Radek Čech, Emmerich Kelih, Jan Mačutek

Impact of semantics on case diversification

Qualico 2014 (**May 29 - June 1, 2014)**
20th anniversary of
IQLA and Journal of Quantitative Linguistics (JQL)
Palacký University Olomouc (Czech Republic)

Distribution of cases vs. semantics

- Why (and how) should semantics influence a frequency distribution of noun cases?
- intuitive assumptions
  - animate nouns should typically represent agentive semantic role
  - e.g. in Czech, agentive role expressed as a syntactic subject, which is usually nominative
  - a morpho-syntactic status of inanimate nouns is not so clear which should be reflected by different characteristics of case distribution

Masculine anim. vs. inanim nouns in Czech

Anim. sg. (Czech SYN2010)

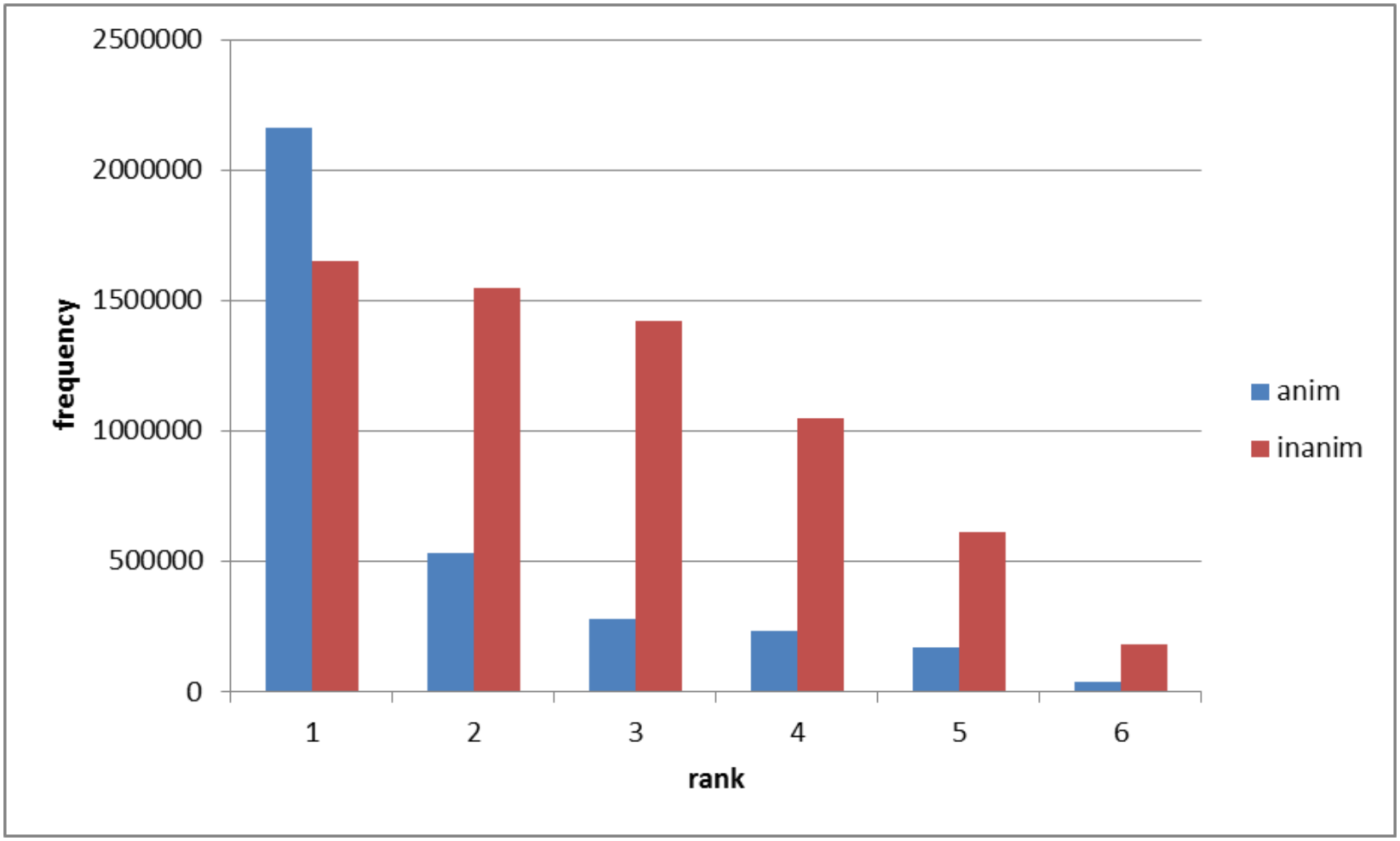
case	frequency
nom.	2161013
gen.	532579
acc.	278806
instr.	233327
dat.	170042
loc.	39956

