Menzerath-Altmannův zákon: Náhodný model

Jiří Milička

O čem to mělo být původně...

- Benfords Law a co vůbec považujeme za přijatelný model Menzrathova vztahu.
- Důvod, proč máme tak rádi Altmannův model.
- Navržení dalších modelů...

Ale pak mě napadlo, že bych mohl vyzkoušet jeden nápad z terasy nad Moselou...

Nejprve ale k tomu Benfordovi

Benford's Law

- Krásný příklad principu, který platí pro naprosto nesouvisející fenomény.
- Svým způsobem až děsivý.
- Vezměte všechna čísla v korpusu a udělejte si distribuci prvních číslic. Ona distribuce bude pro všechny korpusy přibližně stejná
- A nejen pro korpusy, ale i pro délky řek, počty lidí ve městech, fyzikální konstanty...

Benford's Law – vzorec

$$y = \log_{10}\left(1 + \frac{1}{x}\right)$$

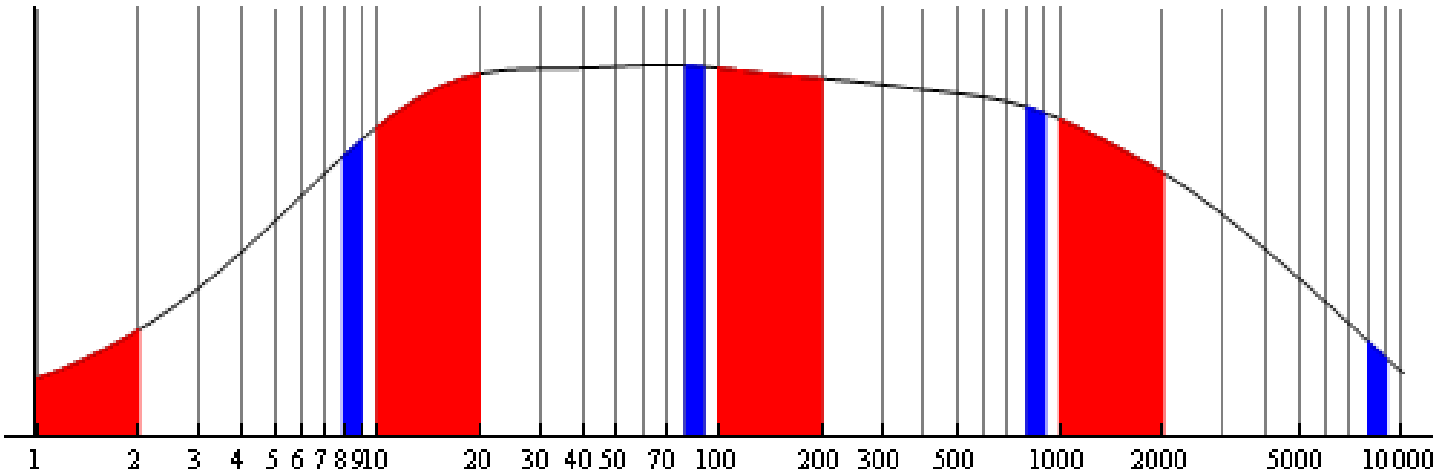
číslo jedna se na první pozici vyskytuje cca ze 30 %, číslo 2 ze 17 % atd. až číslo 9 z cca 5 %

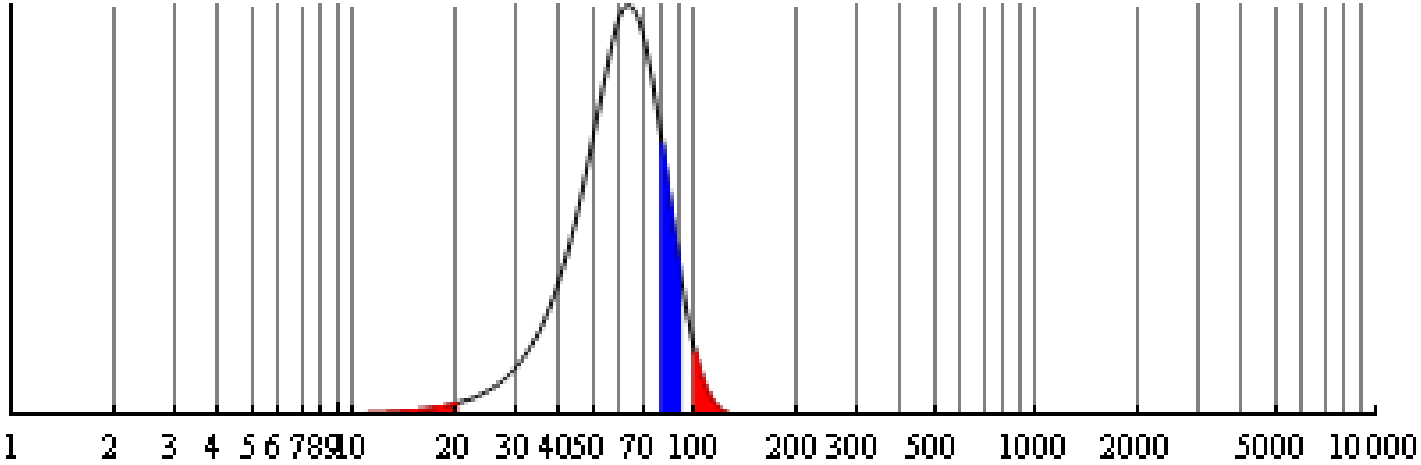
- ROSS, KENNETH A. Benford's Law, a growth industry. *American Mathematical Monthly* 118.7 (2011): 571-583.

„This article is intended for four audiences. The discussion up to Figure 1 could be used to explain Benford's law to people with essentially no mathematical background, while the discussion up to the proof of Theorem 1, just after Table 5, could be used to explain the law to any motivated person who understands logarithms. Upper-division undergraduate students, willing to assume the uniform distribution theorem, should be able to read the proof of Theorem 1. The remaining proofs involve a little measure theory and harmonic analysis. “

(kéž by takto psal články ŁUKASZ DĘBOWSKI, intersubjektivita se dnes podceňuje, ale je strašně důležitá...)

Benford's Law: Explanace





- Tedy platí tam, kde jsou data rovnoměrně (nebo aspoň přibližně rovnoměrně) rozložená přes více řádů (čím rovnoměrněji a čím víc řádů, tím líp)
- Můžete si dataset, který splňuje BL vytvořit tak, že vygenerujete dostatek čísel y, kde

$$y = 10^{\text{náhodné číslo z intervalu}<a,b>}$$

- Z toho plyne, že se zákonem budou řídit čísla z která vyrostla exponenciálně (například města, čím je město větší, tím rychleji roste).
- nejspíš to souvisí i se Zipfovým zákonem.

No, a chovám představu, že podobně obecný princip může stát i za Menzerathovým vztahem...

Menzerath-Altmann

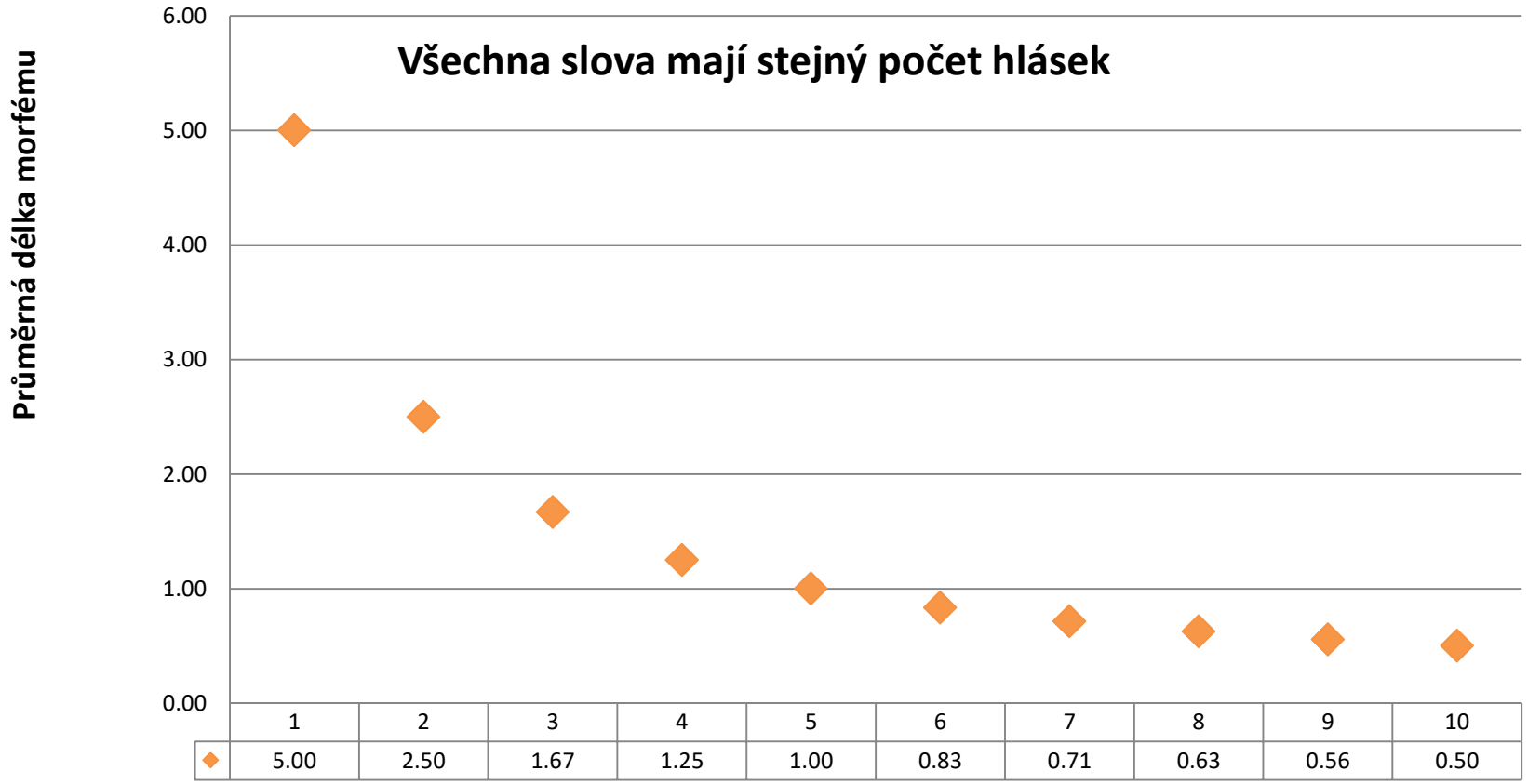
(konečně)

- Menzerath-Altmannův zákon počítá s tím, že průměrná délka konstituentů v daném konstruktu je závislá na počtu konstituentů v daném konstruktu.
- tedy například průměrný počet hlásek v morfech v daném slově je nepřímo úměrný počtu morfů v daném slově.
- Cílem tohoto příspěvku je přesvědčit vás, že ona všem dobře známá křivka neznamená, že počet hlásek v morfu nějak magicky ovlivňuje počet morfů ve slově nebo obráceně.
- Tedy, může to tak být, ale pro vytvoření slavného vztahu $y=Ax^{-b}$ to není třeba.

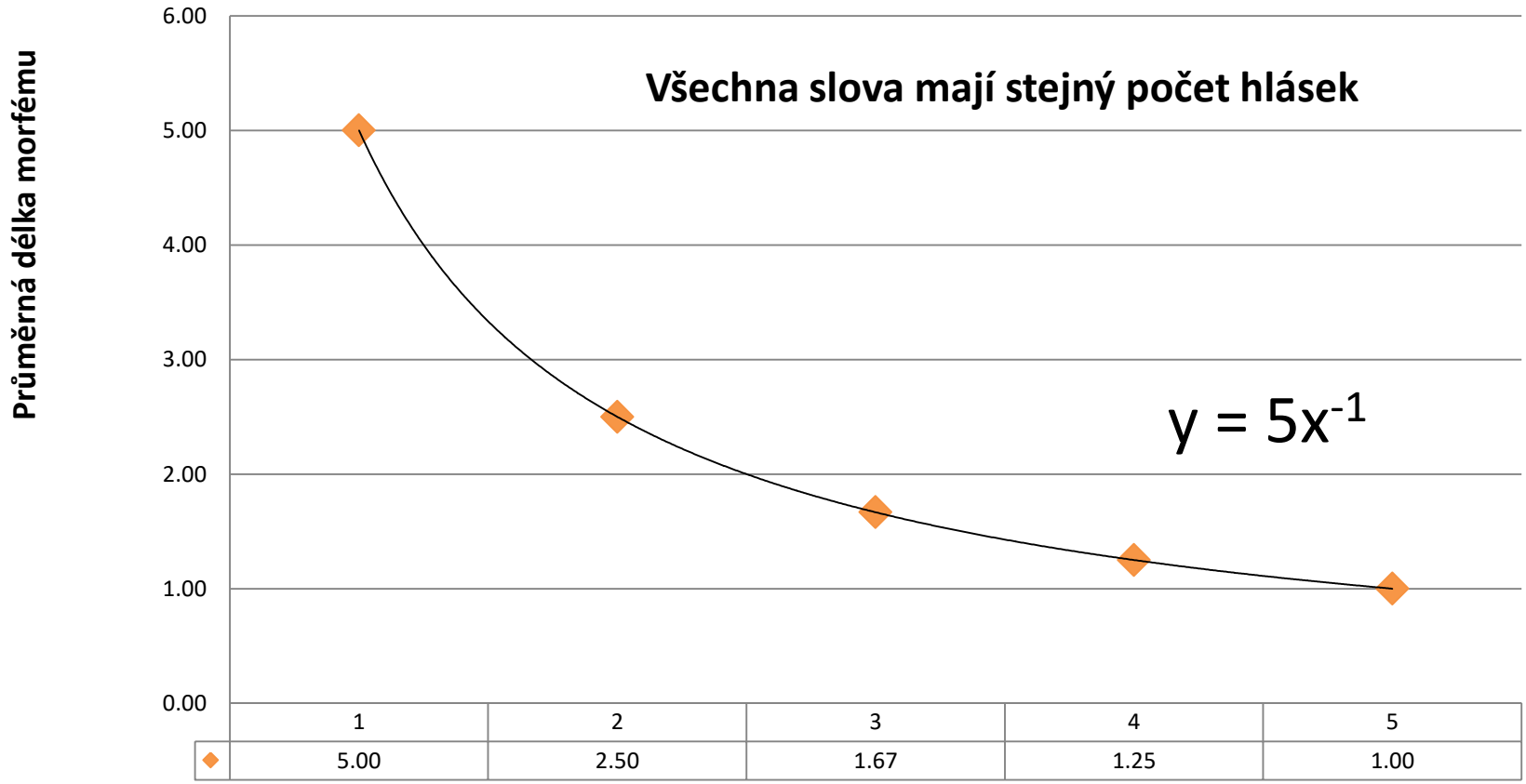
Příklad na začátek...

- Úplné omezení variability délky konstruktu (např. všechna slova mají přesně 5 hlásek)
- Oproti tomu konstituenty variabilitu omezenou nemají (mohou mít jednu až pět hlásek)
- **Nedefinujeme žádnou závislost mezi počtem hlásek v morfému a počtem morfémů ve slově**.
- Tipněte si, jak bude vypadat výsledek

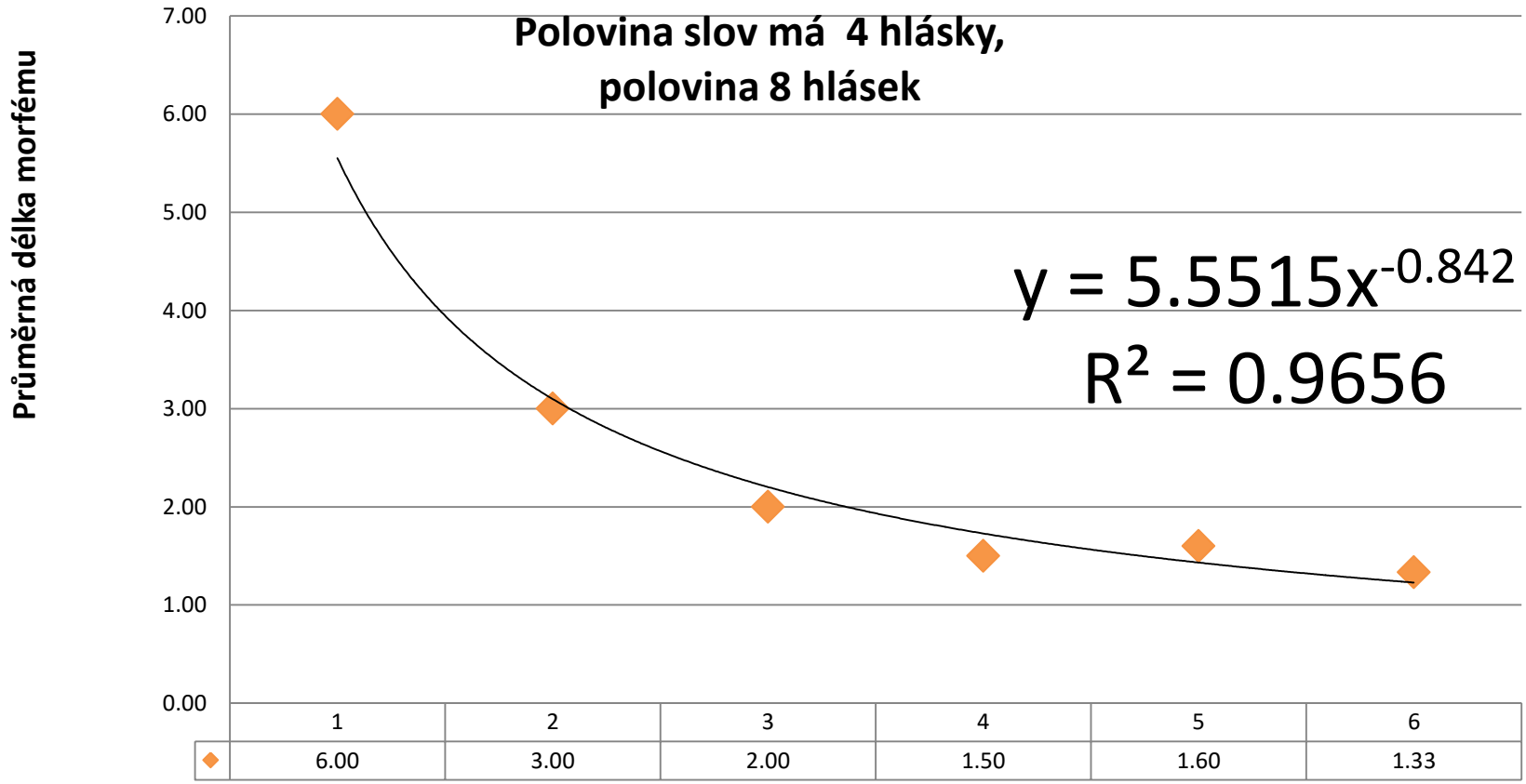
Úplné omezení variability délky konstruktu



