Radek Čech, Emmerich Kelih, Jan Mačutek

Vliv sémantiky na vlastnosti pádové distribuce podstatných jmen v češtině

Korpusová lingvistika 2014 (17. – 19. 9. 2014)
Praha

Distribuce pádů vs. sémantika

- Proč (a jaký) by měla mít sémantika vliv na distribuci pádů substantiv?
- východiska (předběžná)
  - substantiva denotující osoby mají tendenci se vyskytovat nejčastěji v nominativu (vlivem tendence vyskytovat se v sémantické roli agentu)
  - u substantiv denotujících např. neživé předměty nebo abstraktní entity není jejich morfosyntaktický status jednoznačný

Subst. maskulina anim. vs. inanim

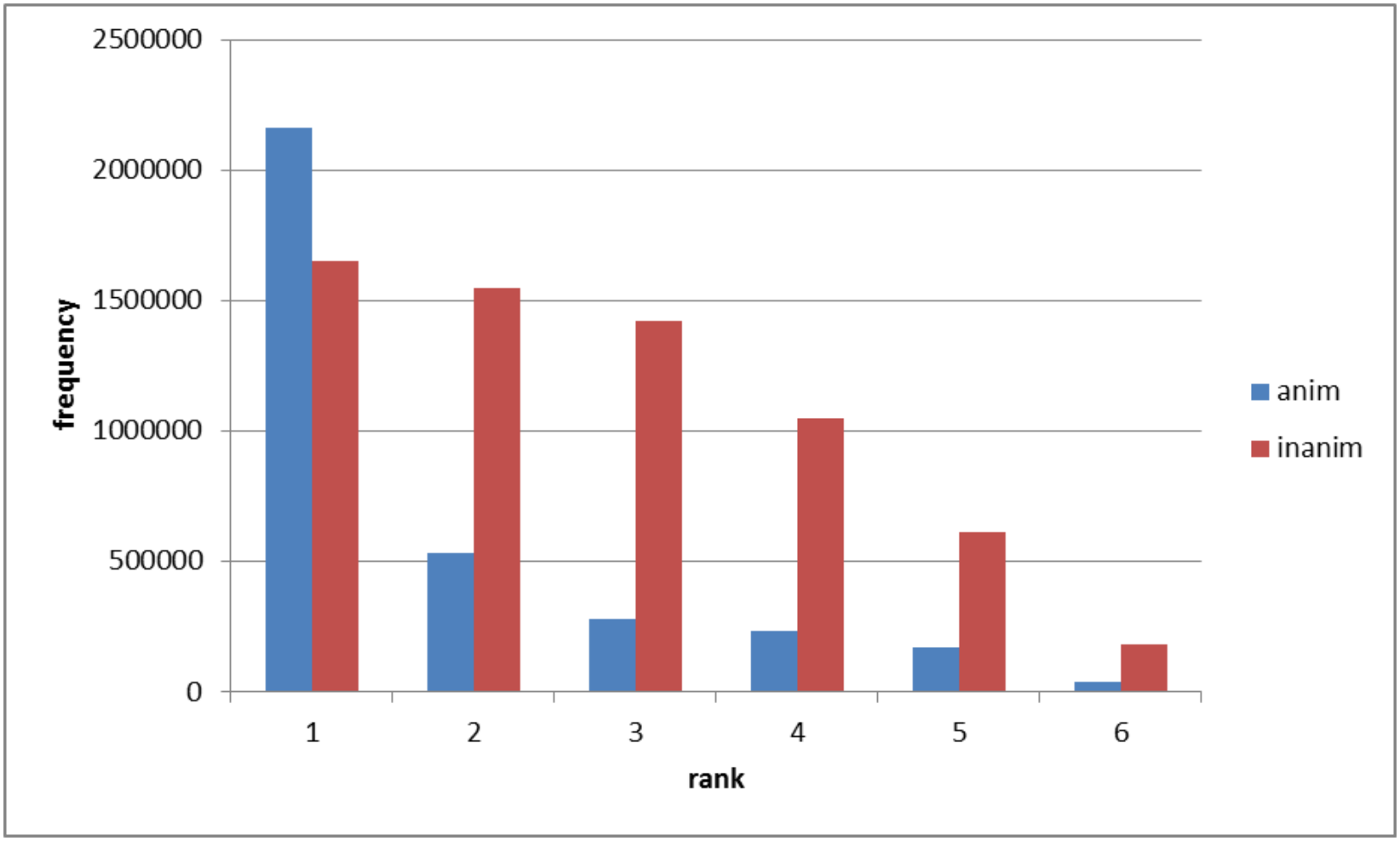
Anim. sg. (SYN2010)