Inanim. sg. (Czech SYN2010)

case	frequency
gen.	1649641
acc.	1546412
nom.	1422769
loc.	1045981
instr.	613918
dat.	184674

N … nominative
G … genitive
D … dative
A … accusative
L … locative
I … instrumental

Masculine anim. vs. inanim nouns in Czech



Distribution of cases vs. semantics

- Is there an impact of semantics on the distribution of cases?
- If so, what are general principles which influence it?
- What are the boundary conditions?
  - grammatical gender
  - grammatical number
  - polysemy etc.

Theoretical assumptions

- distribution of cases as a result of a diversification process
- diversification (generally)
  - one entity (e.g. word) – properties (gender, number etc.) – different categories (m., f., n.)
  - if an entity diversifies, the frequencies of the resulting categories are not uniformly distributed
  - general phenomenon, a characterization of a system

Focus of our study

- diversification of cases for particular nouns
- diversification is influenced by semantic features
- are there differences among grammatical genders?

Hypotheses

- distributions of cases of both animate and inanimate nouns (for all 3 genders) can be modelled by the same mathematical function
- parameters of the function differ significantly between animate and inanimate nouns and among genders differ significantly

Language material

- data taken from Czech synchronic corpus SYN2010
  - 100 mil. tokens
  - morphologically tagged
  - written language
    - journalistic texts (33%)
    - fiction (40%)
    - scientific texts (27%)

Language material

- five most frequent lemmas for animate and inanimate nouns
- together 10 for each gender (masc., fem., neut.)
- 30 nouns observed totally
- concrete nouns
- no proper names
- singular only

Language material

masc. anim.	masc. inanim	fem. anim.	fem. inanim.	neuter. anim.	neuter. inanim.
člověk (human being)	dům (house)	žena (woman)	hlava (head)	dítě (child)	město (town)
muž (man)	stůl (table)	matka (mother)	ruka (hand)	děvče (girl)	tělo (body)
pan (Mr.)	měsíc (month)	paní (Mrs.)	škola (school)	miminko (baby)	auto (car)
otec (father)	vzduch (air)	dívka (girl)	ulice (street)	děcko (kid)	divadlo (theatre)
ředitel (director)	byt (flat)	dcera (daughter)	tvář (face)	děťátko (baby)	srdce (heart)

Results - rank of cases

rank	masc. anim.					masc. inanim					fem. anim.					fem. inanim.					neuter. anim.					neuter. 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 … nominative
G … genitive
D … dative
A … accusative
L … locative
I … instrumental

Results

(example)

