Institut für Slawistik

universität wien

Emmerich Kelih

Parts of speech – theoretical problems, empirical tendencies and modelling

Qualico 2018 – Wrocław

IQLA

International Quantitative Linguistics Association

Aims of the study

- General remarks about parts of speech (POS)
- Quantitative contributions - POS frequency studies
- Empirical part: Analysing POS-f in Slovene texts
- Theoretical modelling
- Summary

General remarks about *parts of speech*, *word classes, lexical categories …*

nouns	boy, machine, beauty
pronouns	I, me, you
adjectives	happy, three, both
verbs	go, frighten, be
prepositions	in, under, with
conjuctions	and, because, if
adverbs	happily, soon, often
interjections	ooh, ooops, ahh

- vagueness of definitions
- shift from meaning-based definitions to **morphological** and **syntactical properties** of the words

DET _____
DET ADJ ____

____ + tense morpheme
____ + third person singular morpheme
____ + progressive morpheme
AUX ____

- reduction of words to a smaller subset of word classes, based on semantical, syntactical and morphological properties

– where POS play a role?

- language acquisition
- cognitive linguistics (mental lexicon)
- corpus linguistics (availability of data): various synchronous and diachronic aspect
- Sociolinguistics – language contact (borrowing hierarchies)
- crucial role in authorship attribution and stylometry

More specific aspects in QL

1. **POS-frequencies**
2. calculation of ratios/indices (V/N, ADJ/V) etc.
3. **theoretical modeling**

POS-frequencies

Hudson, R. (1994). About 37% of all word-tokens are nouns. *Language 70, 331–339.*

	Number of words		% of nouns	
Genre	Brown	LOB	Brown	LOB
A Press: reportage:	88690	89138	42.2	41.2
B Press: editorial:	54505	54447	36.1	35.8
C Press: reviews:	35346	34321	37.0	37.7
D Religion:	34590	34387	34.8	34.9
E Skills, trades, hobbies:	72590	76913	37.2	35.4
F Popular lore:	97223	89090	35.7	35.9
G Belles lettres, biography, essays:	152064	155336	35.5	34.4
H Miscellaneous informational:	62477	60761	37.9	34.5
J Learned & scientific writings:	162211	161900	35.0	33.3
K General fiction:	58380	59204	36.7	35.8
L Mystery and detective:	48208	49145	36.7	35.5
M Science fiction:	12042	12119	35.5	36.3
N Adventure & western:	58416	59391	37.3	36.4
P Romance & love story:	58625	59382	36.5	36.1
R Humor:	18277	18203	36.7	35.6

Table 3. Total words and noun percentage for 15 genres in the Brown and LOB corpora.

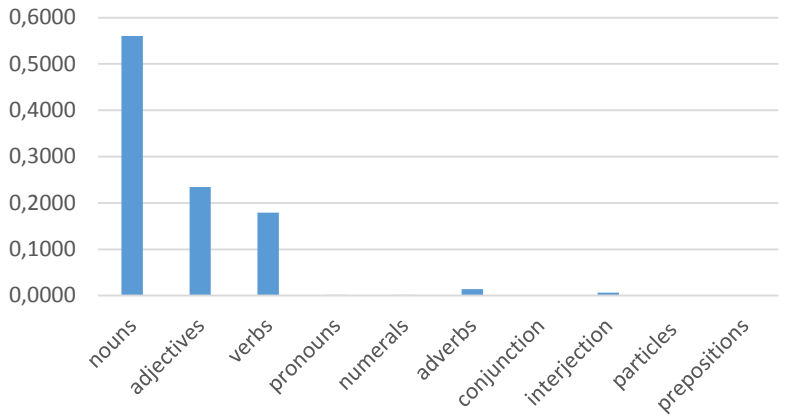
- the overall percentage for all nouns varies between 33% and 42%.
- statistically they (differences) are extremely significant

"At present we cannot explain these regularities, but they are a challenge that our grandchildren may (possibly) be able to meet." (Hudson 1994: 338)