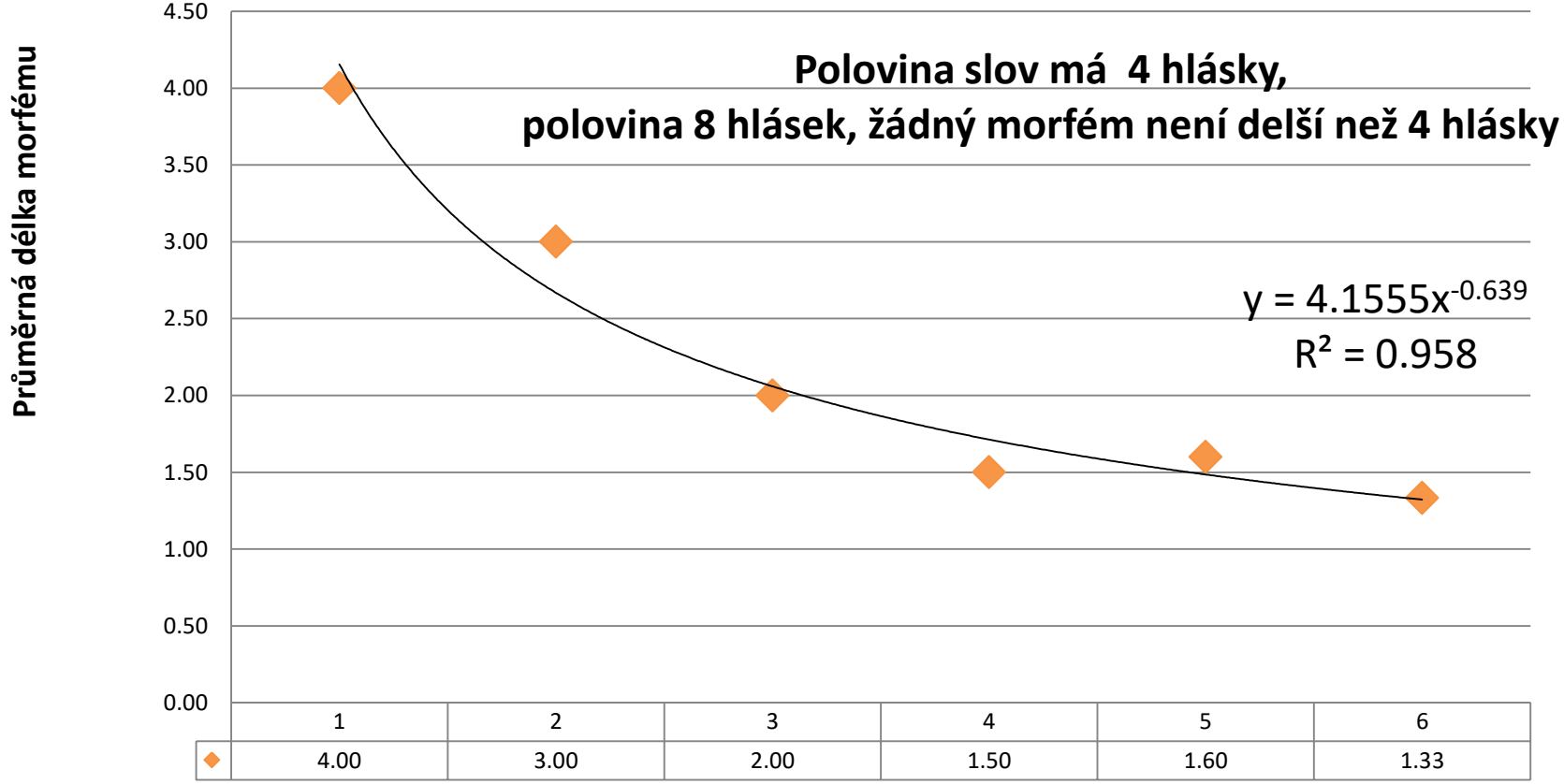
Úplné omezení variability délky konstruktu



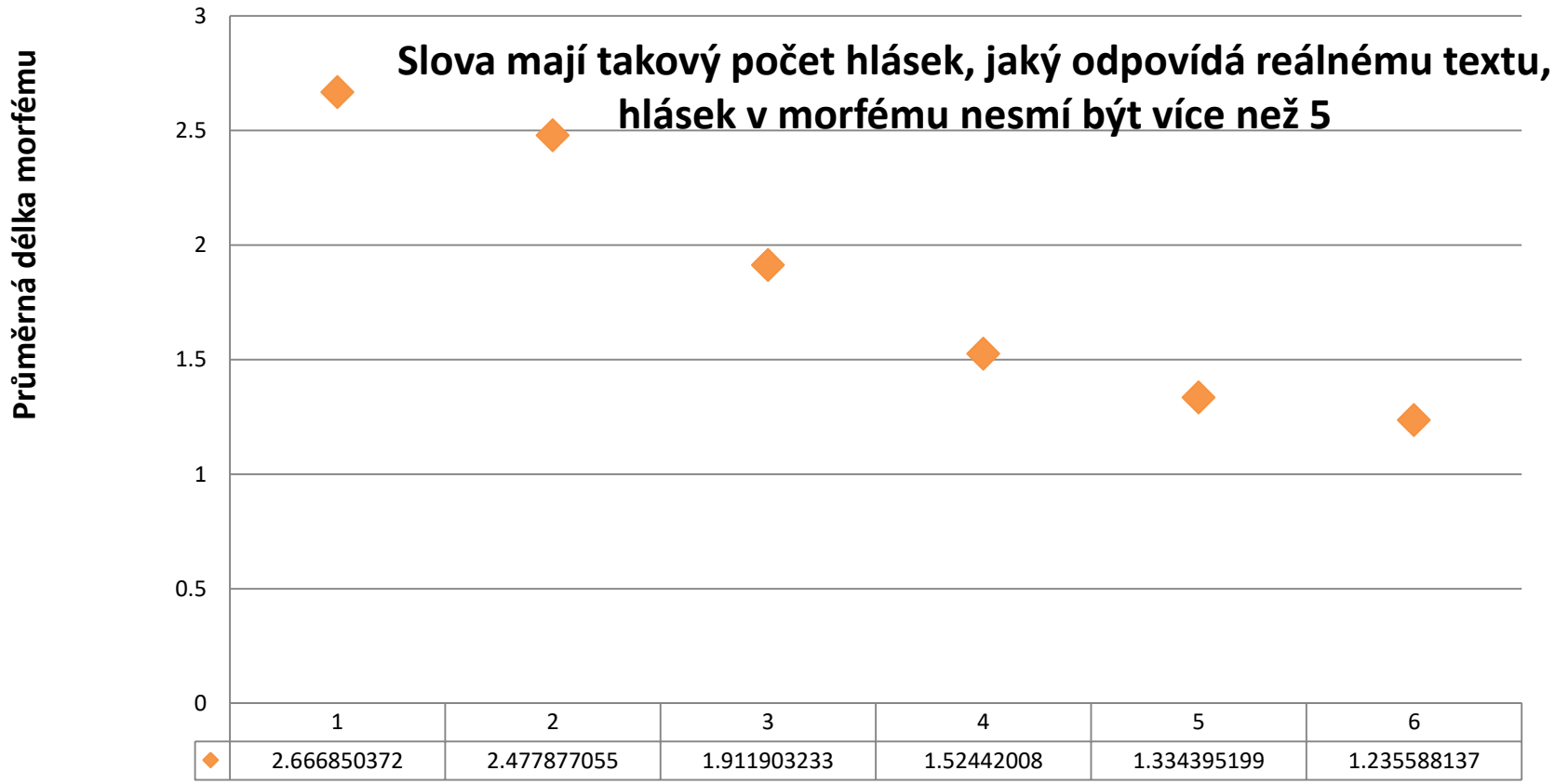
Částečné omezení variability délky konstruktu

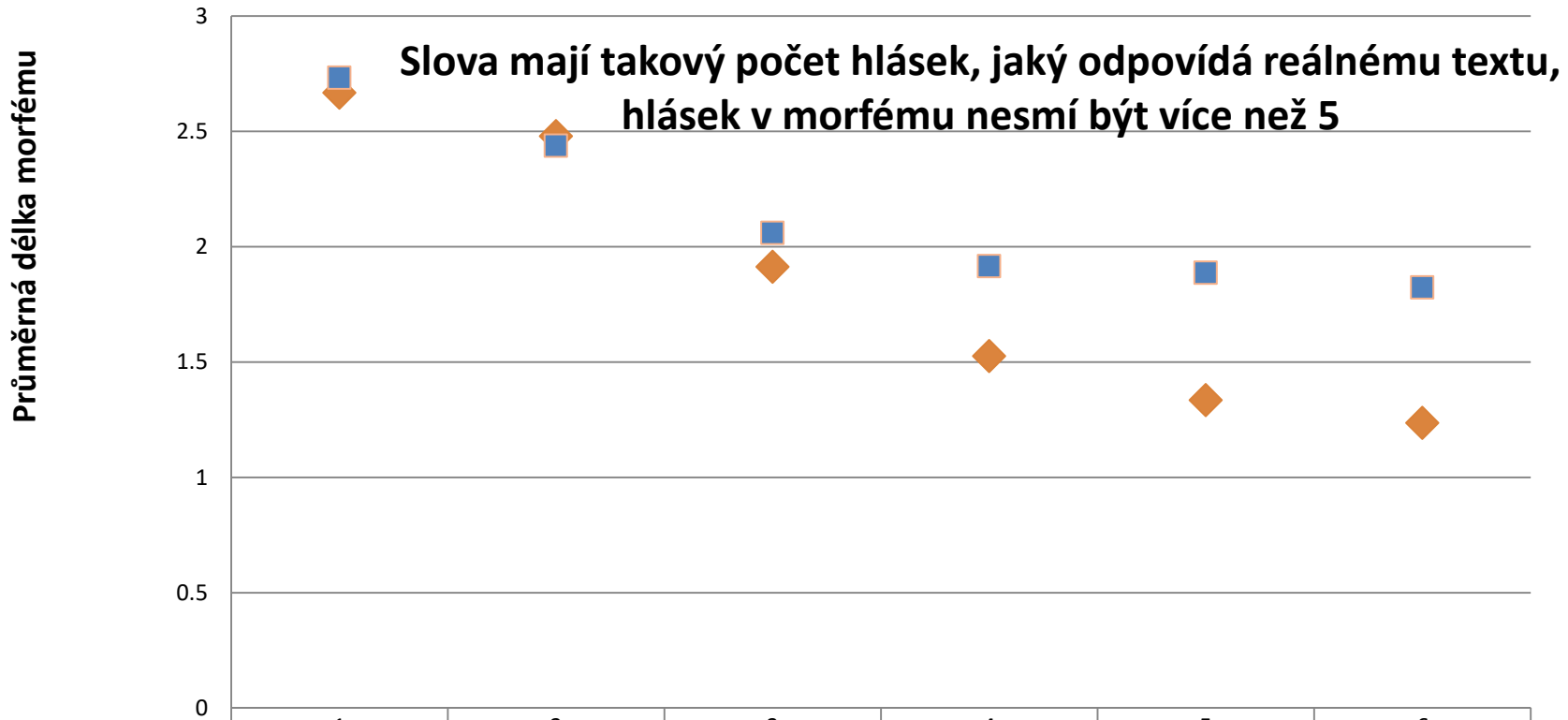


Částečné omezení variability délky konstruktu + částečné omezení variability délky konstituentu



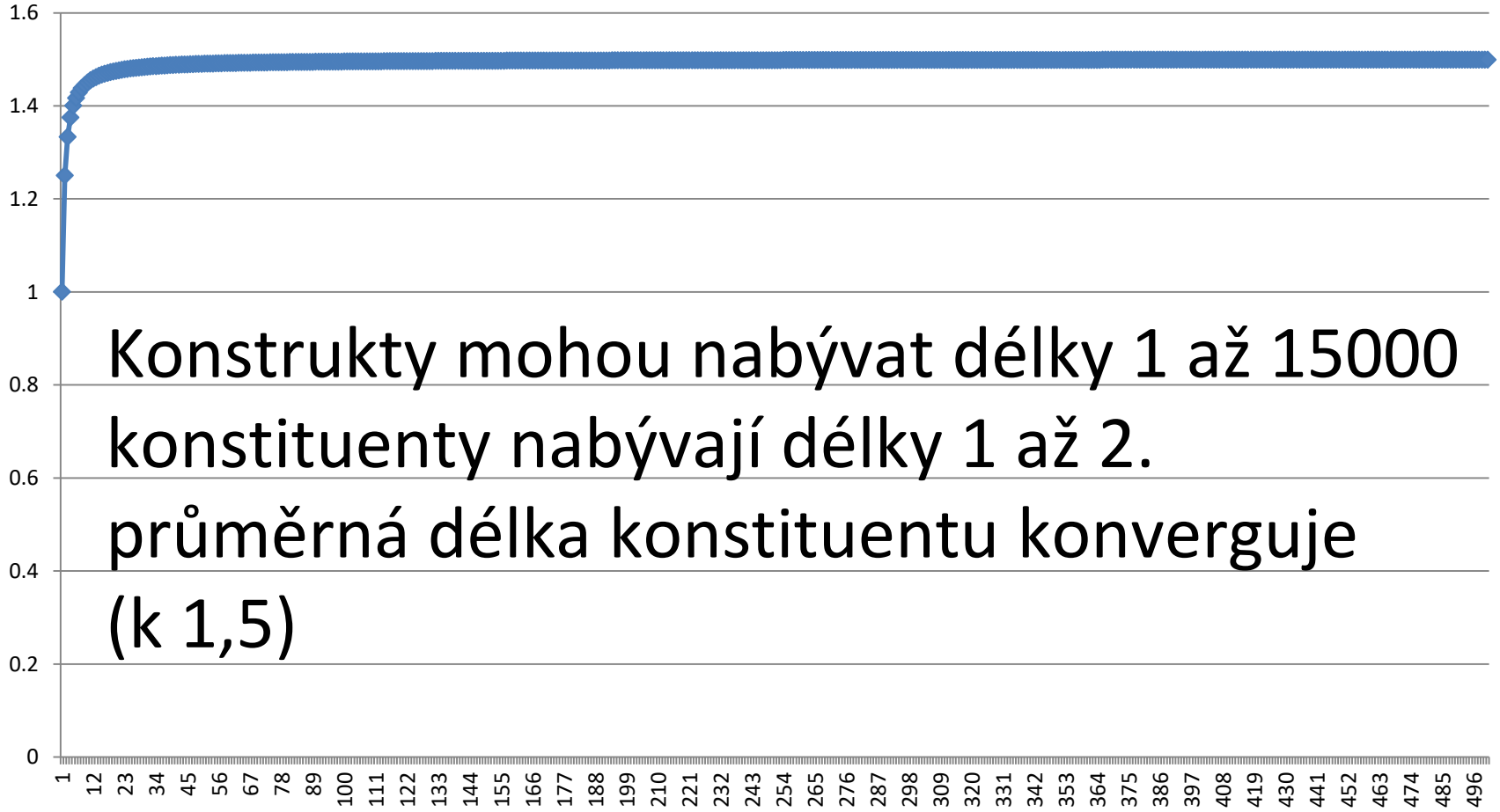
Podle distribuce v reálném textu





	1	2	3	4	5	6
◆	2.666850372	2.477877055	1.911903233	1.52442008	1.334395199	1.235588137
■ orig	2.732	2.437	2.059	1.915	1.886	1.822

Velká variabilita konstruktu omezená variabilita konstituentu



Konstrukty mohou nabývat délky 1 až 15000

konstituenty nabývají délky 1 až 2.

průměrná délka konstituentu konverguje (k 1,5)

Randomizace

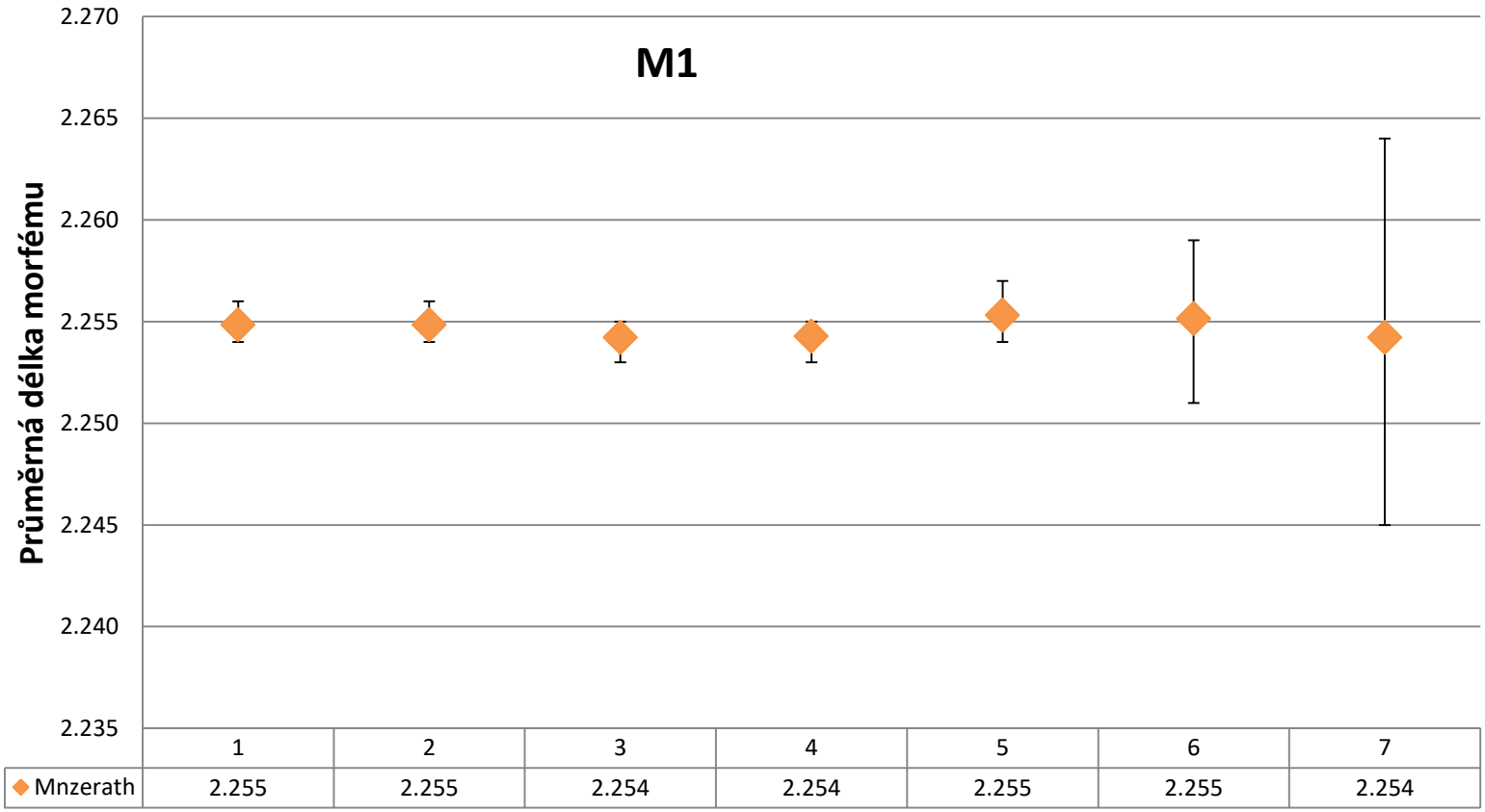
- Chceme do modelu zahrnout jak distribuci konstruktů tak distribuci konstituentů, použijeme náhodný model
- BENEŠOVÁ, M., ČECH, R. (2015). Menzerath-Altmann law versus random model. In Mikros, G. K., Mačutek, J. (eds.) *Sequences in language and text*. De Gruyter, 57-70.