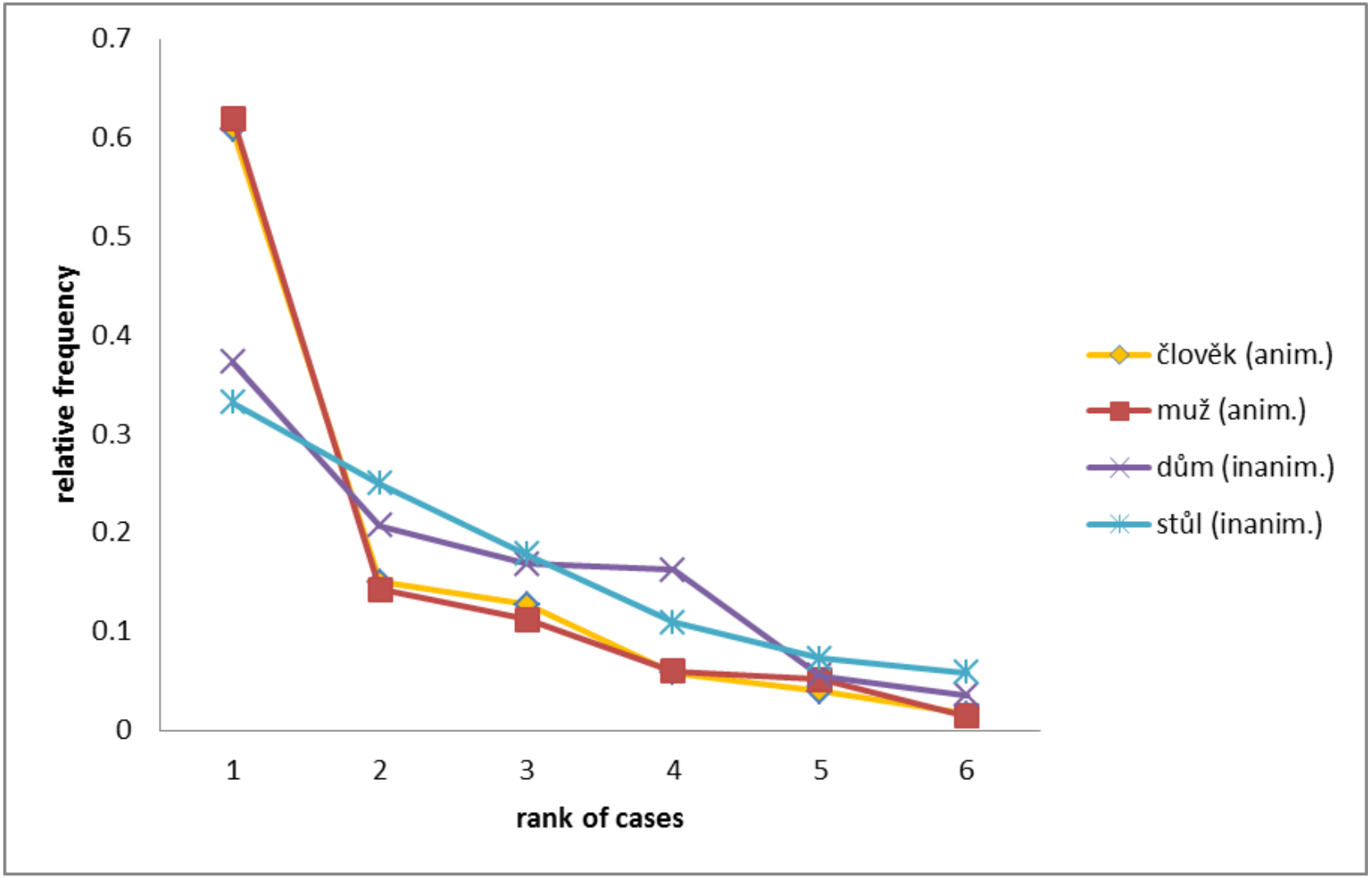
lemma *člověk* (human being)