POS-frequencies in dictionaries

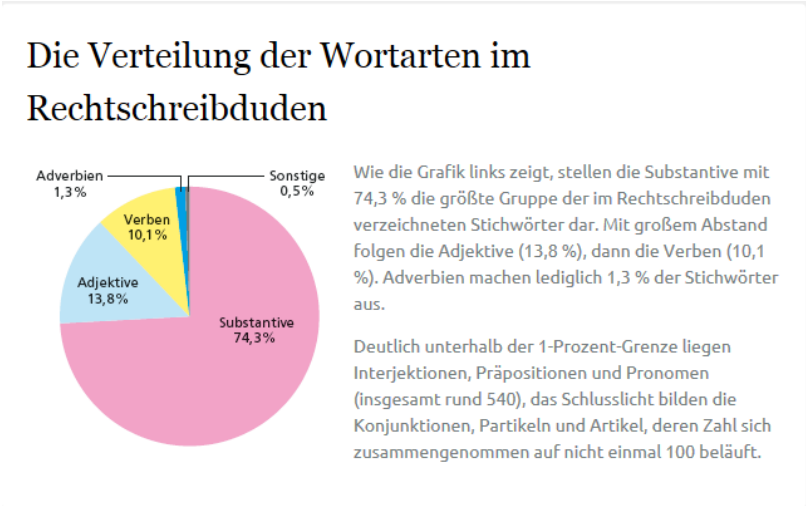
monolingual dictionary: Slovene

POS	abs.	%
nouns	51448	0,56
adjectives	21516	0,23
verbs	16479	0,18
pronouns	130	0,00
numerals	122	0,00
adverbs	1325	0,01
conjunction	76	0,00
interjection	615	0,01
particles	9	0,00
prepositions	115	0,00



Jakopin, Primož (1995): Nekaj številk iz Slovarja Slovenskega Knjižnega Jezika. In: *Slavistična revija* 43 (3), 341–375.

POS-frequencies in German orthographic dictionary (Duden)



https://www.duden.de/sprachwissen/sprachratgeber/Die-Verteilung-der-Wortarten-im-Rechtschreibduden

What is well known is the variation found in different genres/text types

Case (pilot) study on POS-frequencies in Slovene

a. 10 NP - newspaper articles (*Delo*, *Dnevnik*, short notes on daily events)
b. 10 CP - cooking recipes (description how to prepare main meals and sweets)
c. 10 DI - dialogues from different speakers, written by Drago Jančar
d. 10 SO - sonnets written by Milan Jesih

- spectrum of different text types
- texts are approximattely equally long (every corpus < 5000 T)

A-priori ideas/theses?

- higher amount of nouns in NP and CP („description“)
- verbs should have an important role in NP (something is happening, going on etc.)
- no particular ideas about POS-f in sonnets („poetic licence“)
- for dialogues “heterogeneity” is expected

Which classification is used? (Toporišič 2000)

1. NOUNS	– nom
2. VERBS	– verb (inkl. Participles, auxiliaries)
3. ADJECTIVES	– adj (inkl. numerals, possesive and demonstrative pronouns)
4. ADVERBS	– adv
5. PRON	– pron
6. PARTICLES	– part (inkl. interjections, fillers, discourse markers)
7. PREPOSITIONS	– praep (closed set)
8. CONJUNCTIONS	– conj (closed set)