pád	frekvence
nom.	2161013
gen.	532579
acc.	278806
instr.	233327
dat.	170042
loc.	39956

Inanim. sg. (SYN2010)

pád	frekvence
gen.	1649641
acc.	1546412
nom.	1422769
loc.	1045981
instr.	613918
dat.	184674

Subst. maskulina anim. vs. inanim



Distribuce pádů vs. sémantika

- Jaké obecné principy řídící jazykové chování by mohly mít vliv na předpokládaný vztah mezi sémantikou a distribucí pádů?
- Jaké jsou tzv. hraniční podmínky?
  - rod
  - číslo
  - polysémie atd.

Teoretické předpoklady

- distribuce pádů je výsledkem tzv. diverzifikačního procesu
- diverzifikace (obecně)
  - jednotka (např. slovo) – kategorie (pád, rod, číslo atd.) – jednotlivé instance (nom., gen…; mask., fem., neut.…)
  - pokud jednotka v rámci kategorie podléhá diverzifikaci, frekvence nejsou distribuovány rovnoměrně
  - jedná se o obecný jev, který je charakteristický pro jazykový systém

Analýza

- diverzifikace pádů u jednotlivých substantiv
- diverzifikace vs. sémantika
- existují rozdíly mezi rody?

Hypotézy

- pádové distribuce jednotlivých substantiv (anim. a inanim u všech tří rodů) je možné modelovat stejnou matematickou funkcí
- parametry této funkce se budou signifikatntně lišit u
  - anim. vs. inanim.
  - všech třech rodů

Data

- SYN 2010
- 5 nejfrekventovanějších anim. a inanim. substantiv
- 10 v rámci každého rodu (mask., fem., neut.)
- celkem analyzováno 30 substantiv
  - konkrétní substantiva
  - bez vlastních jmen
  - pouze singulár

Analyzovaná substantiva

mask. anim.	mask. inanim	fem. anim.	fem. inanim.	neut. anim.	neut. inanim.
člověk	dům	žena	hlava	dítě	město
muž	stůl	matka	ruka	děvče	tělo
pan	měsíc	paní	škola	miminko	auto
otec	vzduch	dívka	ulice	děcko	divadlo
ředitel	byt	dcera	tvář	děťátko	srdce

Výsledky – pořadí jednotlivých pádů

pořadí	mask. anim.					mask. inanim					fem. anim.					fem. inanim.					neut. anim.					neut. inanim.				
1	N	N	N	N	N	G	G	A	G	G	N	N	N	N	N	I	A	G	L	A	N	N	N	N	N	G	G	A	G	G
2	G	A	G	G	G	L	A	G	L	A	A	G	G	A	A	A	L	L	G	L	A	A	A	A	A	N	A	G	N	N
3	A	G	D	A	I	A	L	L	N	L	G	A	A	G	G	G	I	N	A	I	G	G	G	G	G	L	N	N	L	A
4	D	I	A	I	A	N	D	I	A	N	I	D	D	D	I	L	G	A	N	G	I	I	I	I	I	A	L	I	A	L
5	I	D	I	D	D	I	I	N	I	I	D	I	I	I	D	N	N	I	I	N	D	D	D	D	D	I	I	L	I	I
6	L	L	L	L	L	D	N	D	D	D	L	L	L	L	L	D	D	D	D	D	L	L	L	L	L	D	D	D	D	D

N … nominativ
G … genitiv
D … dativ
A … akuzativ
L … lokál
I … instrumentál

Výsledky
(příklad)