case	frequency
nom.	43543
gen.	10732
acc.	9128
dat.	4186
instr.	2833
loc.	1194

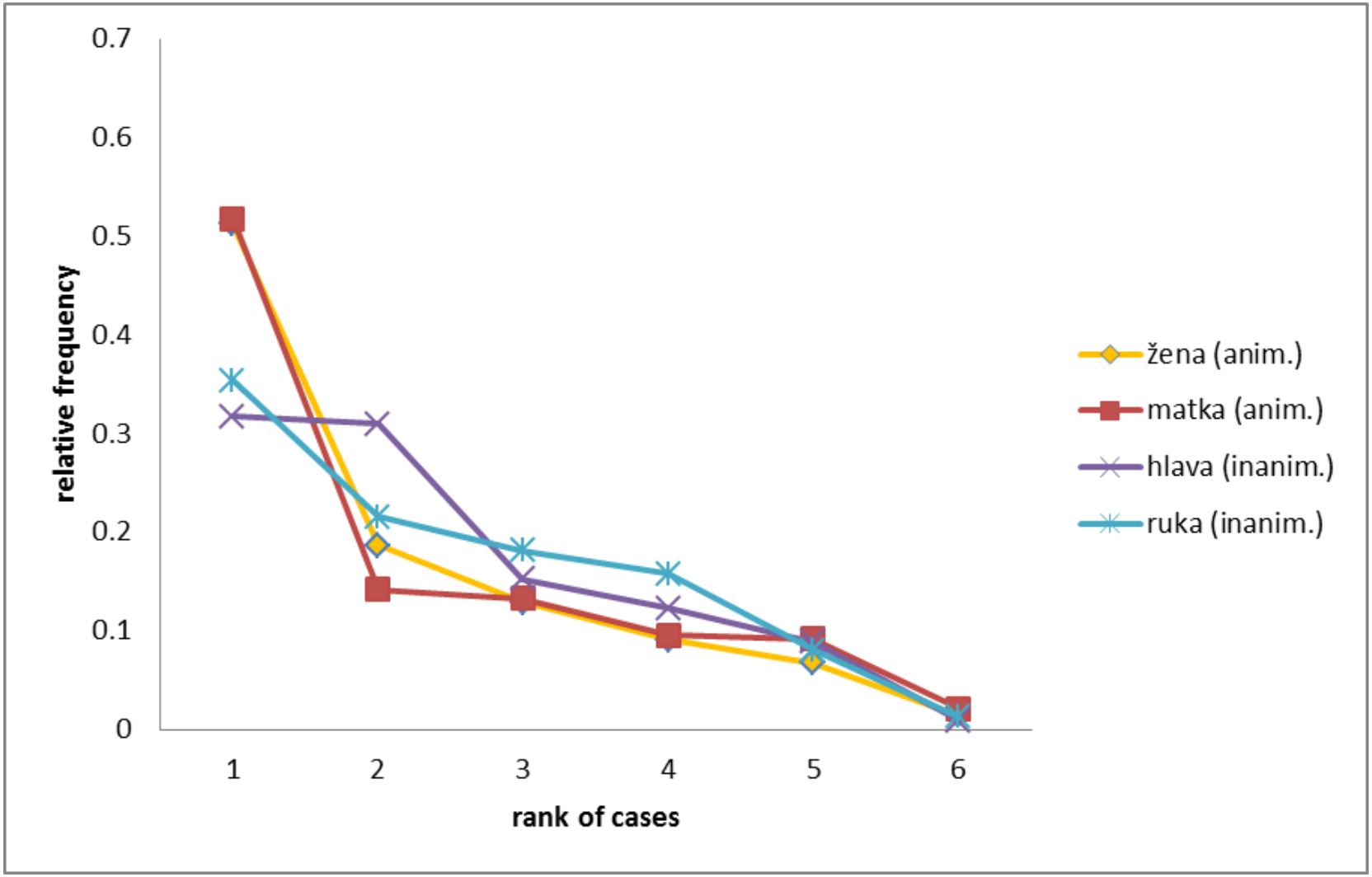
lemma *dům* (house)