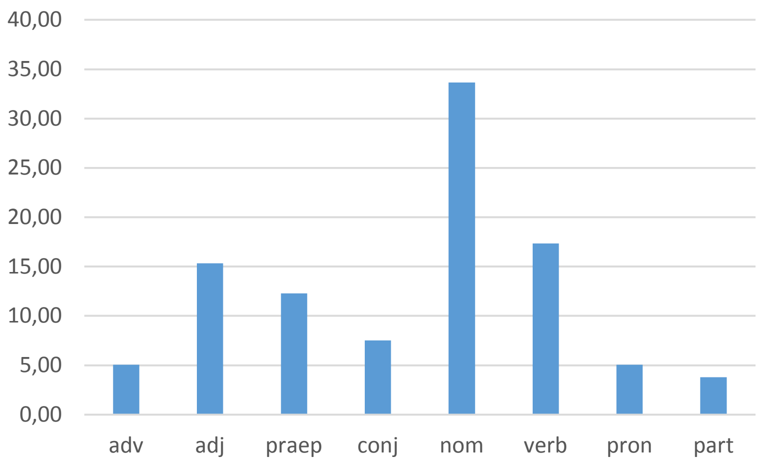
manual tagging

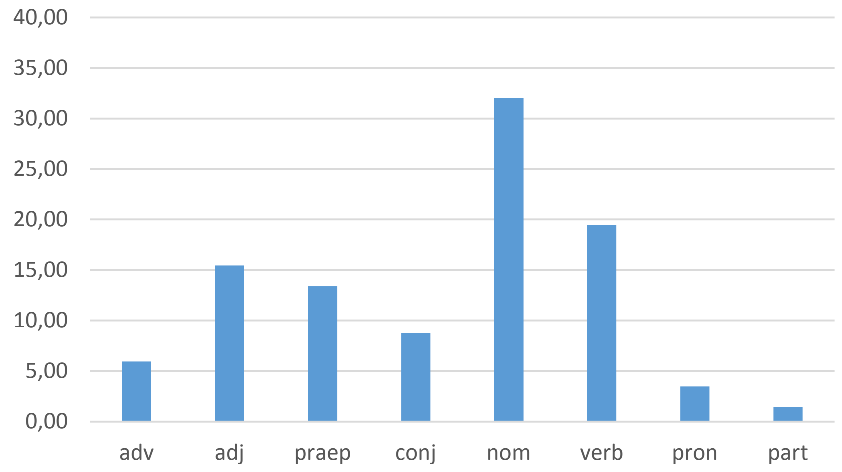
„Karte so na mizi. Igra še ni končana," je dejal predstavnik poljske vlade, ko je komentiral sestanek premierov desetih držav kandidatk za vstop v Evropsko unijo. (NP – text 1)

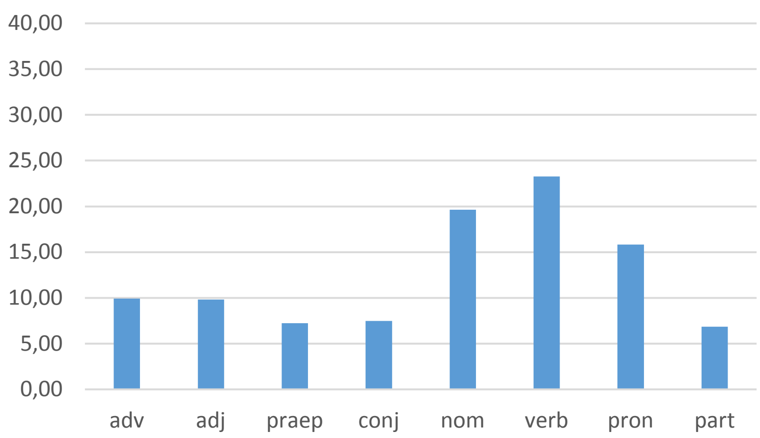
„nom verb praep nom. Nom part part adj“, verb verb nom adj nom, conj verb verb nom nom adj nom nom nom praep non paep nom praep adj nom.

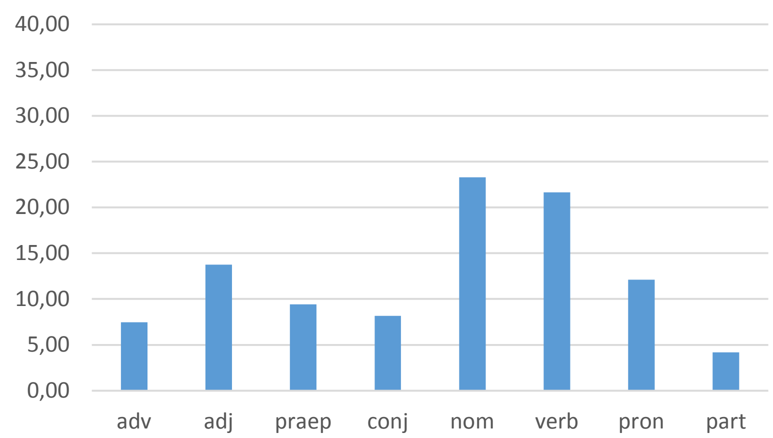
! Whole texts are analysed (every token)

first empirical results – POS-f (%)