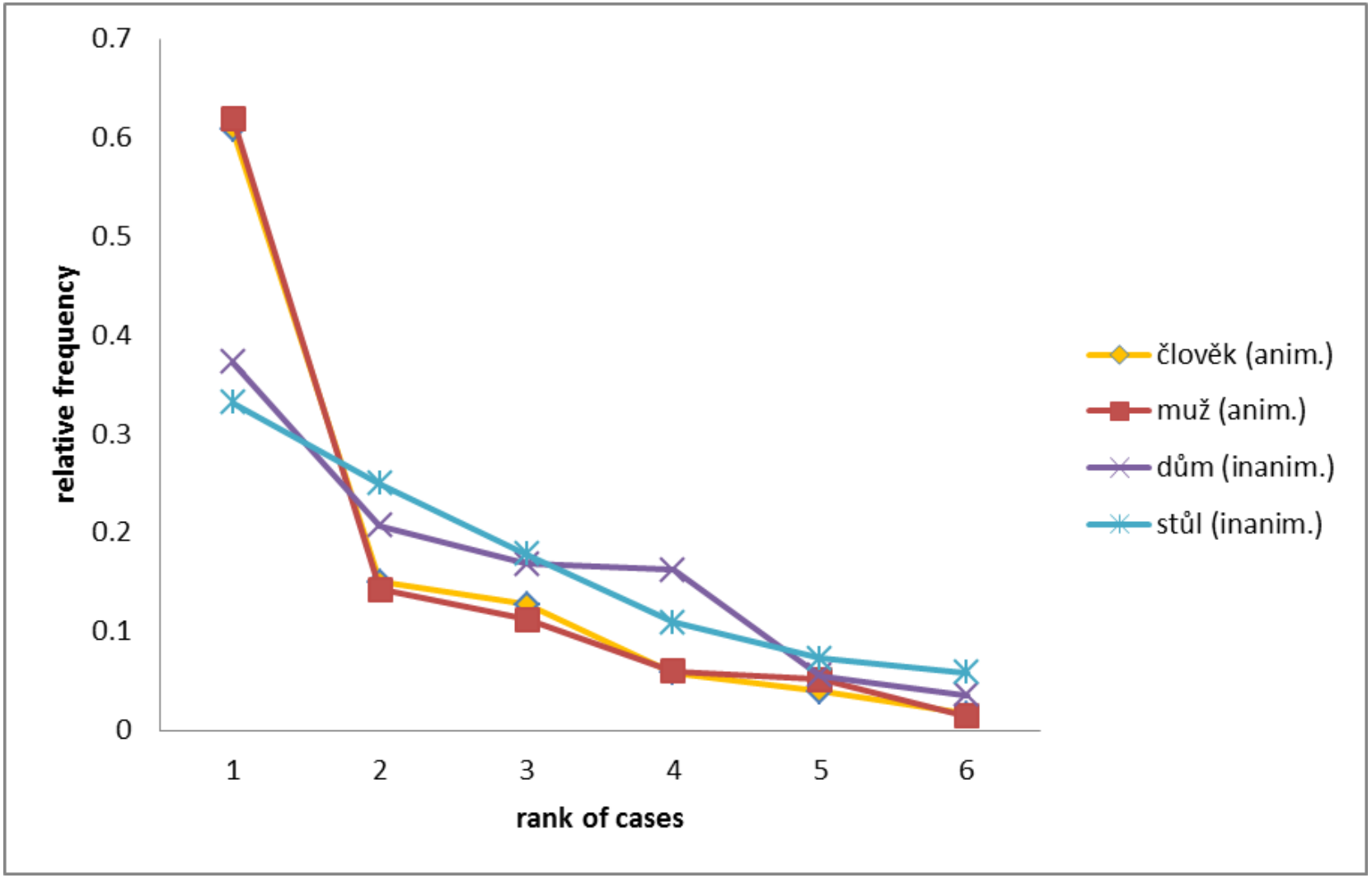
lemma *člověk*

pád	frekvence
nom.	43543
gen.	10732
acc.	9128
dat.	4186
instr.	2833
loc.	1194