case	frequency
gen.	20145
loc.	11194
acc.	9109
nom.	8777
inst.	2966
dat.	1900

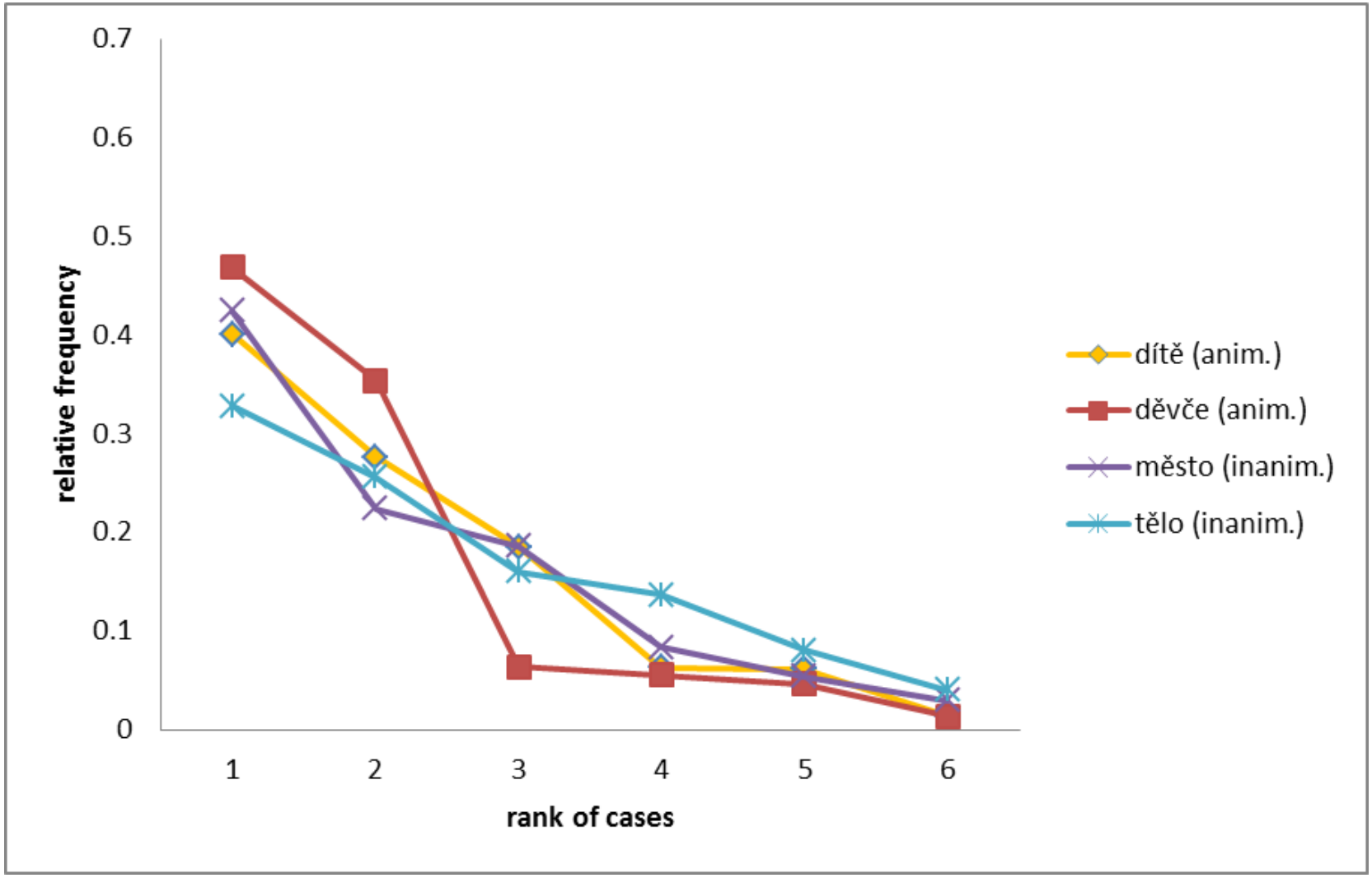
Rank-frequency distribution (masculine)



Rank-frequency distribution (feminine)



Rank-frequency distribution (neuter)



Model

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

x … rank of the case

y … frequency of the case

a, b … parameters

- a special case of the general Wimmer-Altmann model (2005)