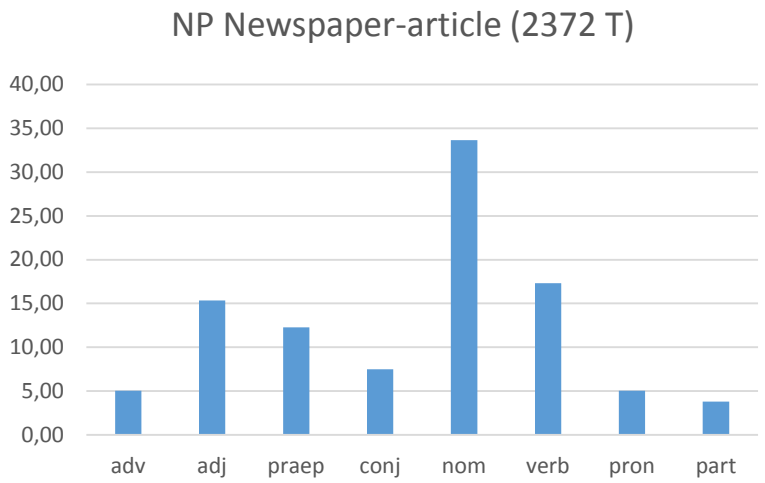


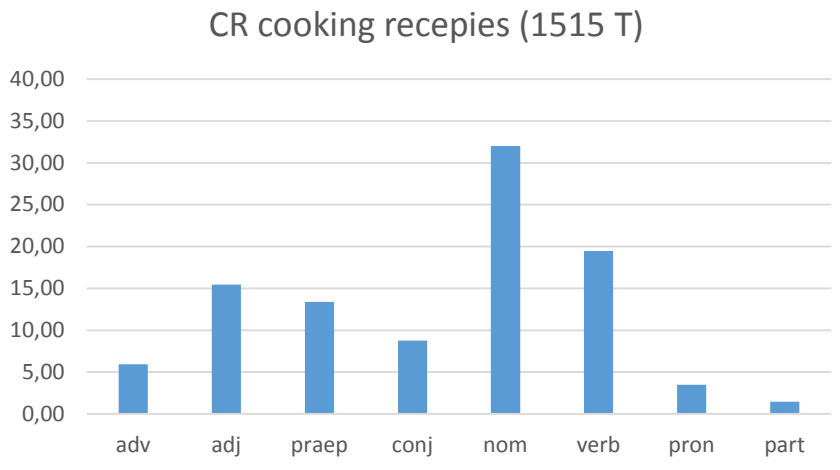


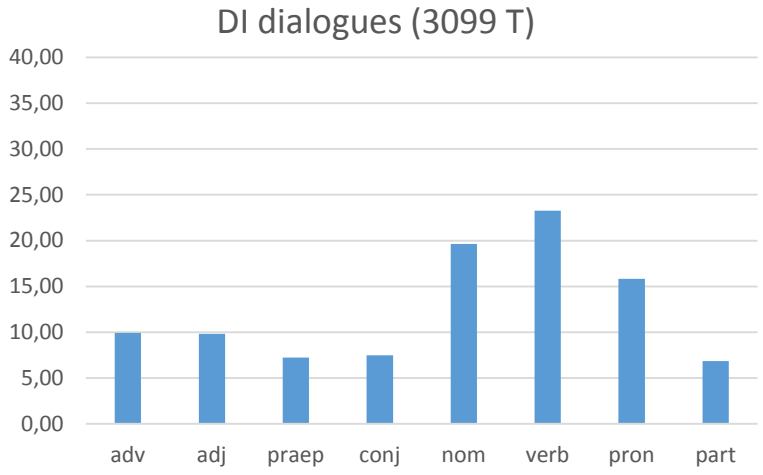


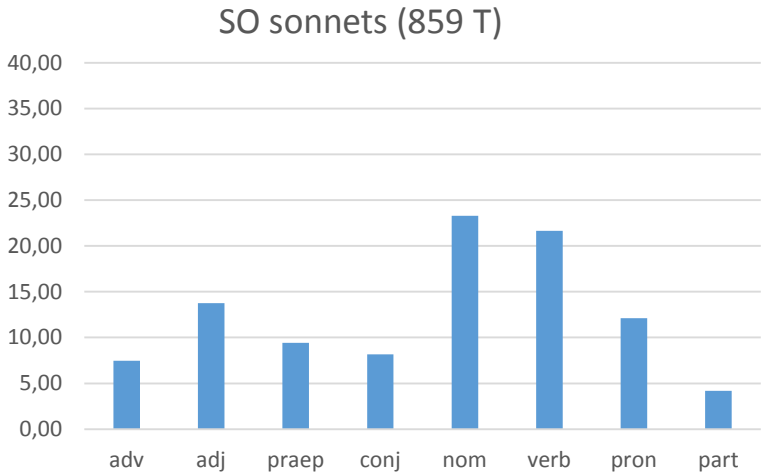


first empirical results – POS-f (%)