M1 (Benešová, M., Čech, R.)

- Vezmeme balík všech vět v textu.
- Vezmeme balík všech souvětí ve stejném textu.
- Pro každé věty v souvětí náhodně vybereme jinou větu z balíku vět.

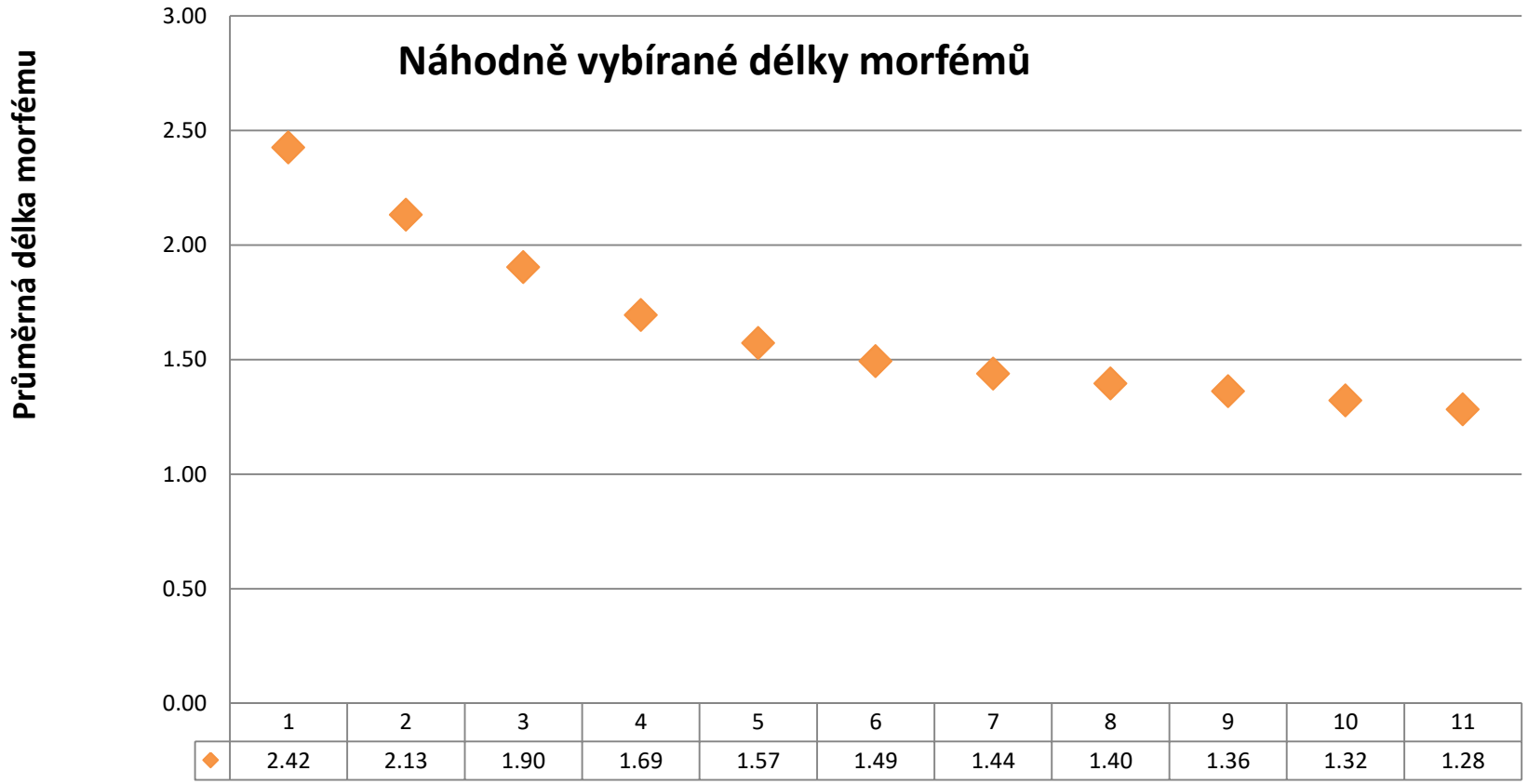
Neuvěříte, co se stalo pak.

Radkův / Martinin model 1000x

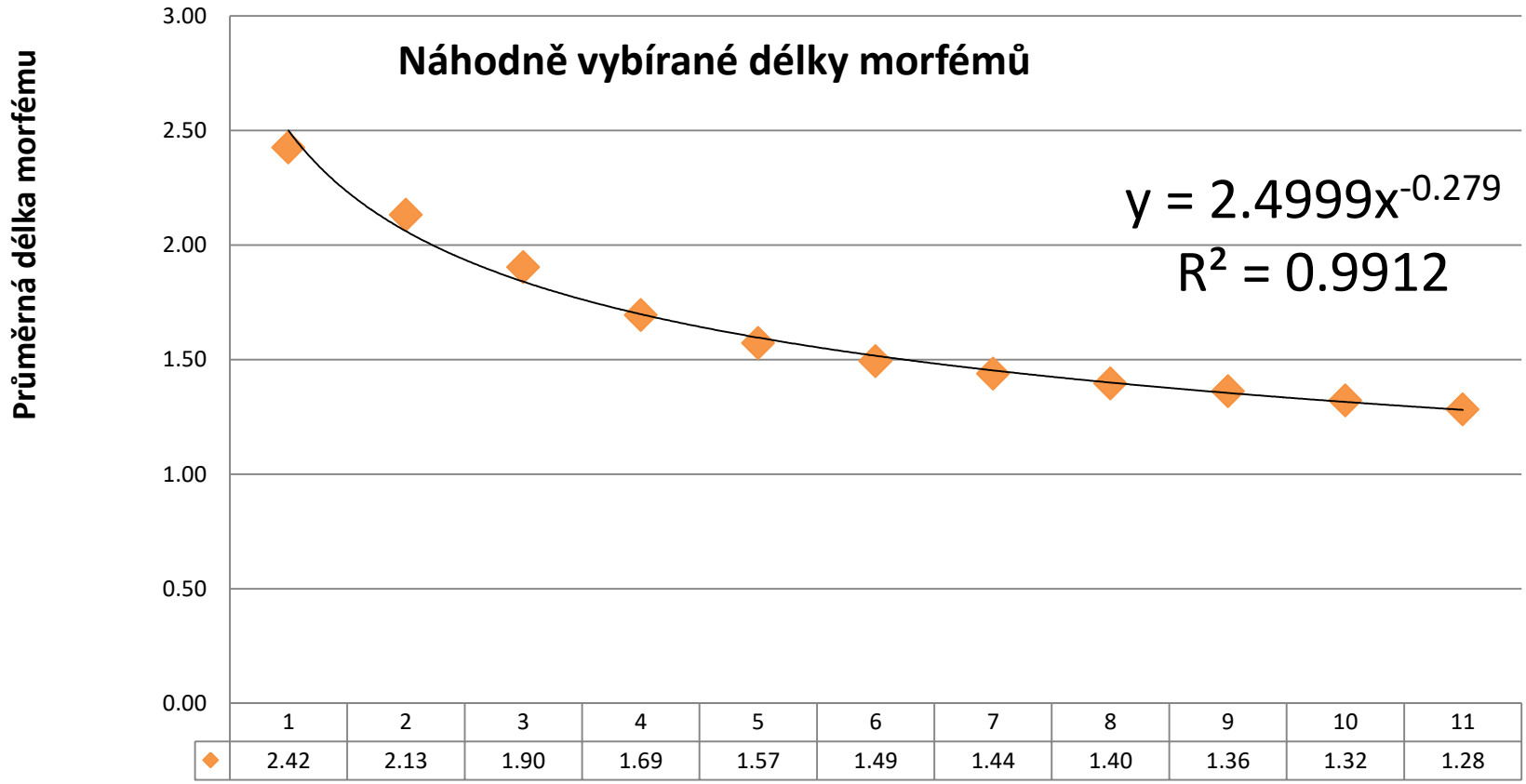


- (provedl jsem 1000x, aby byly vidět conf. int.)
- Definovali si to jako dvě nezávislé proměnné
- pokud by se došlo k jakémukoliv jinému výsledku, tak by to znamenalo, že můžeme zahodit všechno co víme o pravděpodobnosti
- Ještě další dva modely, motivace k M2 nejasná, M3 (Janův) poněkud více tricky: -)

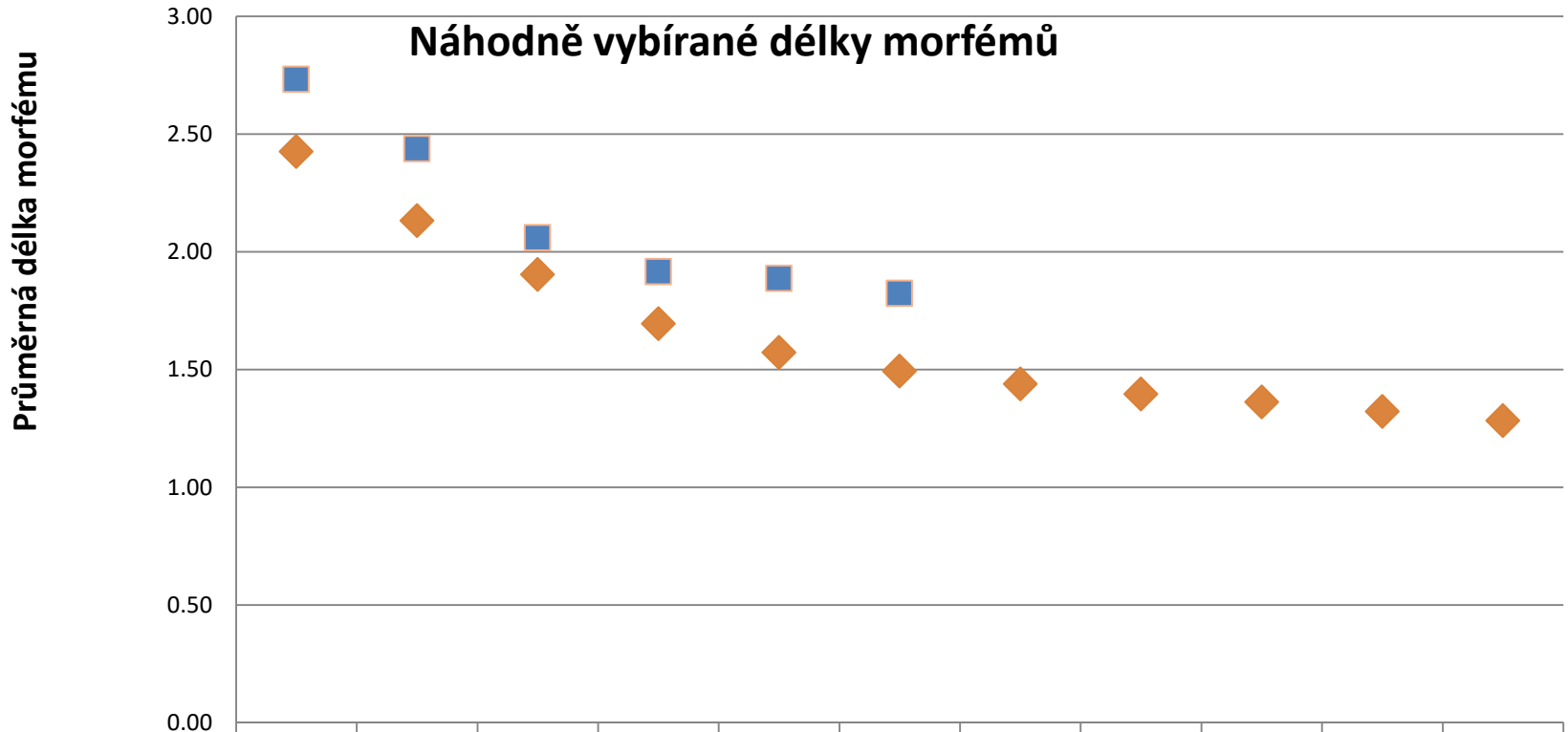
Pevný počet hlásek ve slově, náhodně doplňované morfémy podle distribuce



Pevný počet hlásek ve slově, náhodně doplňované morfémy podle distribuce

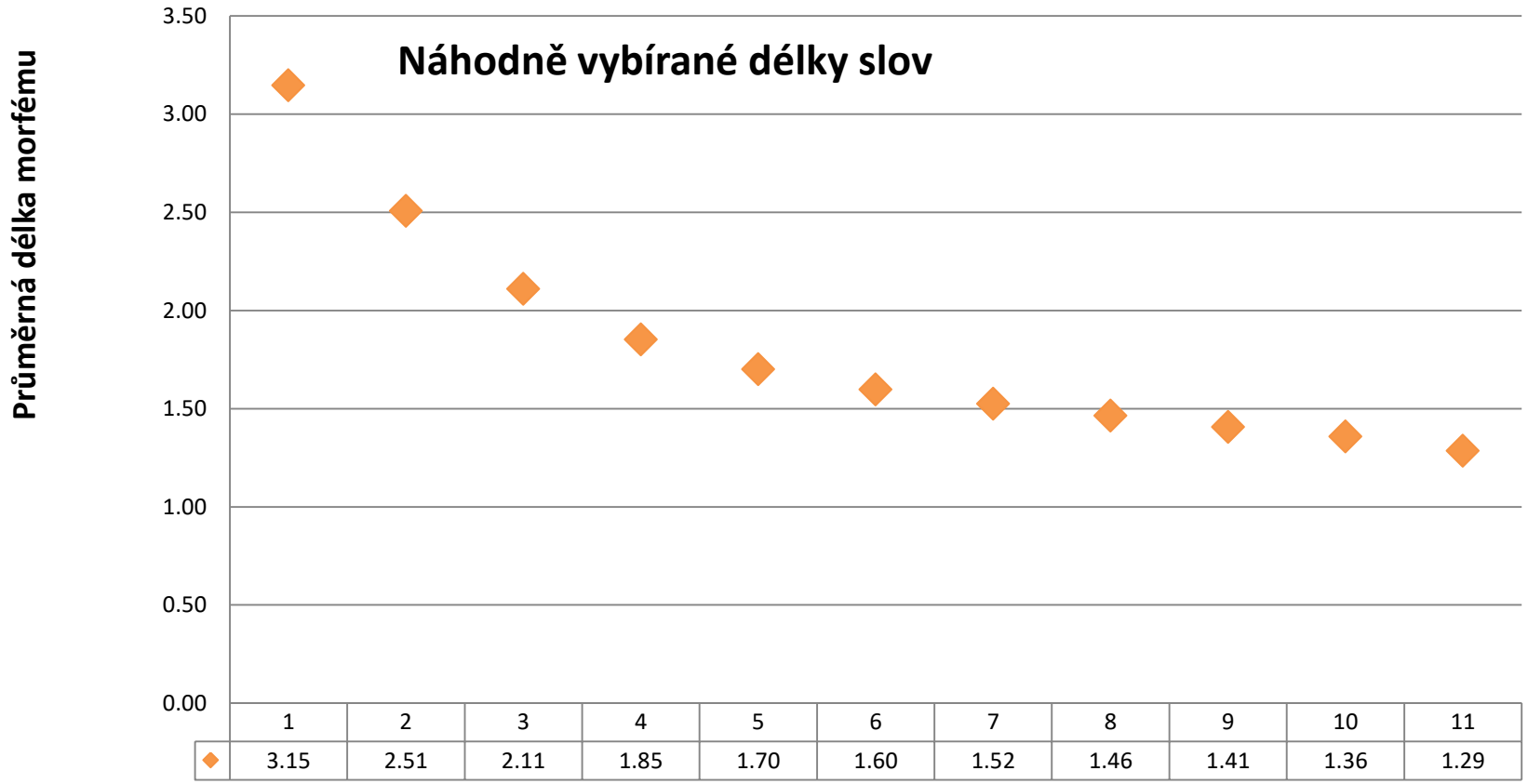


Pevný počet hlásek ve slově, náhodně doplňované morfémy podle distribuce

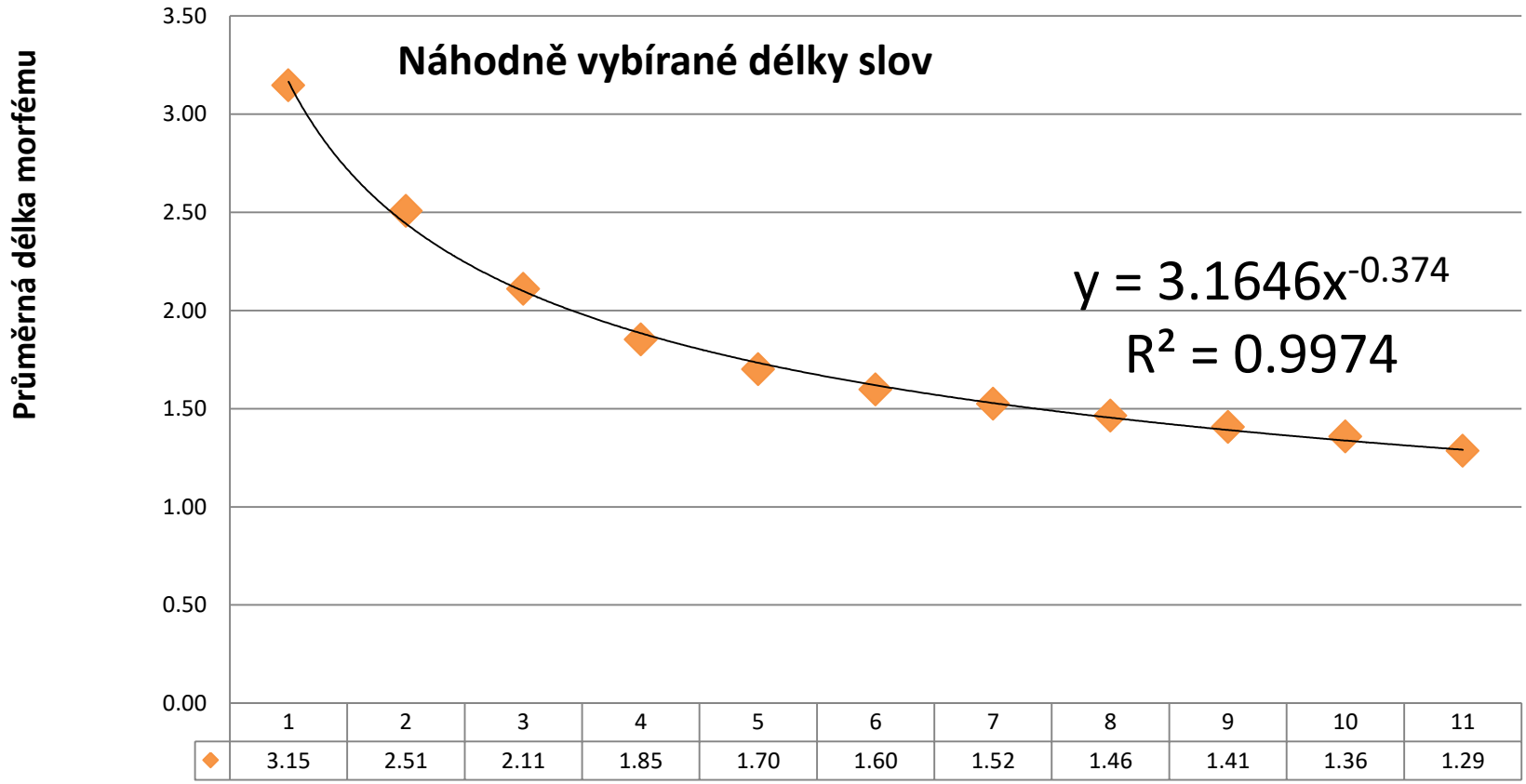


	1	2	3	4	5	6	7	8	9	10	11
◆	2.42	2.13	1.90	1.69	1.57	1.49	1.44	1.40	1.36	1.32	1.28
orig	2.732	2.437	2.059	1.915	1.886	1.822					