Results of fitting - masculine

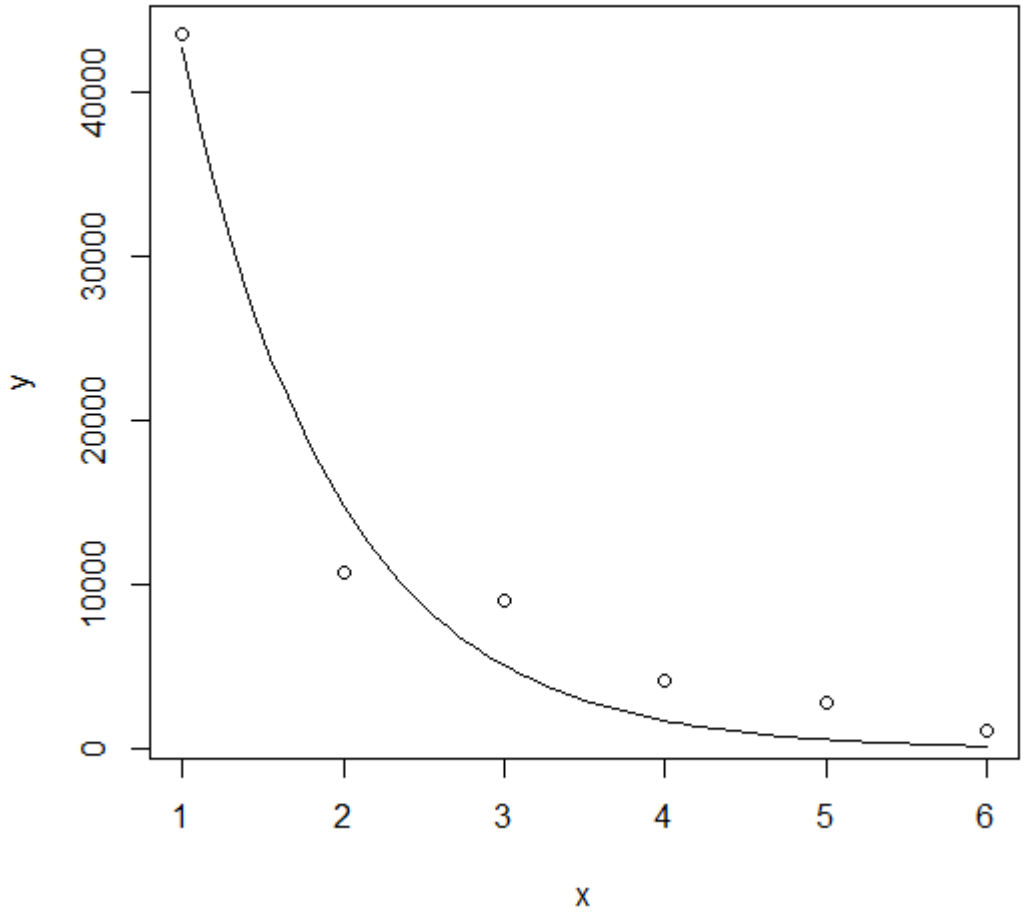
masc. anim.

lemma	*a*	*b*	R^2
člověk (human)	123208.0	1.059	0.9723
muž (man)	91962.4	1.149	0.9711
pan (Mr.)	39832.5	0.685	0.9887
otec (father)	48060.8	1.014	0.9476
ředitel (director)	46335.5	1.142	0.9950