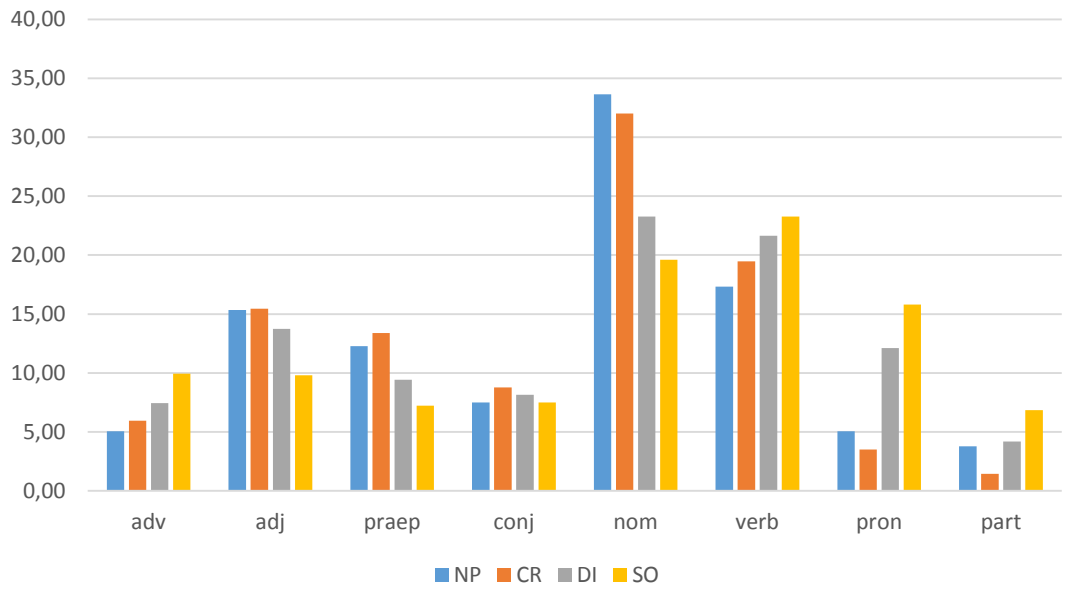








first empirical results – POS-f (%)



- nom: most frequent?
- nom + verb > 50%
- conj: more or less stable
- pron. and part: high amount in DI and SO

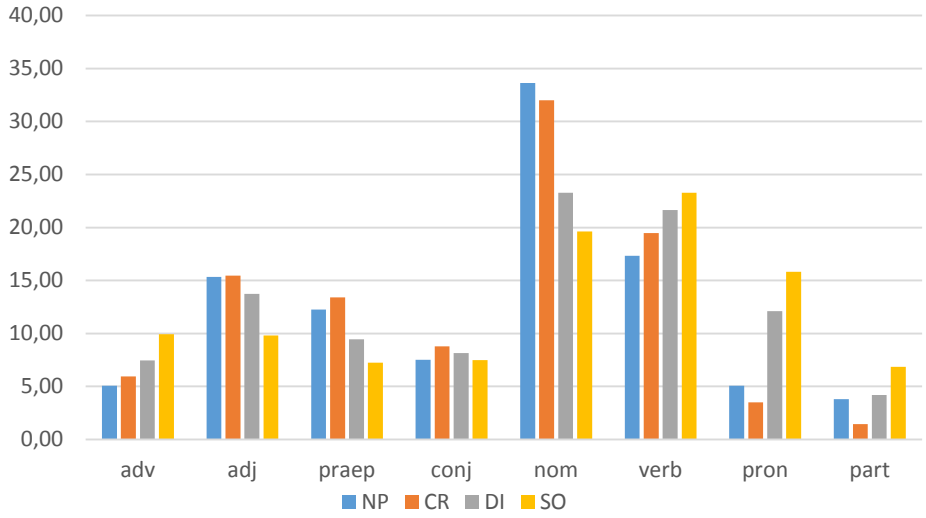
	NP	CR	DI	SO
NP	1	**0,9739**	0,03495839	**0,51396673**
CR		1	0,00857688	**0,28609693**
DI			1	**0,90036474**
SO				1