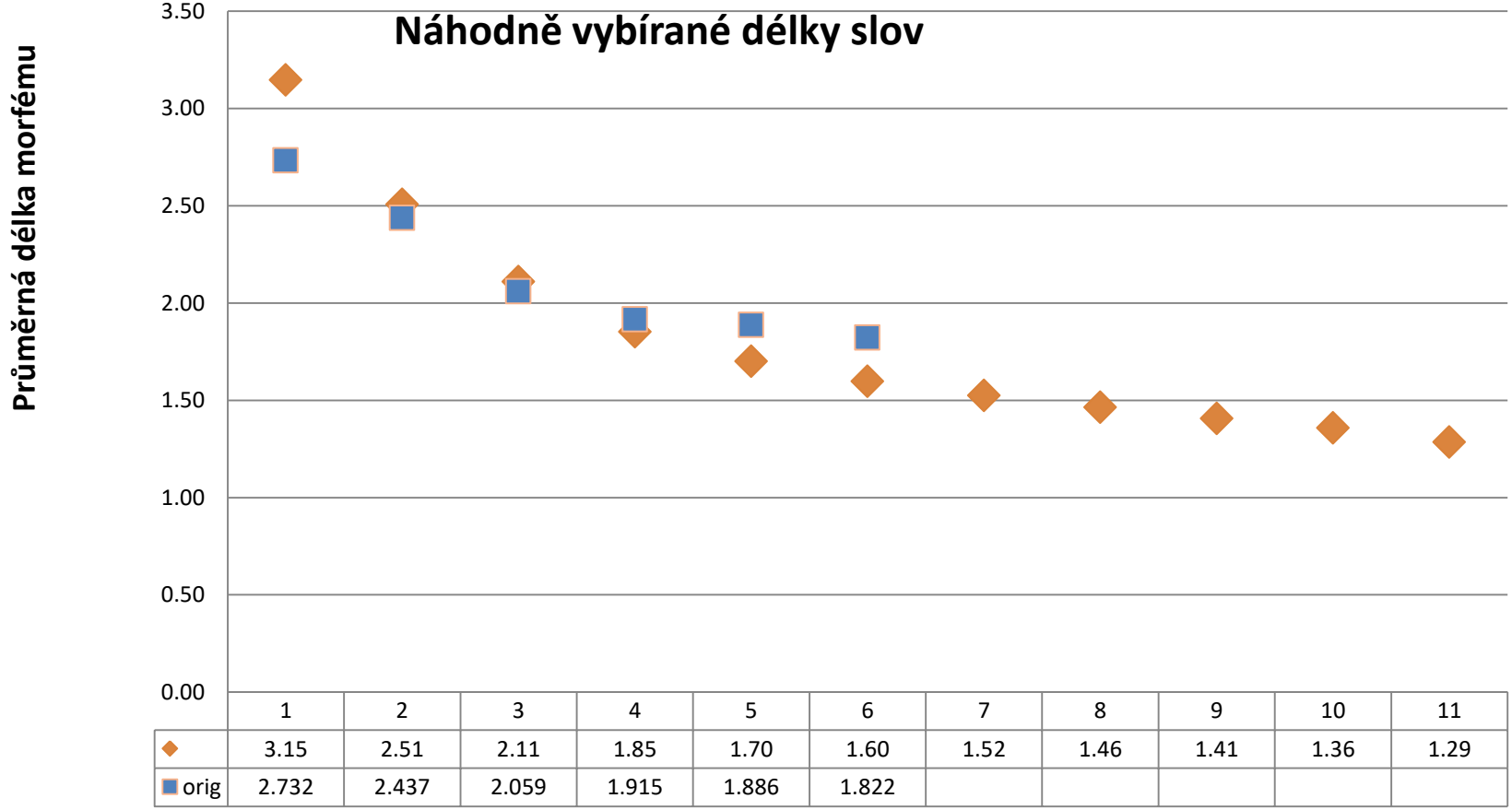
Náhodně promíchané morfy, slova náhodně vybíraná z distribuce



Náhodně promíchané morfy, slova náhodně vybíraná z distribuce

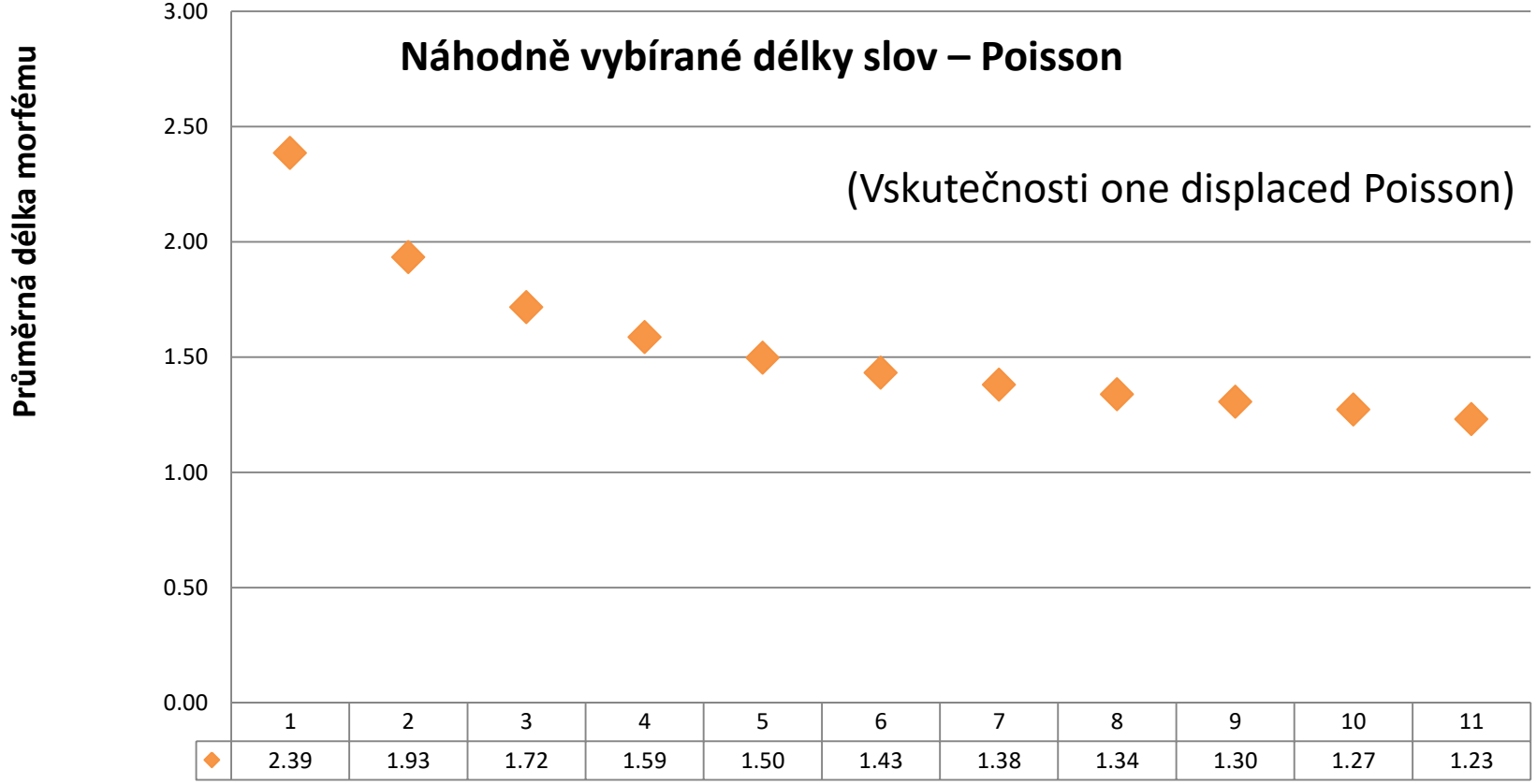


Náhodně promíchané morfy, slova náhodně vybíraná z distribuce

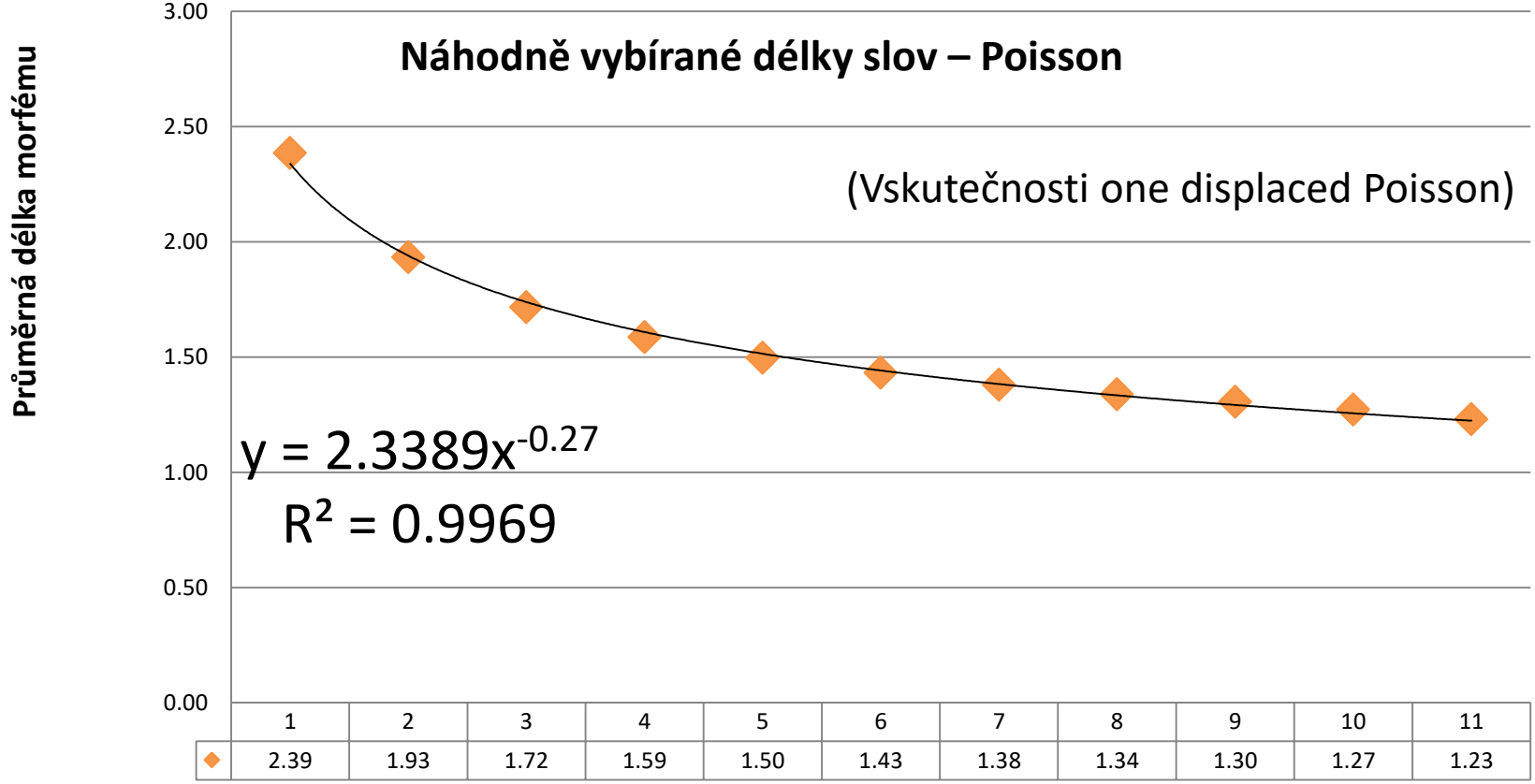


	1	2	3	4	5	6	7	8	9	10	11
◆	3.15	2.51	2.11	1.85	1.70	1.60	1.52	1.46	1.41	1.36	1.29
■ orig	2.732	2.437	2.059	1.915	1.886	1.822					

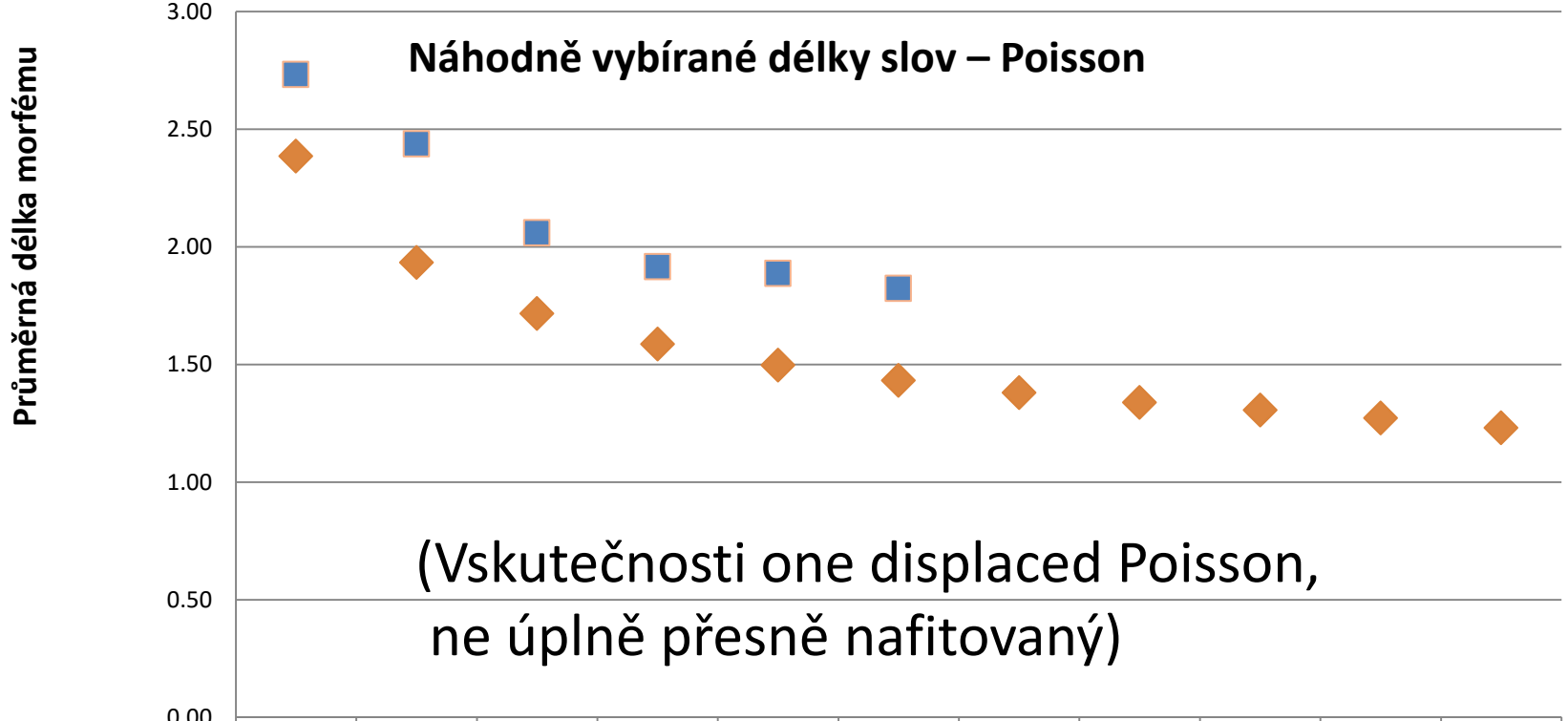
Jak distribuce délek morfů tak distribuce délek slov nahrazena Poissonovým rozdělením



Jak distribuce délek morfů tak distribuce délek slov nahrazena Poissonovým rozdělením