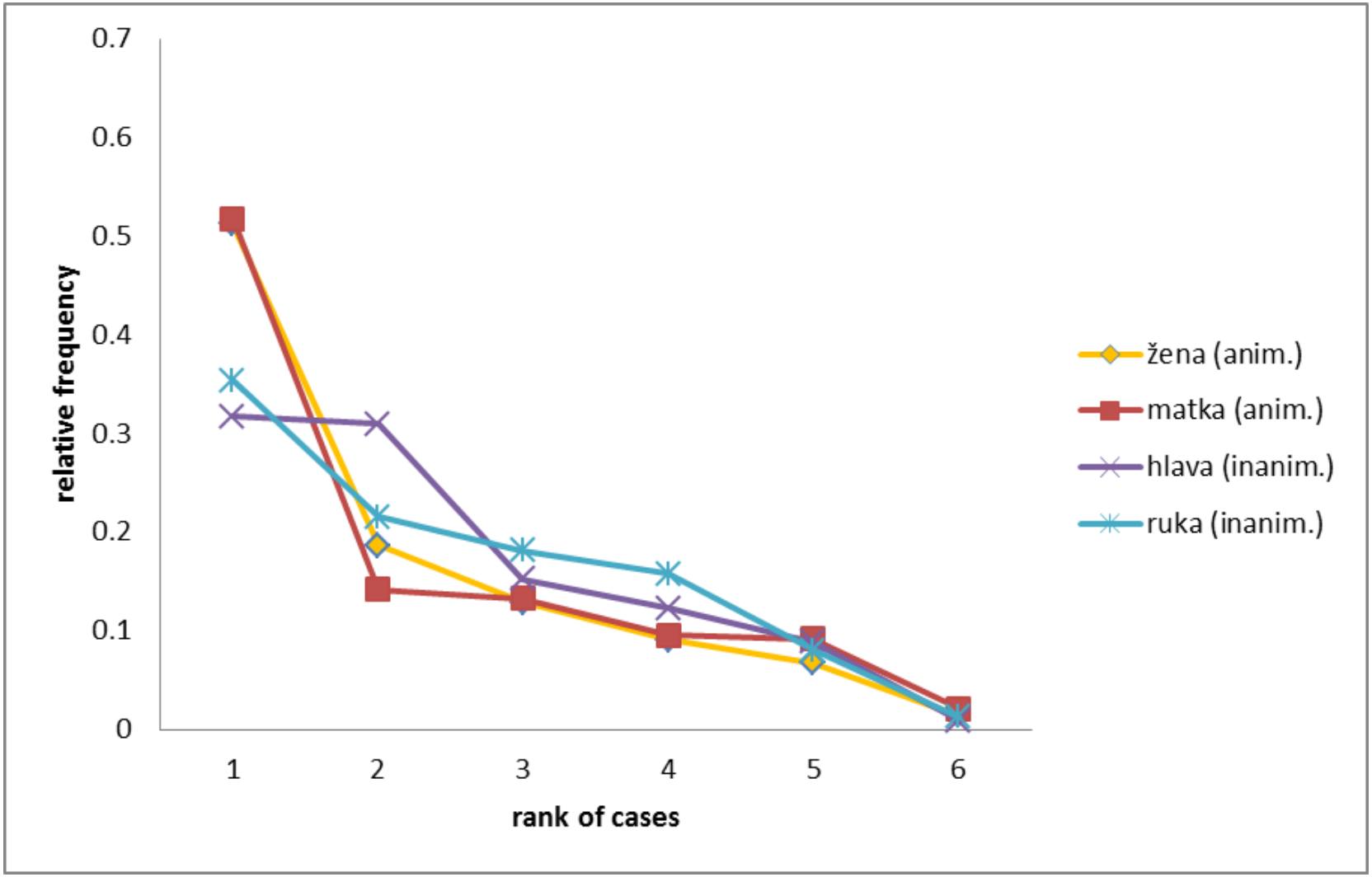
lemma *dům*

pád	frekvence
gen.	20145
loc.	11194
acc.	9109
nom.	8777
inst.	2966
dat.	1900