Kendall's tau_b

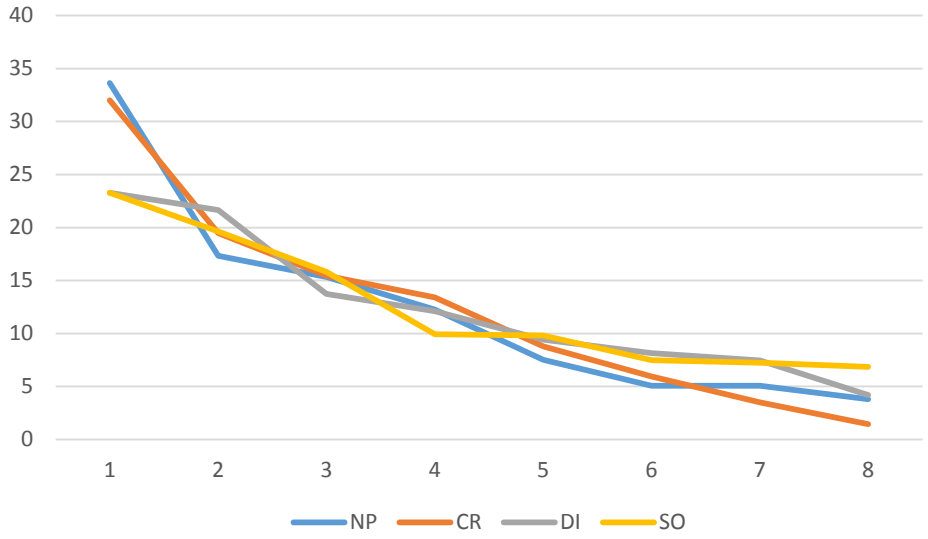
NP and CR, DI and SO: high similarity

NP – DI, CR – DI: less similarity

Modeling: From frequencies to rank-frequency distribution

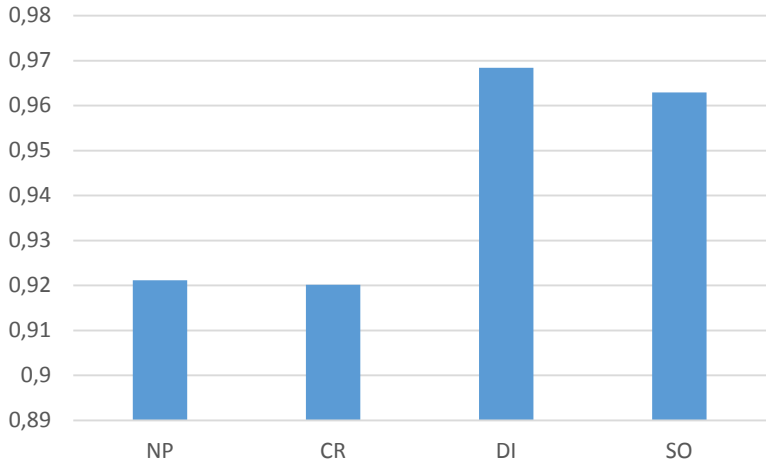


Transforming to a rank-frequency distribution



- Similarity of NP and CR vs. DI and SO
- remarkable characteristic: „uneven“ distribution

Calculating the relative repeat rate for the corpora



$$R = \frac{1}{N^2} \sum_{i=1}^{P} f_i^{\,2}$$

$$R_{rel} = \frac{1 - R}{1 - 1/N}$$

- R_{rel} = 1 (quite impossible, only in case all POS equally distributed)
- low R_{rel}: particular “over-exploitation”, higher functional load of particular POS categories
- high R_{rel}: more “even” distribution of POS
- linguistic meaning? Complexity of the morphosyntactical organization (and even of higher levels?)

And finally: what about a theoretical model?

discrete distributions

negative hyper-geometric d. (Best 2001)

1-displaced mixed Poisson as well as the 1-displaced hypergeometric distribution (Ziegler 1998)

continuous distributions

Modifications of Zipf's law (Hammerl 1991)

Tuzzi, A.; Popescu, I.-I.; Altmann, G. (2010): Quantitative Analysis of Italian Texts. Lüdenscheid: Ram-Verlag.