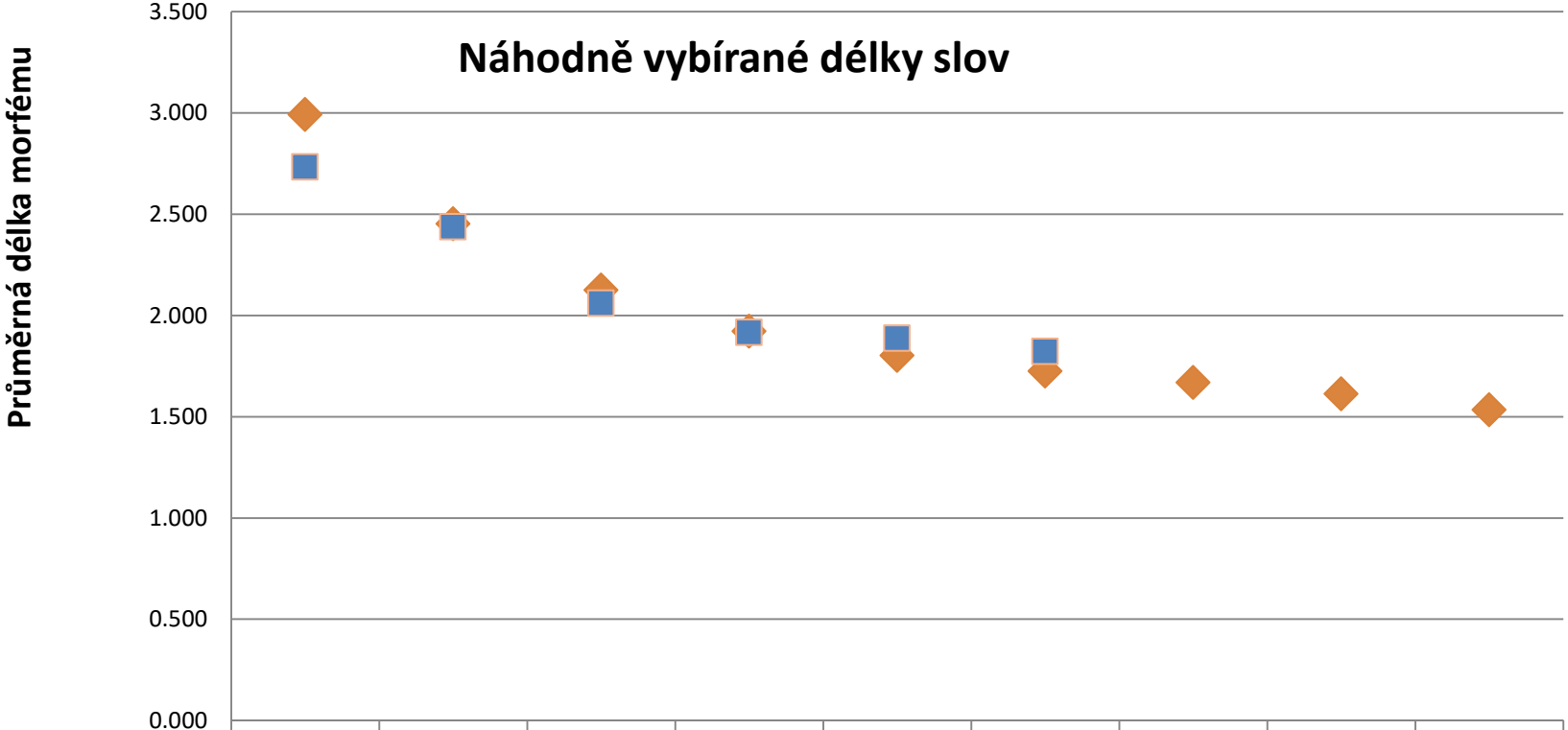


Jak distribuce délek morfů tak distribuce délek slov nahrazena Poissonovým rozdělením



	1	2	3	4	5	6	7	8	9	10	11
◆	2.39	1.93	1.72	1.59	1.50	1.43	1.38	1.34	1.30	1.27	1.23
■ orig	2.732	2.437	2.059	1.915	1.886	1.822					

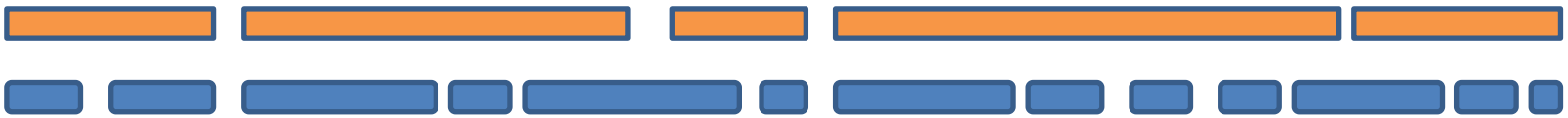
Nepromíchané morfy, slova náhodně vybíraná z distribuce



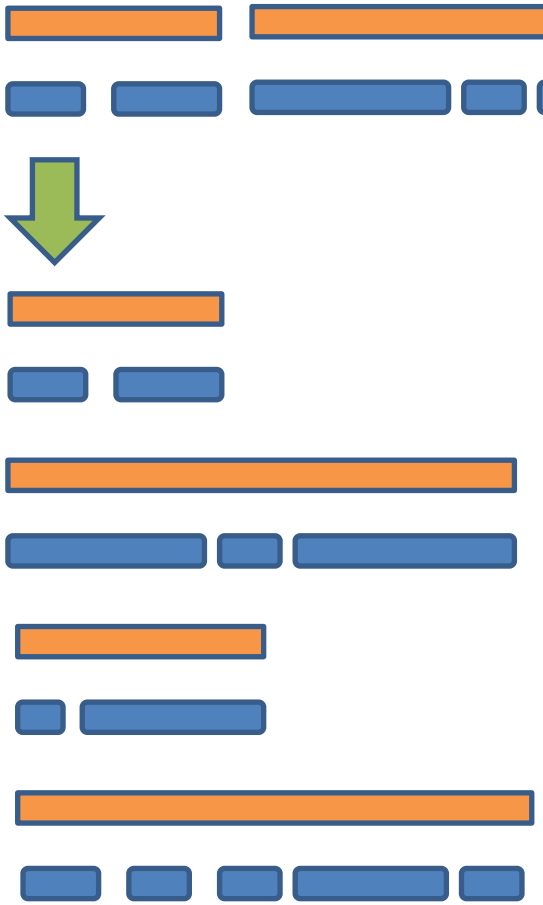
	1	2	3	4	5	6	7	8	9
◆	2.992	2.453	2.125	1.921	1.801	1.724	1.668	1.612	1.533
orig	2.732	2.437	2.059	1.915	1.886	1.822			

Další modely: Náhodně promíchané morfy i slova

(otázka, jak správně synchronizovat délky morfů a délky vět, aby hranice segmentů byly na stejném místě)

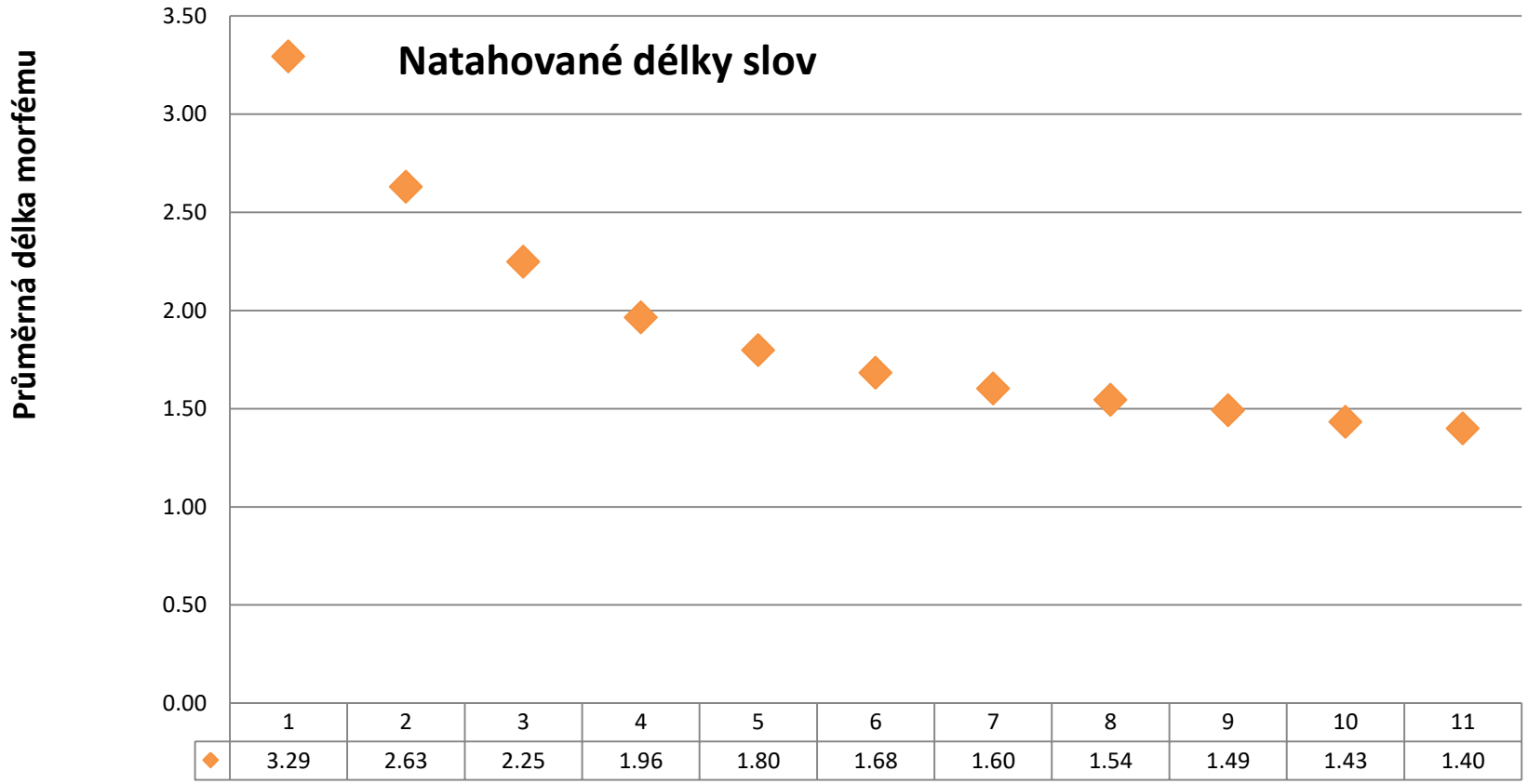


Nedosahující slova natáhneme

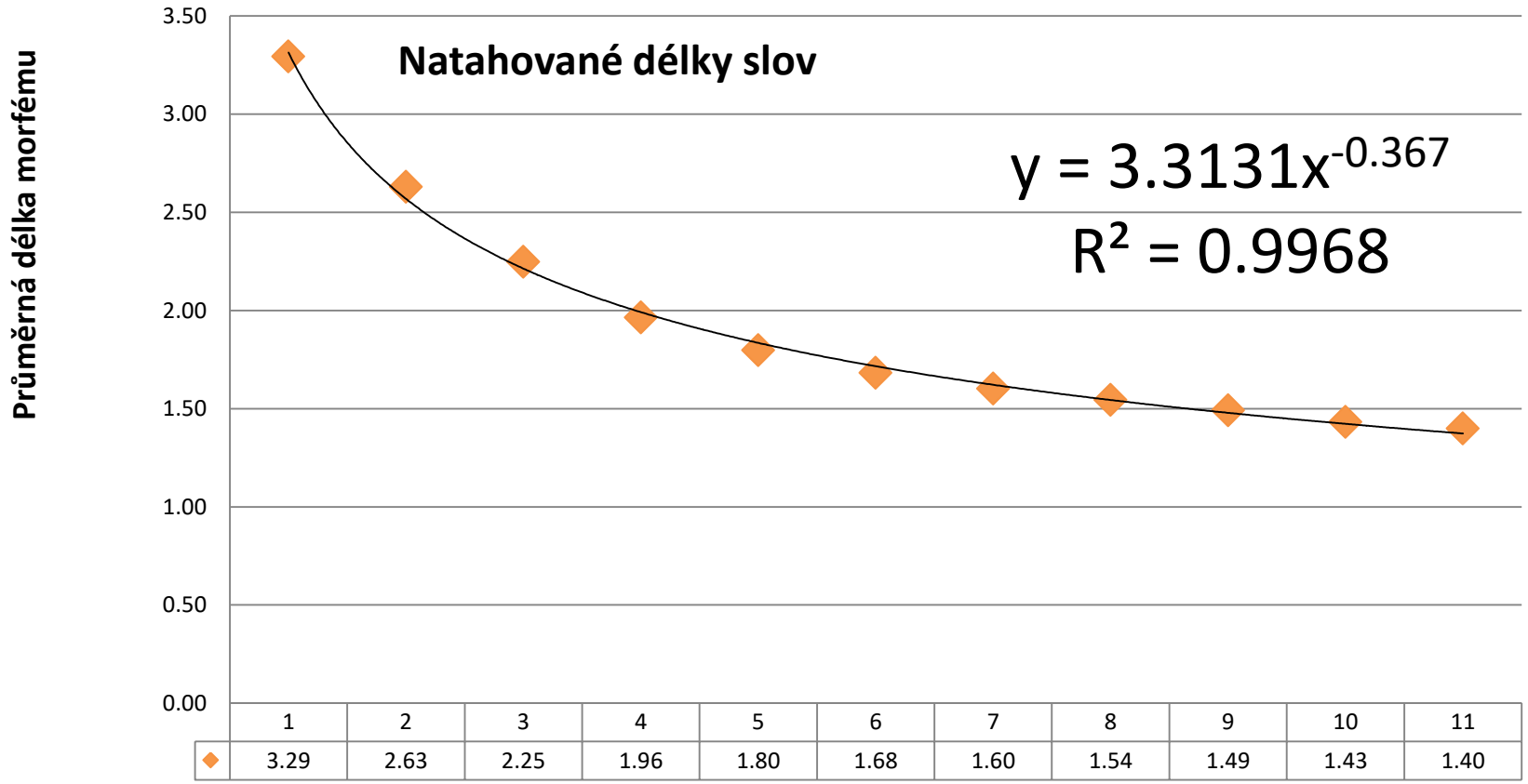


Ovšem na poslední slovo se nám tak nedostává morfémů...

