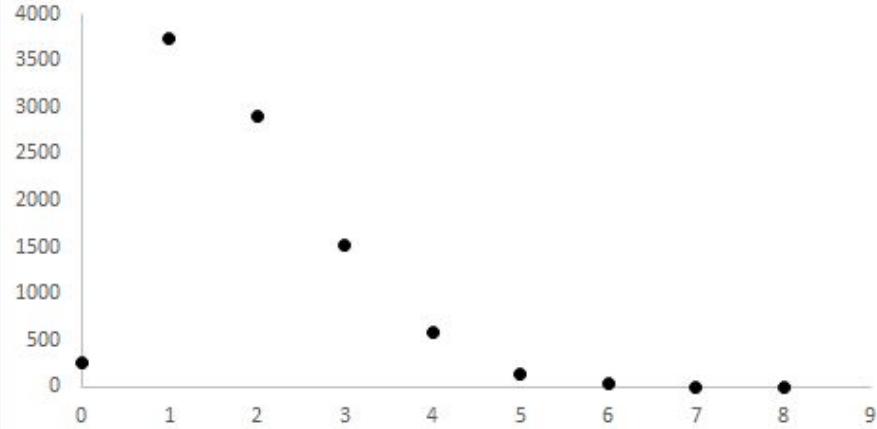
C

journalism 1.703 0.339 0.005

scientific 1.904 0.335 0.012

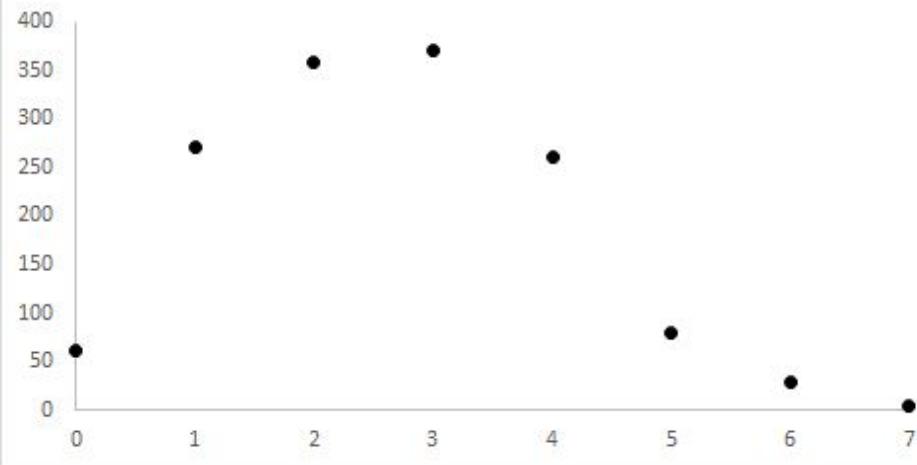
fiction 0.982 0.068 0.008

fiction

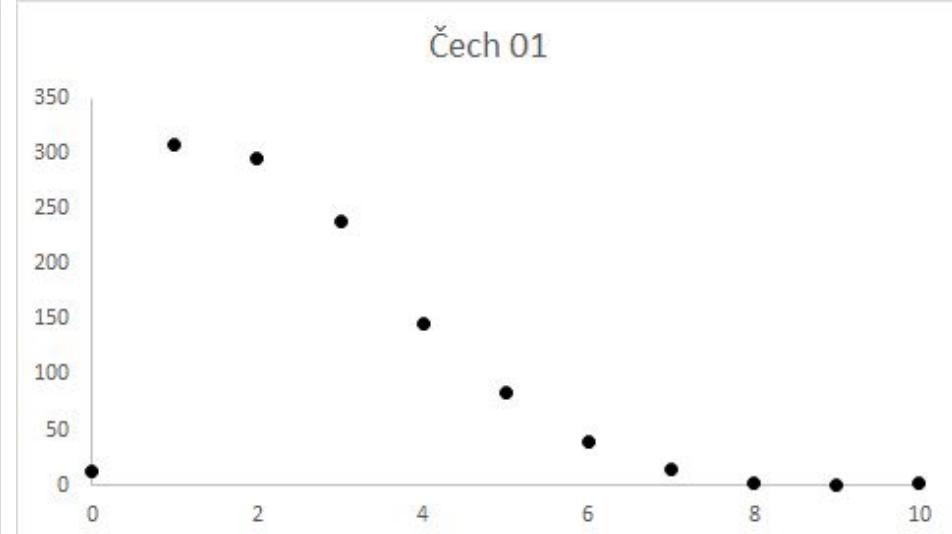


Scientific

Hrbáček 01



Čech 01



Conclusion

Conclusion