"Zipf´s approach is rather restricted to word frequencies"

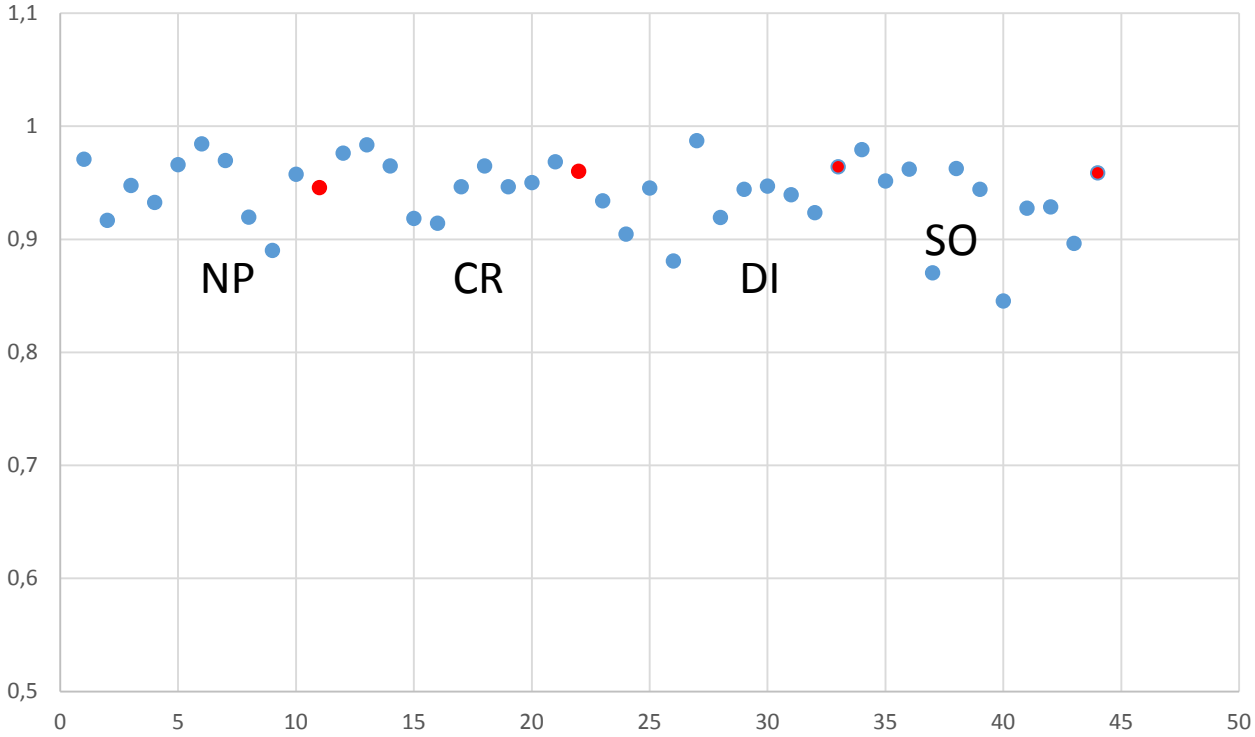
alternative exponential model:

$$y = 1 + a0 * e^{(-a1x)}$$

Popescu, Ioan-Iovitz; Altmann, Gabriel; Köhler, Reinhard (2010): Zipf´s law another view. In: *Quality and Quantity* 44 (4), S. 713–731.