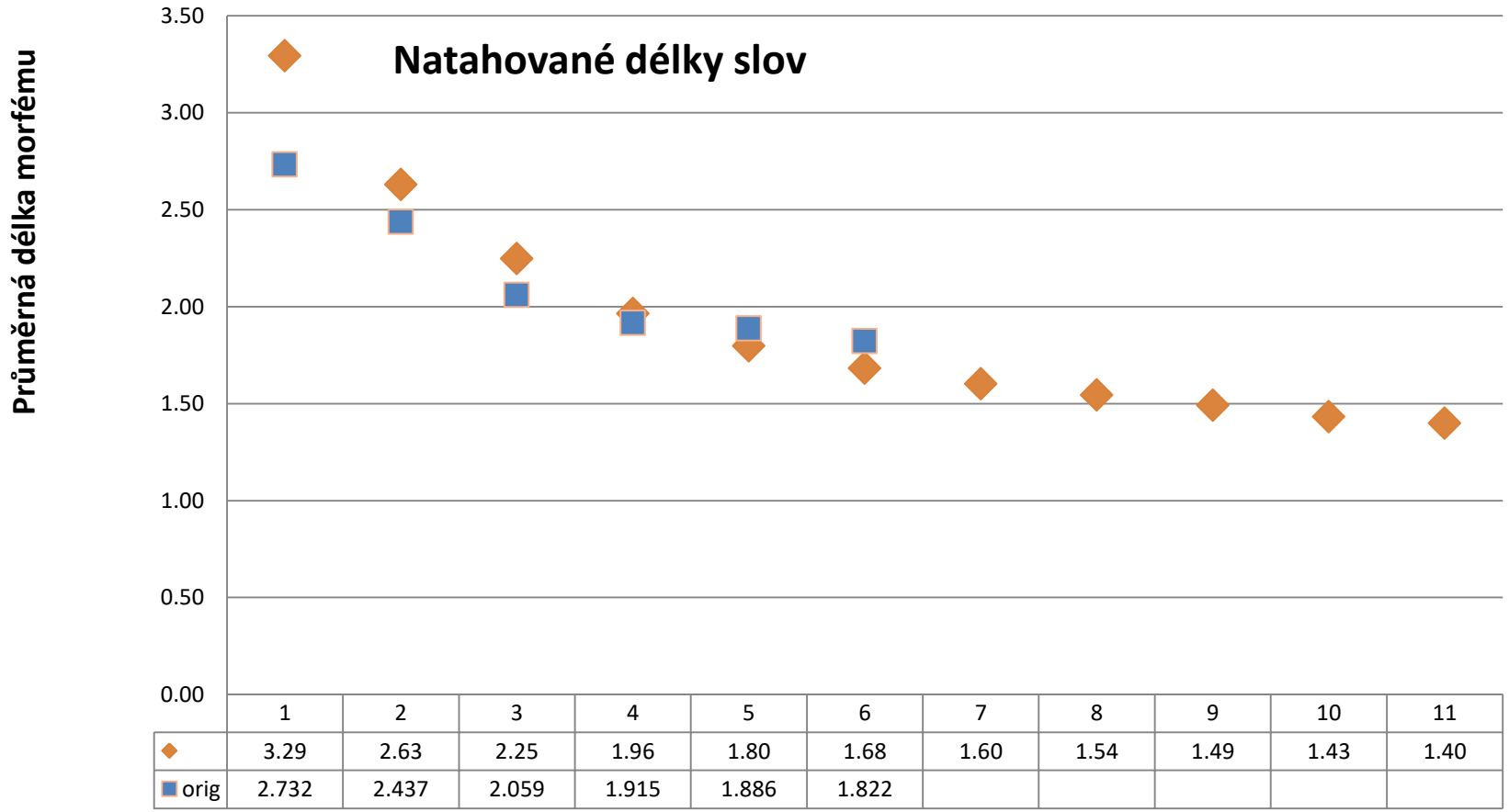
Nedosahující slova natažena



Nedosahující slova natažena

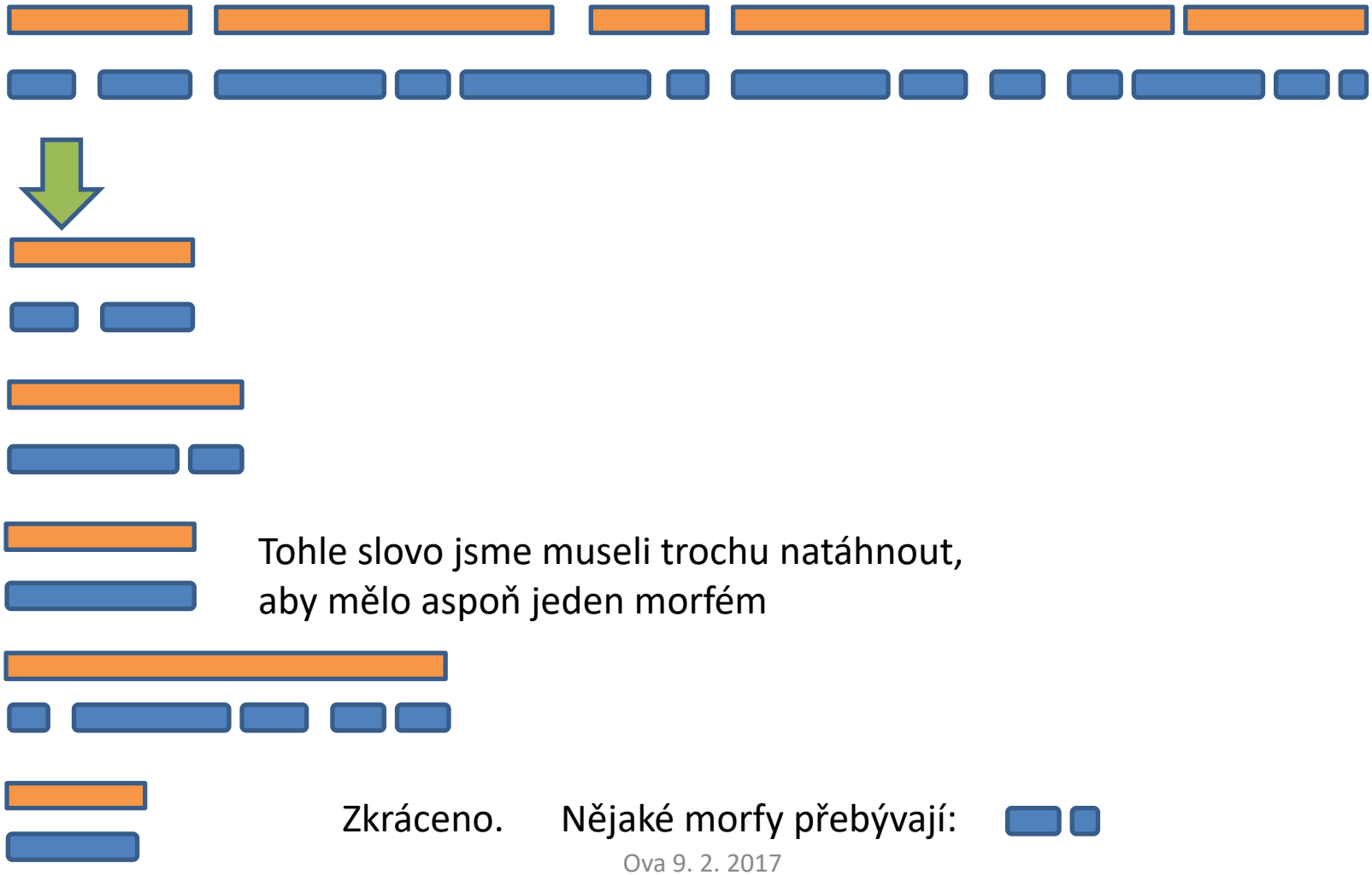


Nedosahující slova natažena

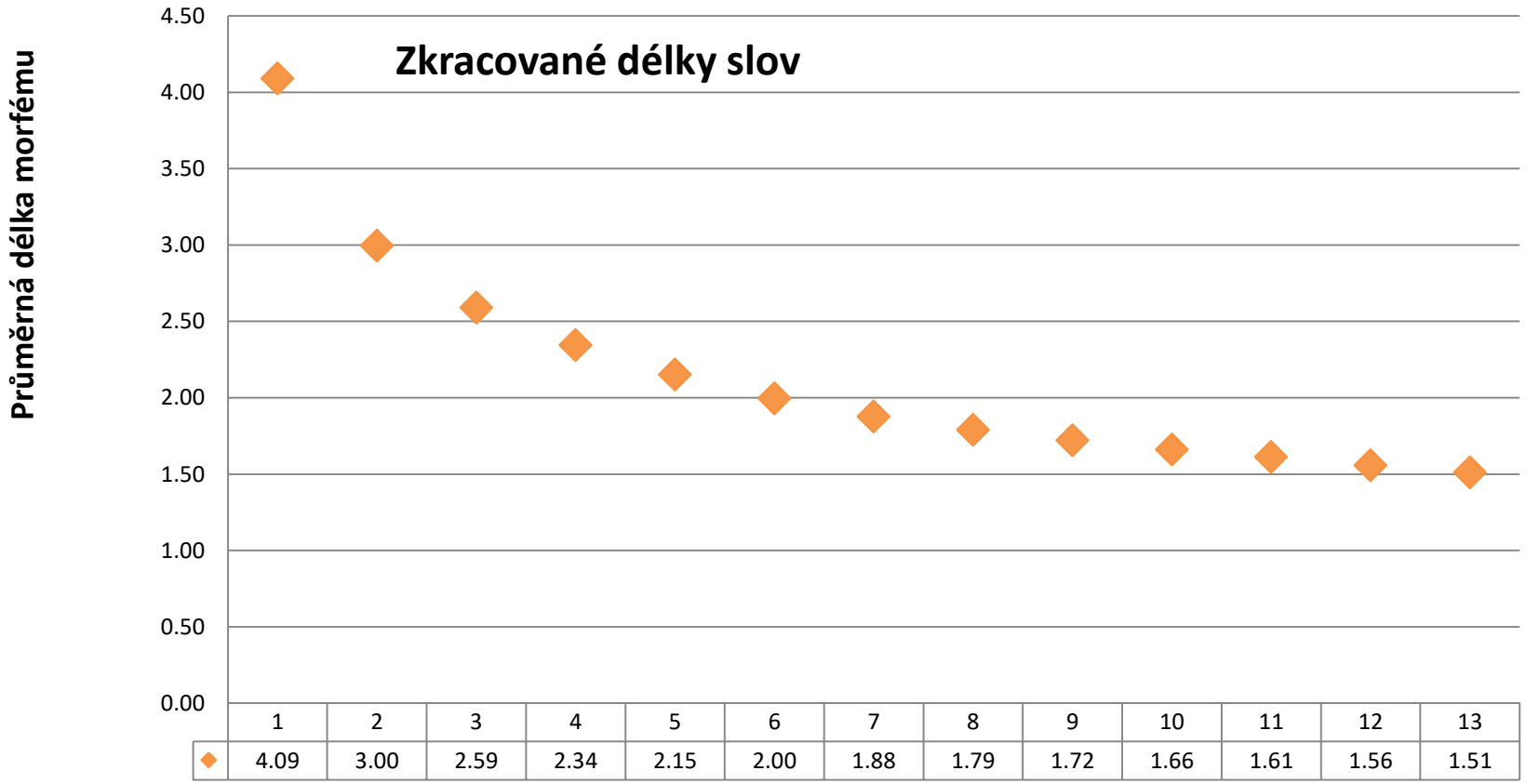


	1	2	3	4	5	6	7	8	9	10	11
◆	3.29	2.63	2.25	1.96	1.80	1.68	1.60	1.54	1.49	1.43	1.40
■ orig	2.732	2.437	2.059	1.915	1.886	1.822					

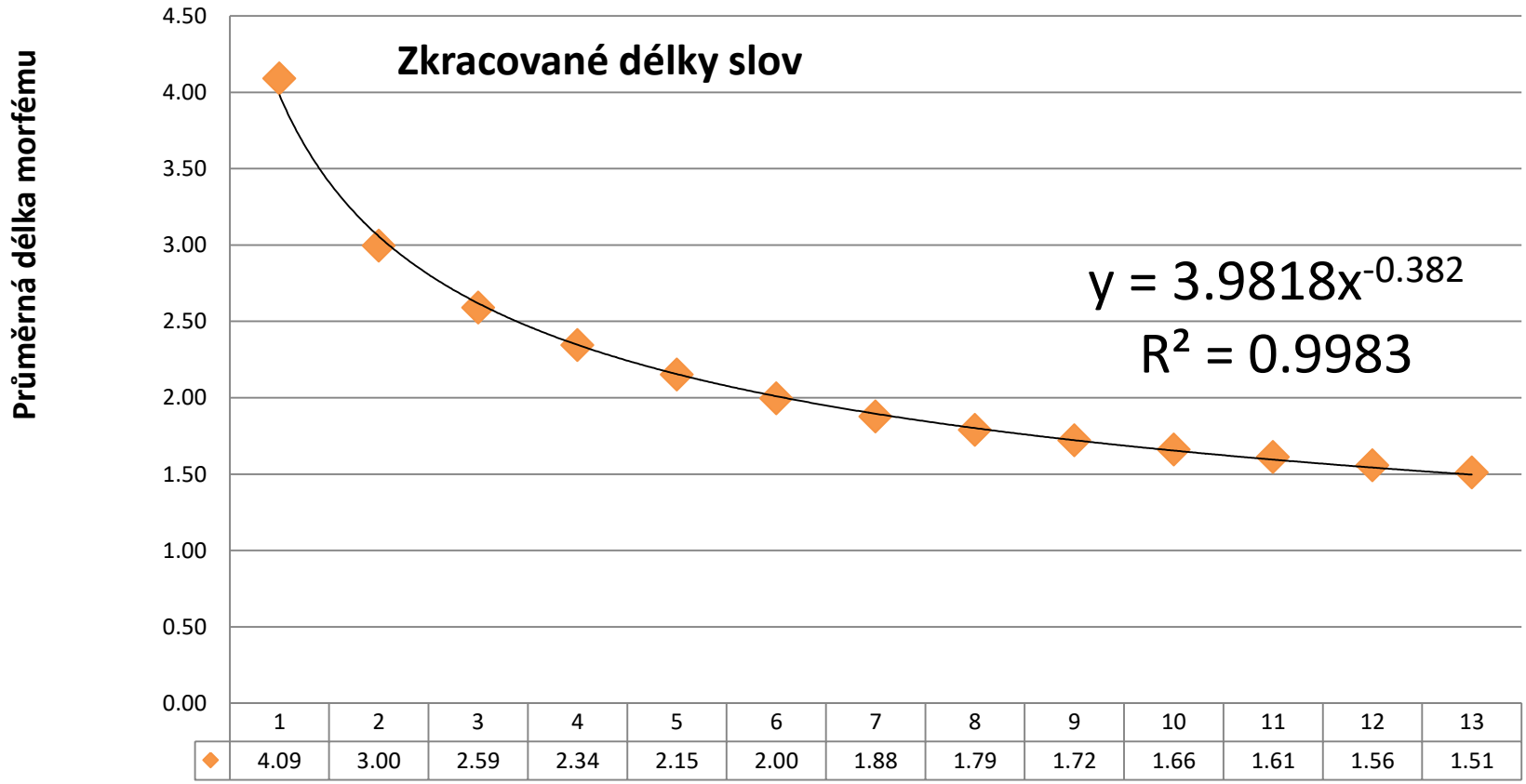
Přesahující slova usekneme



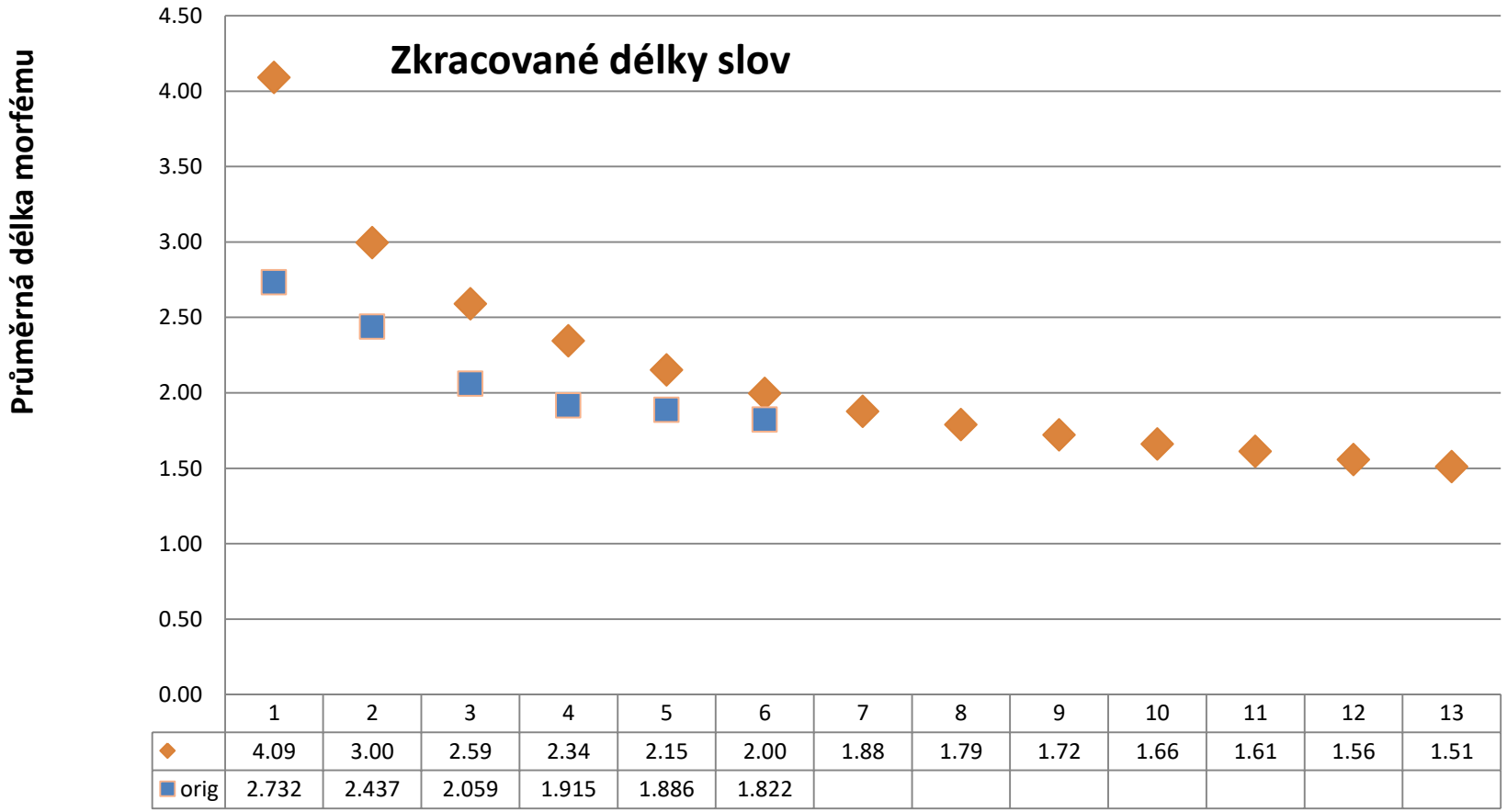
Přesahující slova useknuta



Přesahující slova useknuta

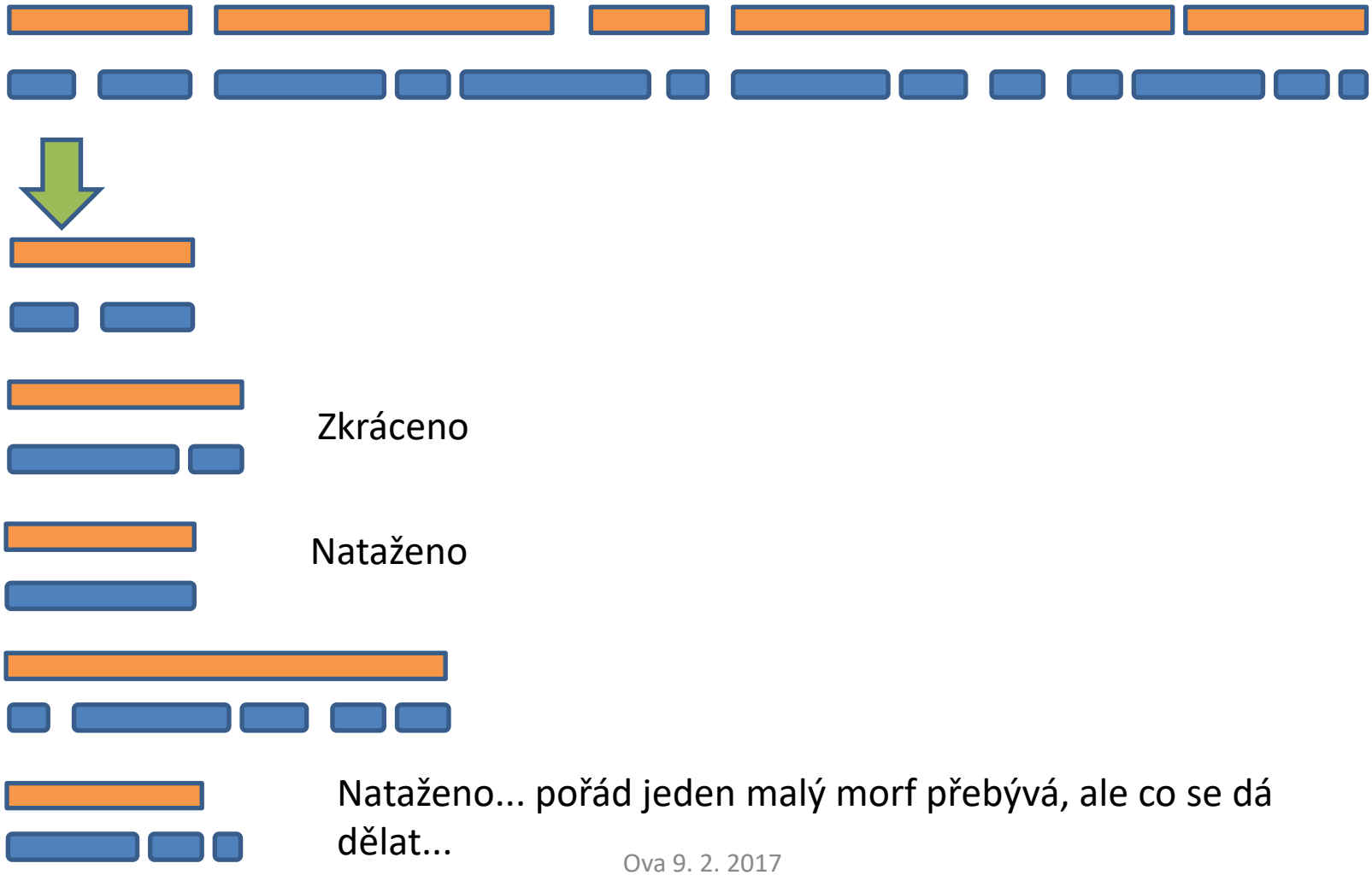


Přesahující slova useknuta

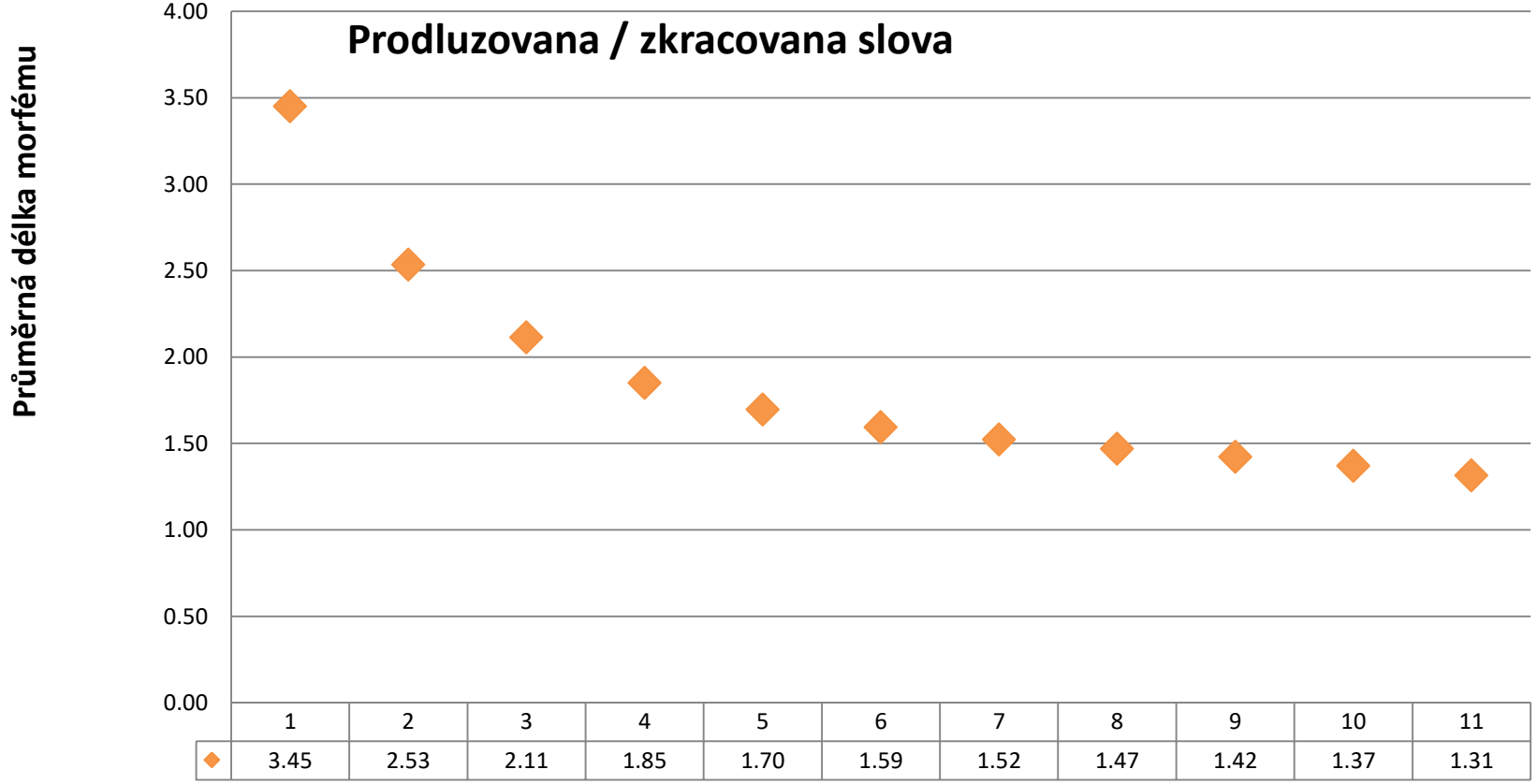


	1	2	3	4	5	6	7	8	9	10	11	12	13
◆	4.09	3.00	2.59	2.34	2.15	2.00	1.88	1.79	1.72	1.66	1.61	1.56	1.51
orig	2.732	2.437	2.059	1.915	1.886	1.822							