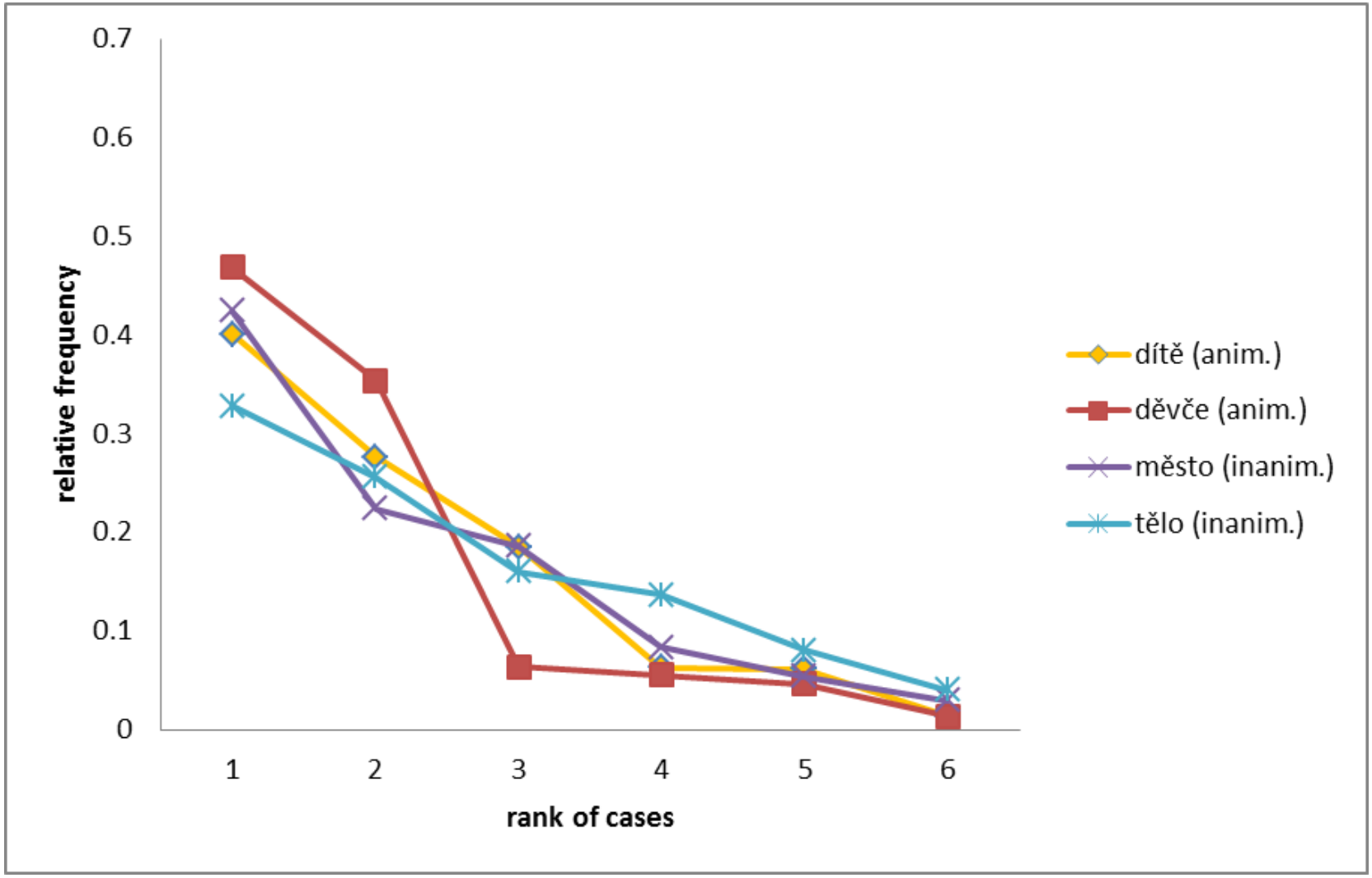
Rankové frekvenční distribuce (mask.)



Rankové frekvenční distribuce (fem.)



Rankové frekvenční distribuce (neut.)