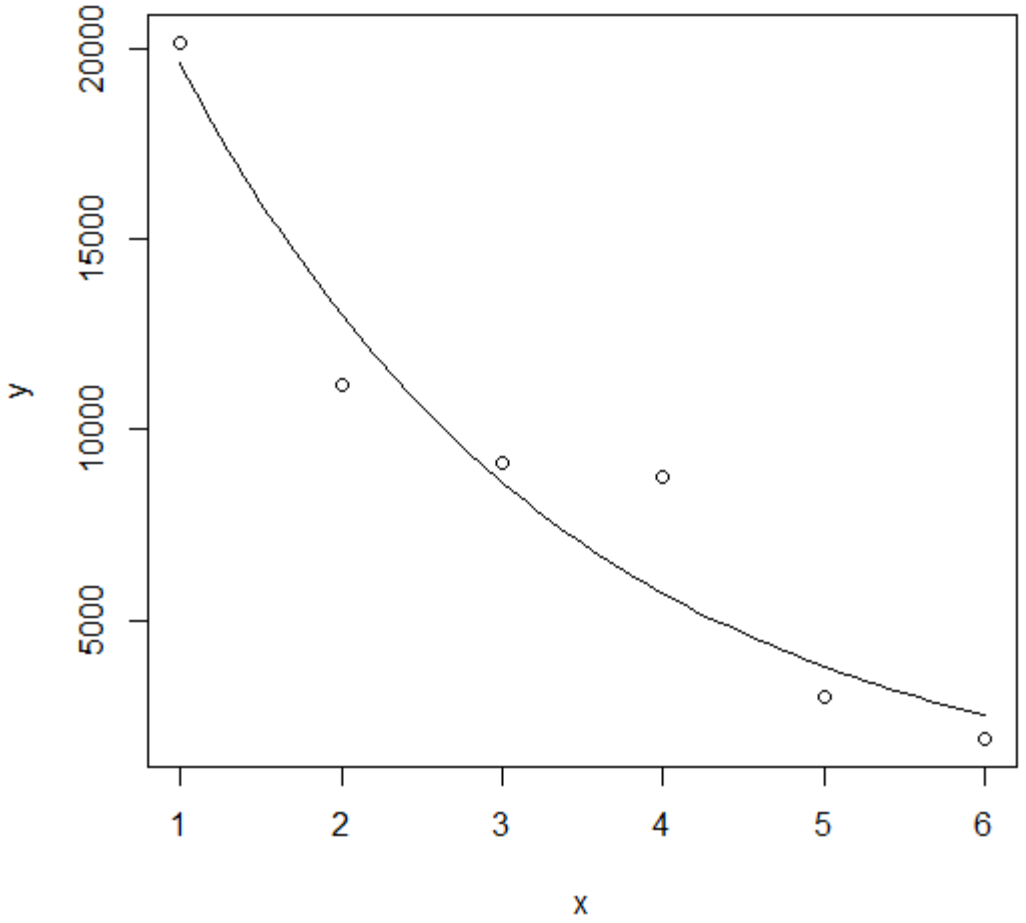
masc. inanim.

lemma	*a*	*b*	R^2
dům (house)	29591.9	0.411	0.9400
stůl (table)	11730.4	0.375	0.9772
měsíc (month)	18996.9	0.738	0.9852
vzduch (air)	8805.4	0.374	0.9268
byt (flat)	10488.9	0.456	0.9221

Fitting the model to lemma „člověk“ (human being)



Fitting the model to lemma „dům“ (house)



Results of fitting - feminine

fem. anim.

lemma	*a*	*b*	R^2
žena (woman)	46525.9	0.723	0.9664
matka (mother)	32093.7	0.772	0.9108
paní (Mrs.)	56238.4	1.247	0.9667
dívka (girl)	17406.2	0.889	0.9771
dcera (daughter)	8179.4	0.471	0.9757

fem. inanim.

lemma	*a*	*b*	R^2
hlava (head)	32671.4	0.380	0.9046
ruka (hand)	25636.7	0.385	0.9366
škola (school)	25894.7	0.506	0.9778
ulice (street)	31193.3	0.772	0.9802
tvář (face)	11521.9	0.308	0.8232

Results of fitting - neuter

neut. anim.