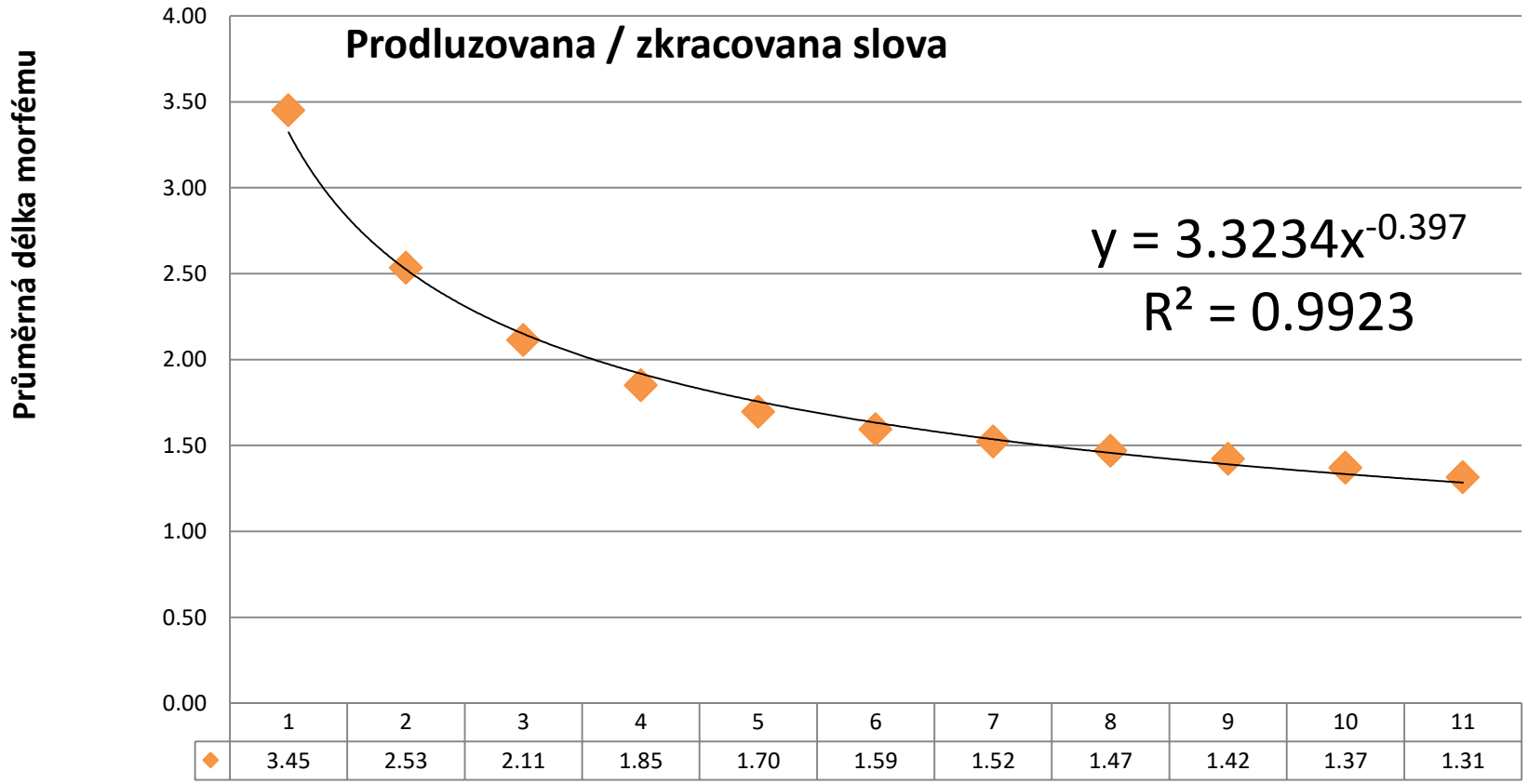
Slova někdy natáhneme, jindy zkrátíme, podle toho, co je jednodušší



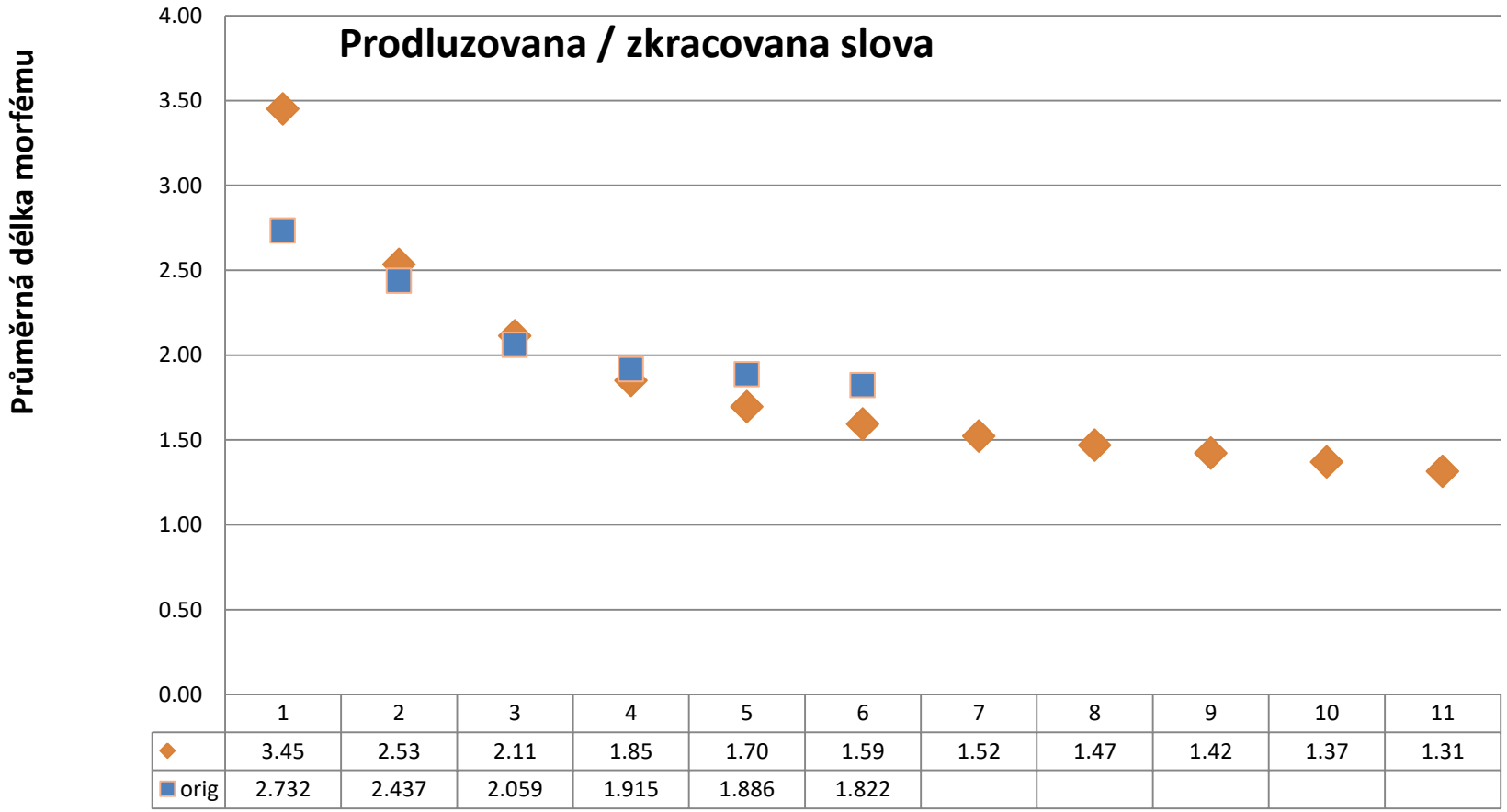
Prokrustovo lože



Prokrustovo lože

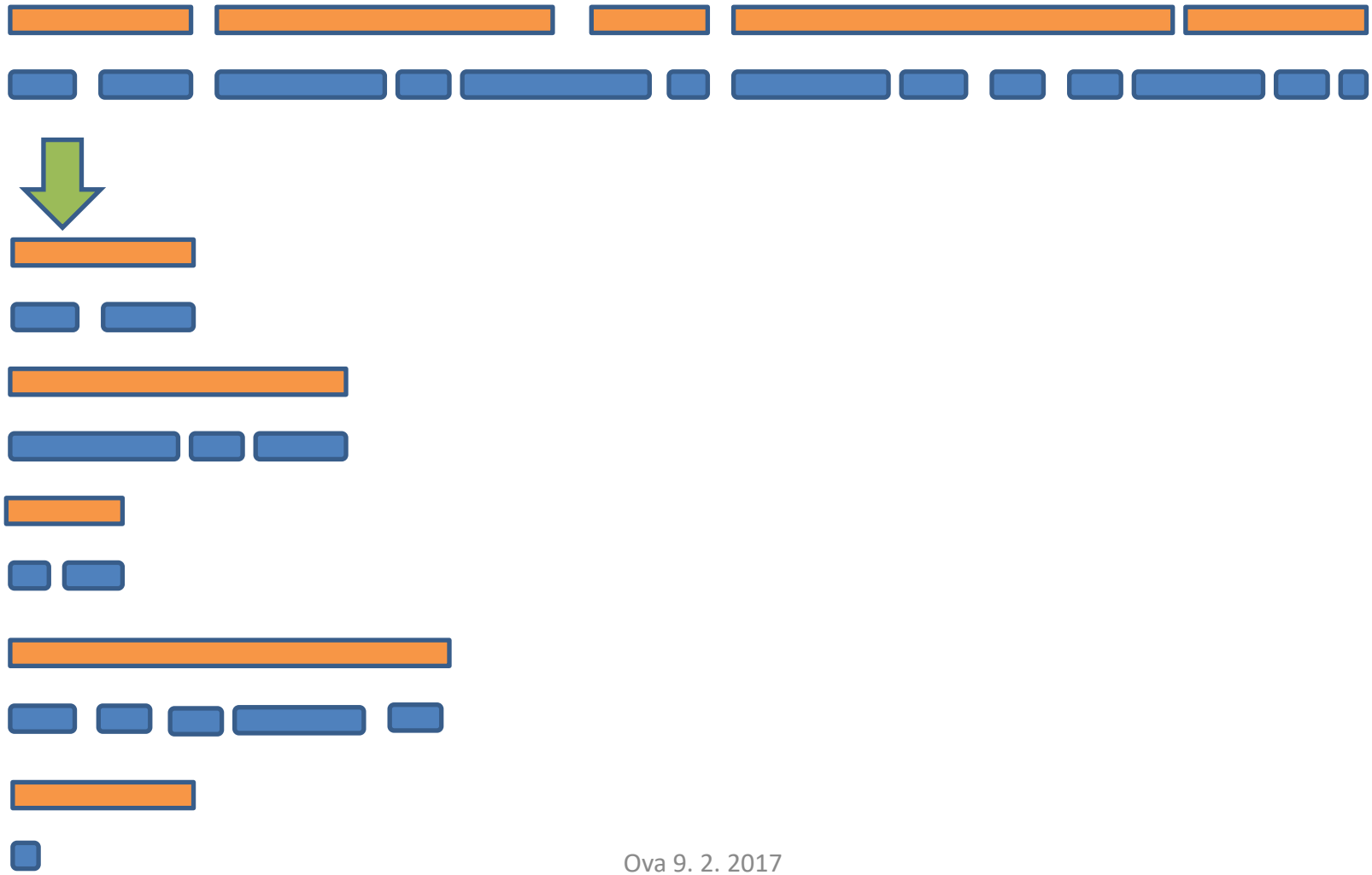


Prokrustovo lože

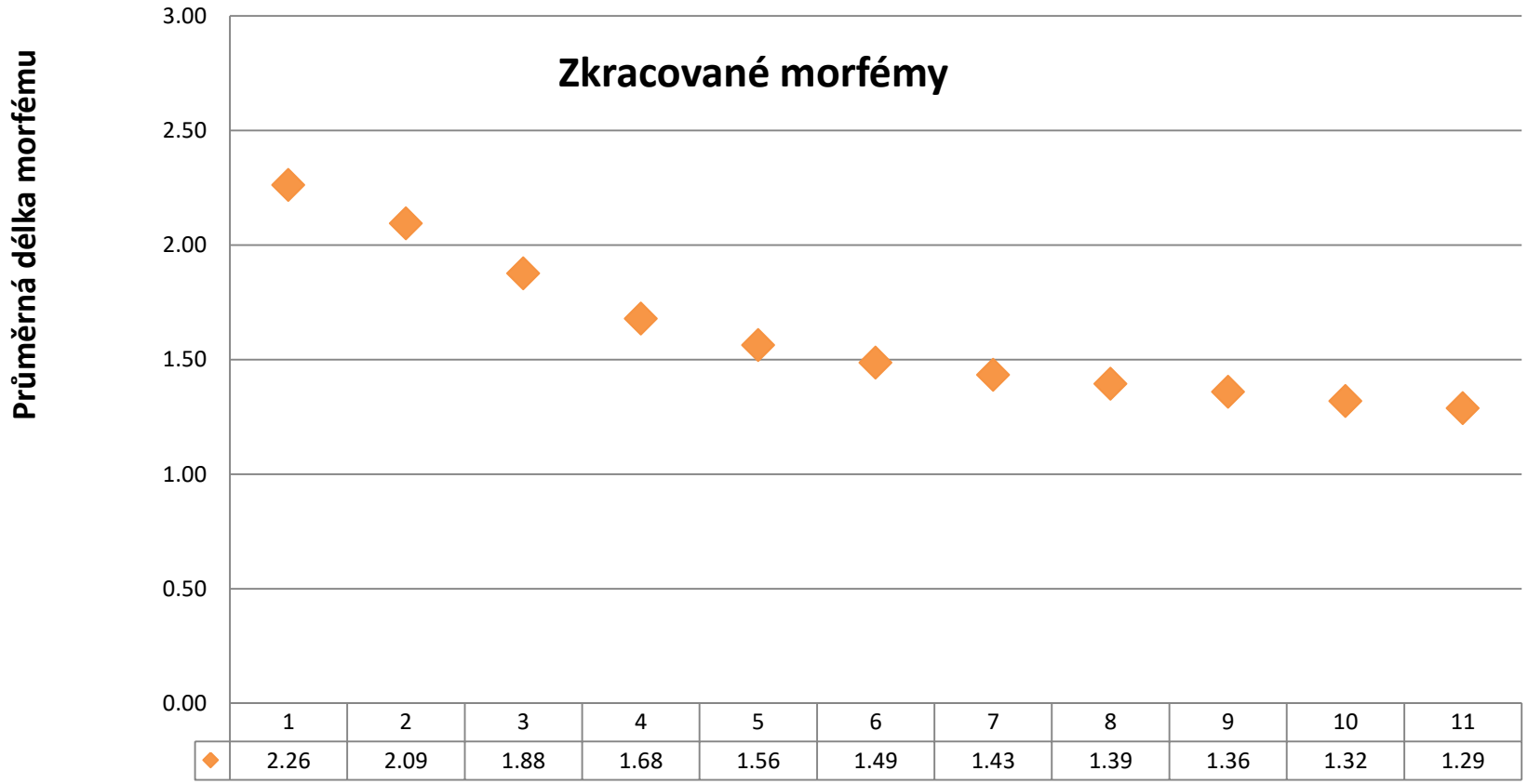


	1	2	3	4	5	6	7	8	9	10	11
◆	3.45	2.53	2.11	1.85	1.70	1.59	1.52	1.47	1.42	1.37	1.31
orig	2.732	2.437	2.059	1.915	1.886	1.822					