Model

$$y = ae^{-bx}$$

x … pořadí pádu

y … frekvence pádu

a, *b* … parametry

- speciální případ Wimmerova-Altmannova modelu

Výsledky aplikace modelu na data (mask.)

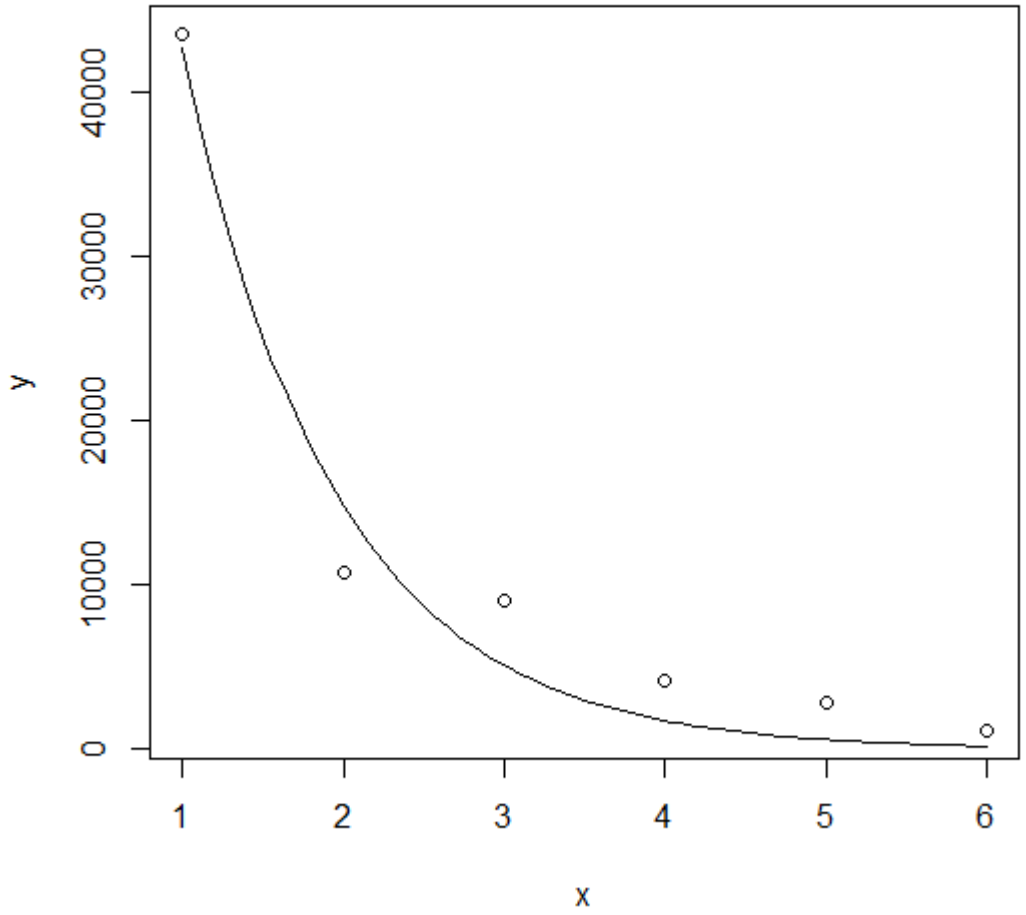
mask. anim.

lemma	*a*	*b*	R^2
člověk	123208.0	1.059	0.9723
muž	91962.4	1.149	0.9711
pan	39832.5	0.685	0.9887
otec	48060.8	1.014	0.9476
ředitel	46335.5	1.142	0.9950