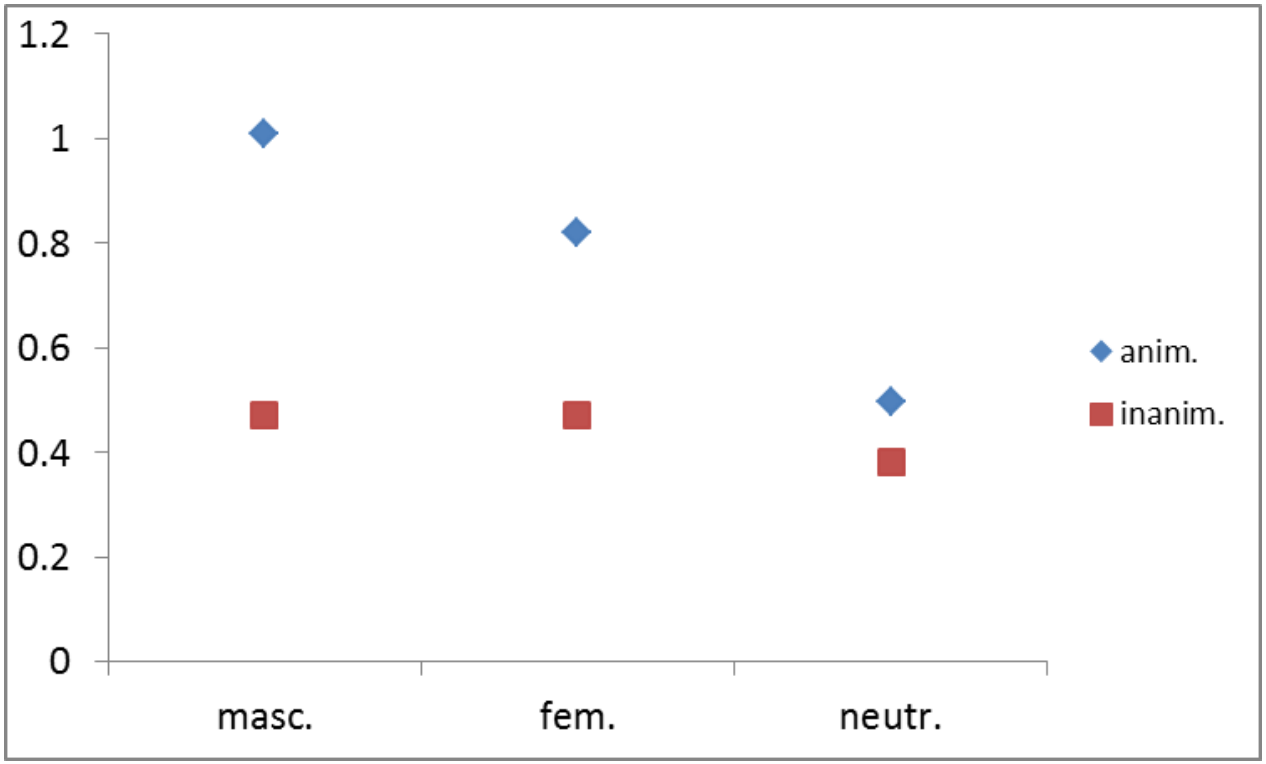
lemma	*a*	*b*	R^2
dítě (child)	19762.8	0.492	0.9743
děvče (girl)	2220.8	0.596	0.9281
miminko (baby)	1135.3	0.403	0.8960
děcko (kid)	587.2	0.430	0.9574
děťátko (baby)	691.4	0.555	0.9854

neut. inanim.

lemma	*a*	*b*	R^2
město (town)	53158.3	0.515	0.9819
tělo (body)	15578.0	0.369	0.9673
auto (car)	10164.0	0.299	0.8910
divadlo (theatre)	11608.6	0.423	0.9291
srdce (heart)	7553.4	0.295	0.9777

Means of parameter *b*

gender	anim.	inanim.
masc.	1.010	0.471
fem.	0.821	0.471
neutr.	0.495	0.380



Conclusion

- nominative is the most frequent case for animate nouns regardless of gender
- preliminary interpretation of parameter *b* is possible
- information about the steepness
- the steeper the curve (the higher the parameter *b*), the higher the exploitation of the most frequent case
- high case diversification of animate

Conclusion

- almost no differences among genders for inanimate nouns
- as nominative indicate agentive role of the noun, masculine animate nouns take this role with a higher probability than the other nouns