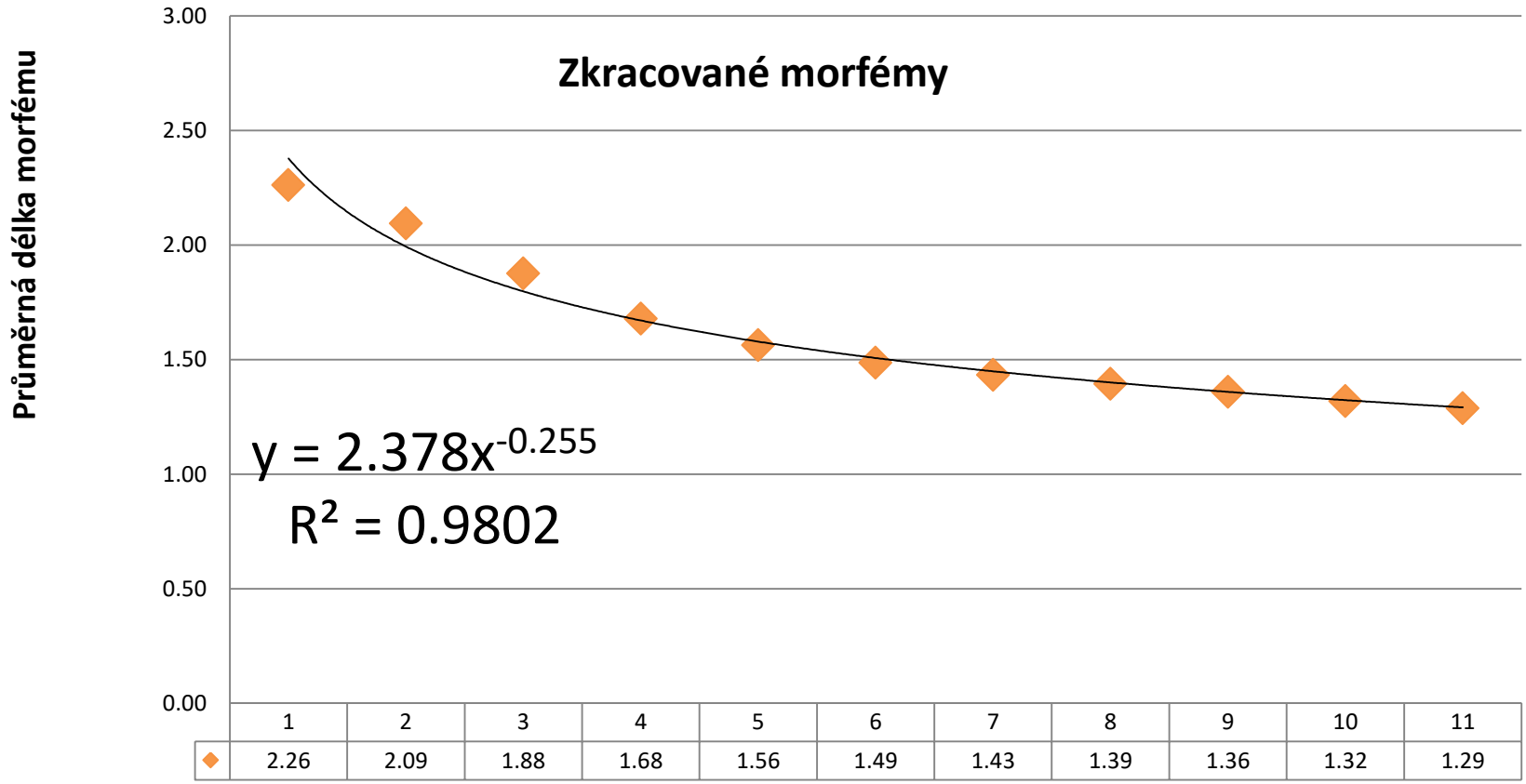
Zkracujeme morfy



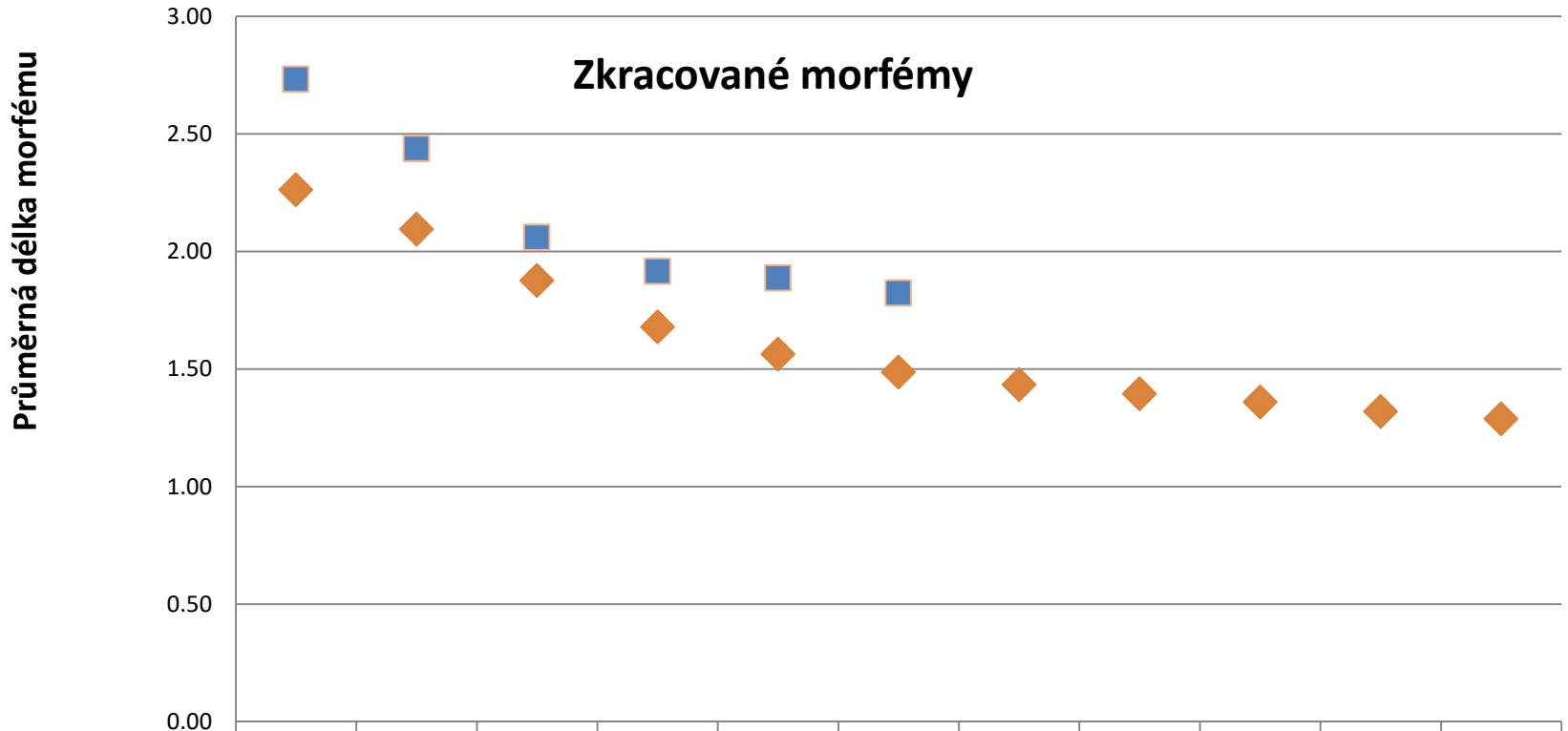
Zkrácené morfy



Zkrácené morfy



Zkrácené morfy



	1	2	3	4	5	6	7	8	9	10	11
◆	2.26	2.09	1.88	1.68	1.56	1.49	1.43	1.39	1.36	1.32	1.29
orig	2.732	2.437	2.059	1.915	1.886	1.822					

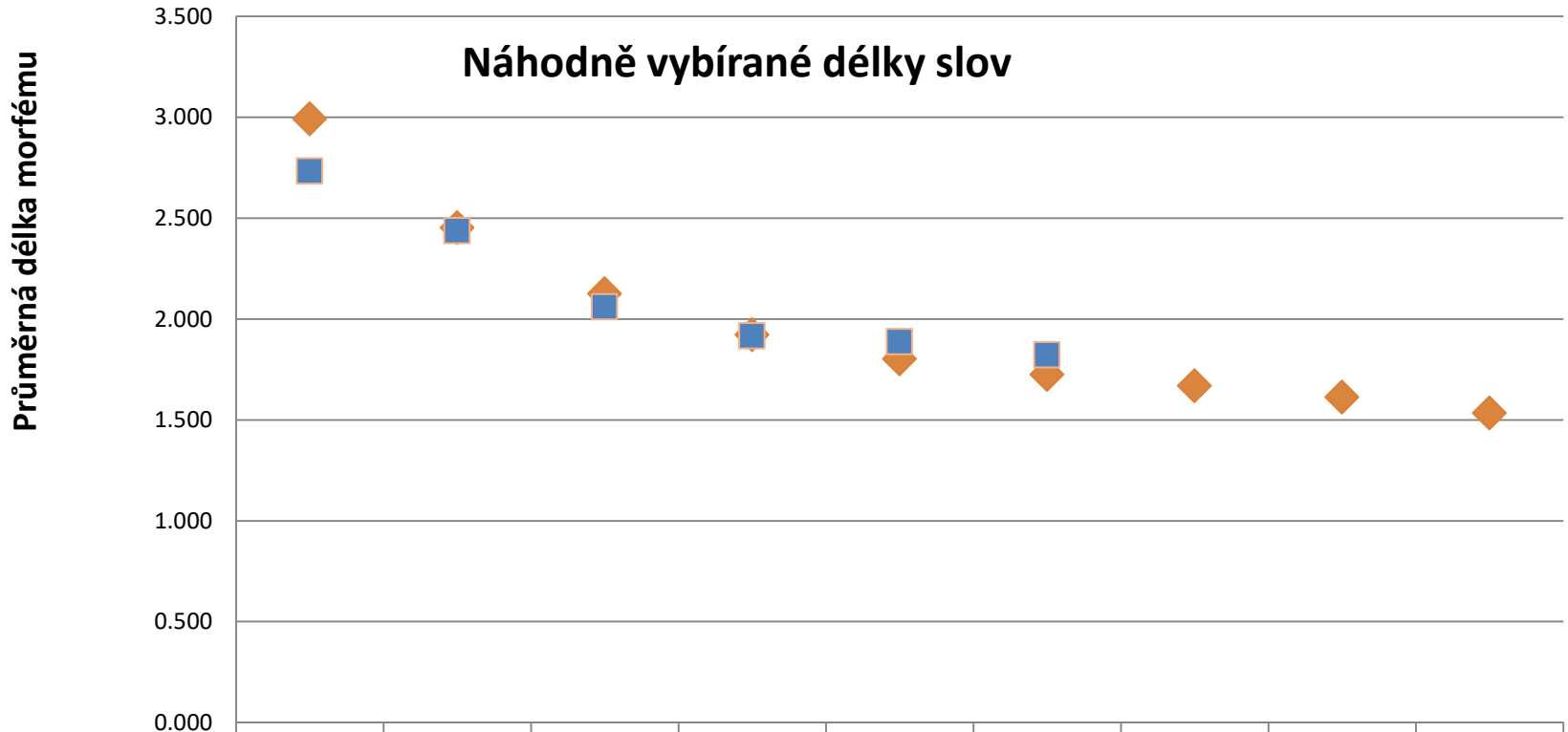
Konec Menzerathova zákona?

Práve naopak:

- teprve teď víme, proč platí i mimo lingvistiku
- Můžeme ho začít mimo lingvistiku cíleně vyhledávat
- Stává se z něj něco jako Benfordův zákon: něco co platí pro spoustu nesouvisejících fenoménů, které spojuje nějaký princip. Nyní víme, že tím principem může být **malá variabilita konstruktů a velká variabilita konstituentů**.

- Nicméně to neznamená, že tento princip je jediný, který hraje roli. Například morfémy nejsou v textu rozloženy rovnoměrně a tak má křivka poněkud jiné parametry než jak plyne z náhodných modelů, pokud ponecháme původní rozložení, jak jsme viděli:

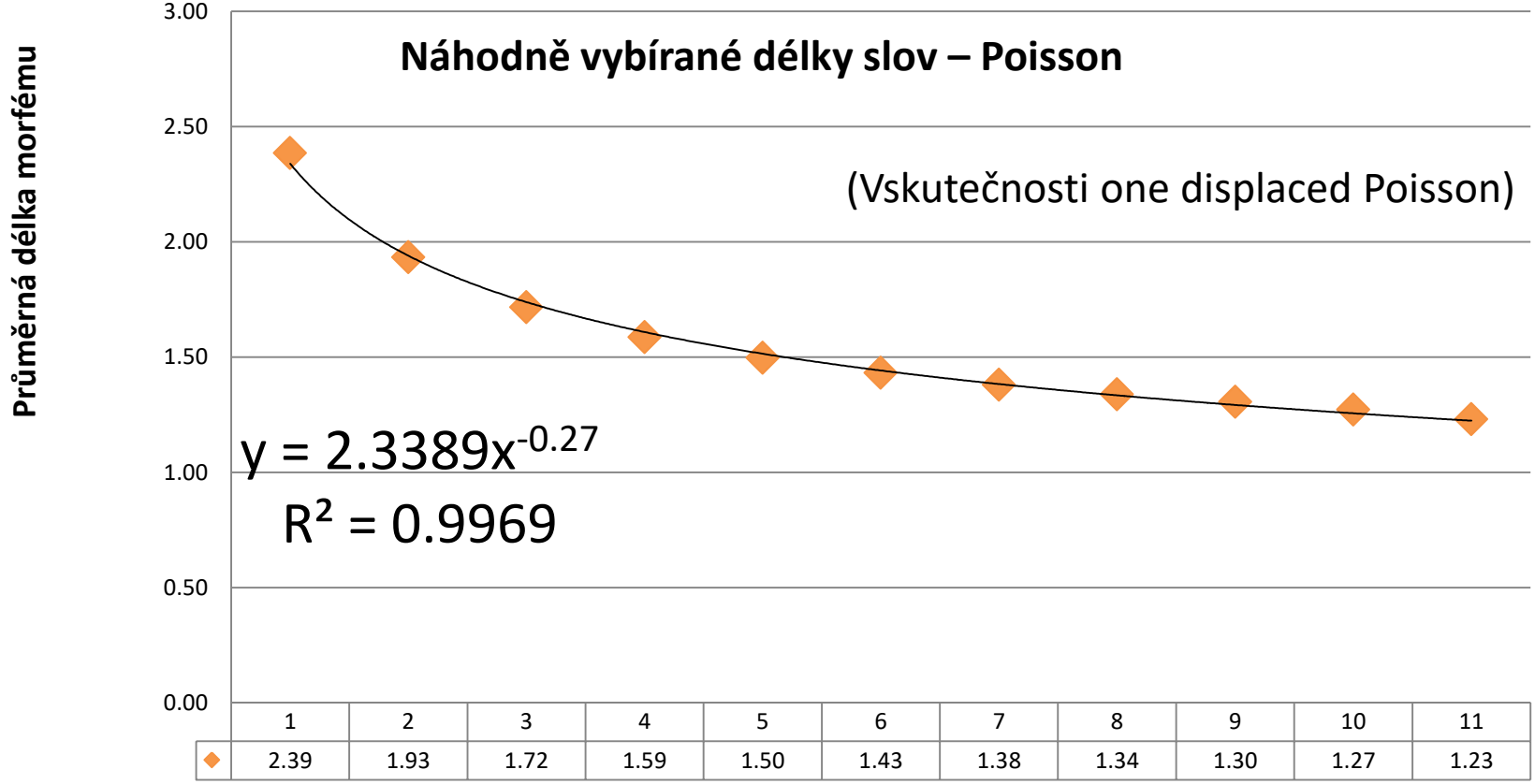
Nepromíchané morfy, slova náhodně vybíraná z distribuce (opakování)



	1	2	3	4	5	6	7	8	9
◆	2.992	2.453	2.125	1.921	1.801	1.724	1.668	1.612	1.533
orig	2.732	2.437	2.059	1.915	1.886	1.822			

- Ovšem není třeba počítat s tím, že text má nějakou magickou distribuci konstruktů a konstituentů, docela dobře stačí, když mají jak konstrukty tak konstituenty obyčejné poissonovské rozdělení (pochopitelně konstrukty mají vyšší průměr než konstituenty a tím i menší variabilitu) – opět připomínám graf:

Jak distribuce délek morfů tak distribuce délek slov nahrazena Poissonovým rozdělením

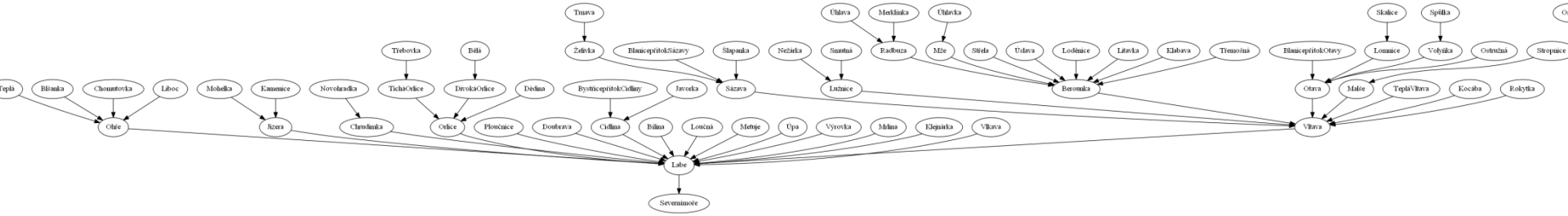


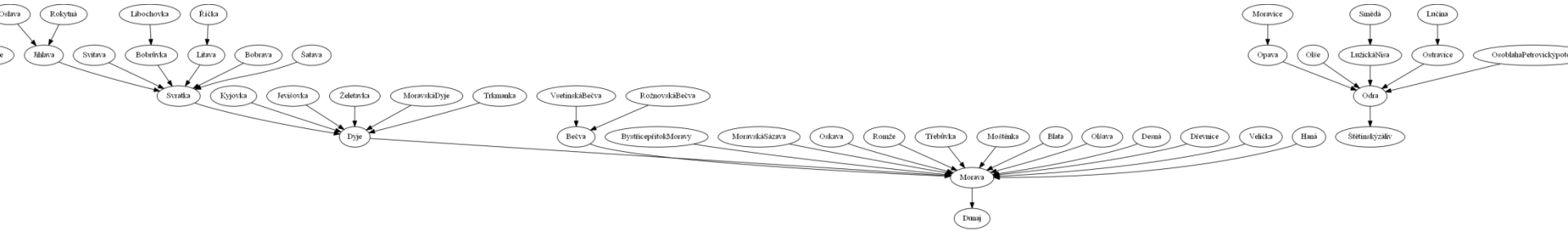
- Otázka pro matematiky: je klasický Altmannův vzorec

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

skutečně nejlepší, pokud modelujeme Menzerathův vztah takto (třeba pomocí dvou poissonů? šlo by to odvodit? Já osobně tady končím...)

- Každopádně jsme vyvrátili, že Menzerathův vztah bude platit pouze *in informationsverarbeitenden Systemen*, ale můžeme ho začít hledat kdekoli.
- Paradoxně aby to platilo, tak podmínka je, aby nebyl vztah mezi počtem konstituentů v konstruktu a průměrným počtem subkonstutuentů (alespoň tedy ne takový, který by šel proti principu, který jsme sledovali) – například řeky: čím víc má řeka přítoků, tím je větší, čili tím víc přítoků budou mít i její přítoky.



Děkuji za pozornost!!!

jiri@milicka.cz