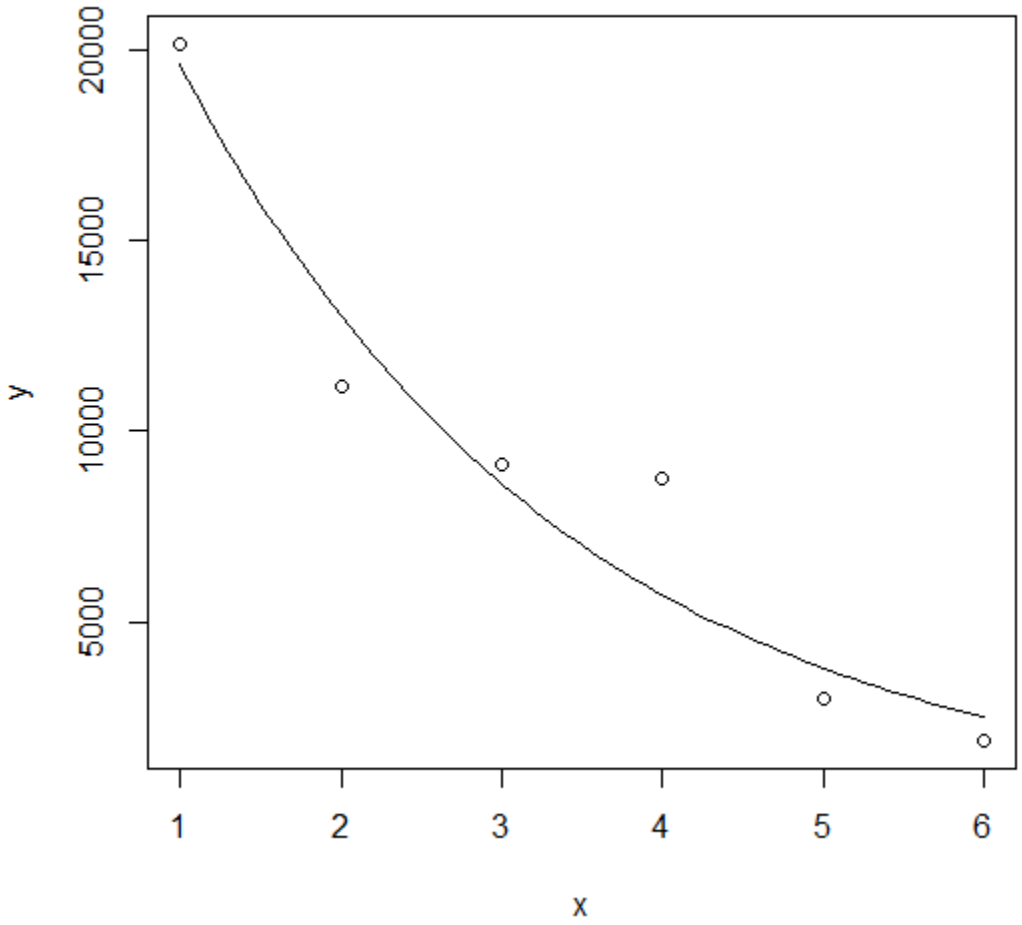
mask. inanim.

lemma	*a*	*b*	R^2
dům	29591.9	0.411	0.9400
stůl	11730.4	0.375	0.9772
měsíc	18996.9	0.738	0.9852
vzduch	8805.4	0.374	0.9268
byt	10488.9	0.456	0.9221

Aplikace modelu na lemma „člověk“



Aplikace modelu na lemma „dům“



Výsledky aplikace modelu na data (fem.)

fem. anim.

lemma	*a*	*b*	R^2
žena	46525.9	0.723	0.9664
matka	32093.7	0.772	0.9108
paní	56238.4	1.247	0.9667
dívka	17406.2	0.889	0.9771
dcera	8179.4	0.471	0.9757

fem. inanim.

lemma	*a*	*b*	R^2
hlava	32671.4	0.380	0.9046
ruka	25636.7	0.385	0.9366
škola	25894.7	0.506	0.9778
ulice	31193.3	0.772	0.9802
tvář	11521.9	0.308	0.8232

Výsledky aplikace modelu na data (neut.)

neut. anim.

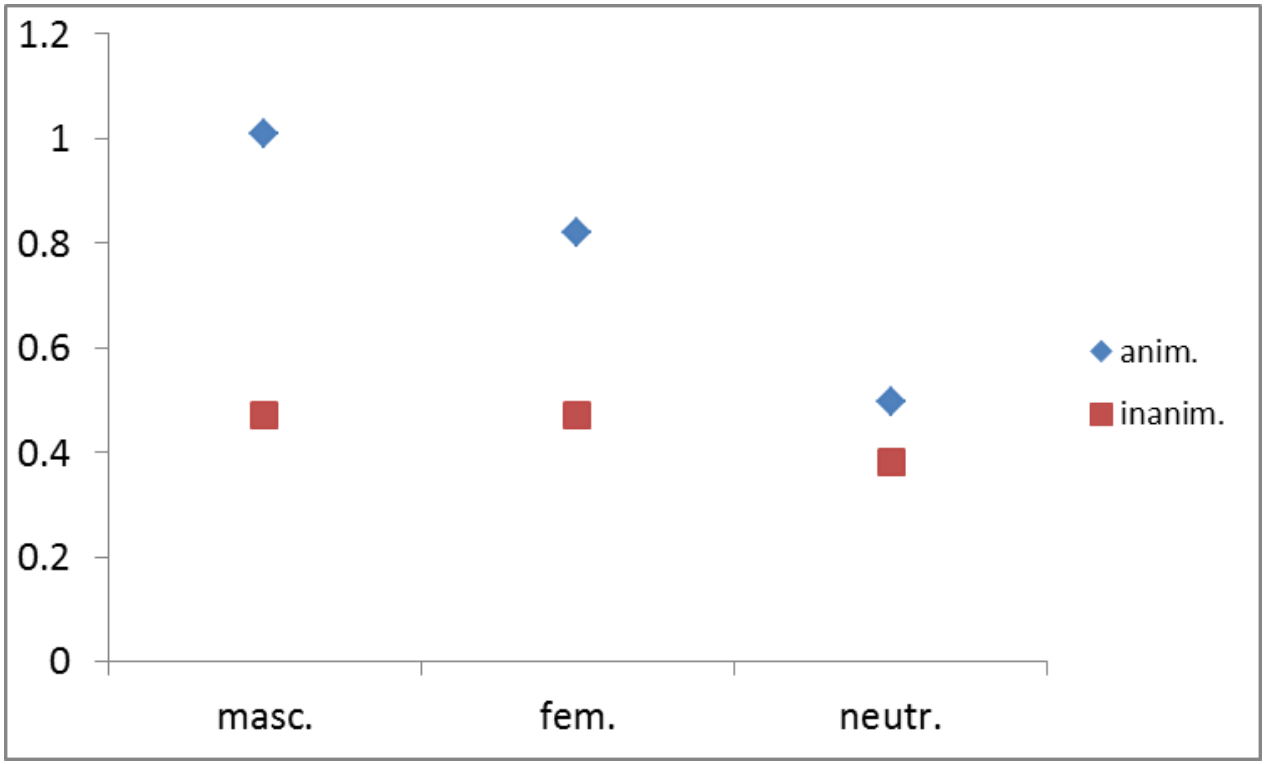
lemma	*a*	*b*	R^2
dítě	19762.8	0.492	0.9743
děvče	2220.8	0.596	0.9281
miminko	1135.3	0.403	0.8960
děcko	587.2	0.430	0.9574
děťátko	691.4	0.555	0.9854

neut. inanim.

lemma	*a*	*b*	R^2
město	53158.3	0.515	0.9819
tělo	15578.0	0.369	0.9673
auto	10164.0	0.299	0.8910
divadlo	11608.6	0.423	0.9291
srdce	7553.4	0.295	0.9777

Průměry parametru *b*

rod	anim.	inanim.
mask.	1.010	0.471
fem.	0.821	0.471
neut.	0.495	0.380



Závěry

- nominativ je nejfrekventovanějším pádem pro anim. bez ohledu na rod
- předběžně lze tvrdit, že parametr *b* je interpretovatelný ve smyslu hypotézy
- čím je křivka strmější (tj. čím je vyšší hodnota *b*), tím významnější role pádu s nejvyšší frekvencí
- anim. substantiva se diverzifikují silněji než inanim.
- nezjistili jsme téměř žádné rozdíly mezi jednotlivými rody u inanim. substantiv