And finally: The results

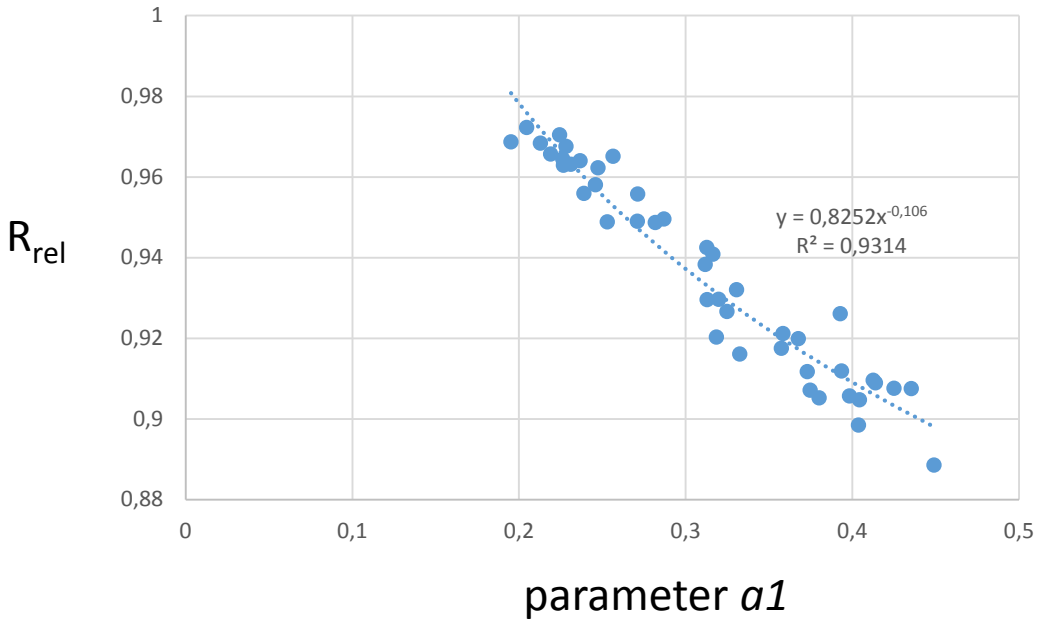
R^2 (coefficient of determination) for all 40 texts, plus for 4 corpora > 0.9



NP (1), CR (0), DI (1), SO (3) R^2 < 0,9

- overall results are satisfying
- one common model for all text types

And now really, finally: Interpretation of Parameter



- strong interrelation between parameter a1 and R_{rel}

  „The higher parameter a1, the lower the R_{rel}.“

- interpretable parameter in the used model

Summary – perspectives

- POS-frequencies give information about morphological, syntactical and lexical organisation of texts

- vagueness of definitions is not part of a “problem”, but rather an reasonable property for text interpretation

- POS-frequencies are regularly distributed, homogenous/heterogenous behaviour in regard to particular text types

- one common model for all types of texts
- interpretable parameter
- proposed layer model has to be tested empirically in other texts/languages
- different levels has to be tested: word frequencies – word classes – syntactical categories – morphological categories
- a synergetic interpretation of POS-f, relation to other features

References:

Best, Karl-Heinz (2001): Zur Gesetzmäßigkeit der Wortartenverteilungen in deutschen Pressetexten. In: *Glottometrics* 1, S. 1–26.

Hammerl, Rolf (1991): Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells. Trier: Wissenschaftlicher Verlag Trier.

Hudson, R. (1994): About 37% of all word-tokens are nouns. In: *Language* 70, S. 331–339.

Ito, Masamitsu (2005): Quantitative linguistics in Japan. In: Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski (Hg.): Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27), 82–95

Liang, Junying; Liu, Haitao (2013): Noun distribution in natural languages. In: *Poznan Studies in Contemporary Linguistics* 49, S. 509–529.

Popescu, Ioan-Iovitz; Altmann, Gabriel; Köhler, Reinhard (2010): Zipf´s law another view. In: *Quality and Quantity* 44 (4), S. 713–731.

Tuldava, Juhan (1998): Probleme und Methoden der quantitativ-systemischen Analyse. Trier: Wissenschaftlicher Verlag Trier (Quantitative Linguistics, 59).

Tuzzi, Arjuna; Popescu, Ioan-Iovitz; Altmann, Gabriel (2010): Quantitative Analysis of Italian Texts. Lüdenscheid: Ram-Verlag (Studies in Quantitative Linguistics, 6).

correlations – interrelations

Ito, Masamitsu (2005): Quantitative linguistics in Japan. In: Reinhard Köhler, Gabriel Altmann und Rajmund G. Piotrowski (Hg.): Quantitative Linguistik. Quantitative Linguistics. Ein internationales Handbuch. An International Handbook. Berlin, New York: de Gruyter (Handbücher zur Sprach- und Kommunikationswissenschaft, 27), 82–95.

> "Kabashima demonstrated the existence of a linear dependence between the occurrence of nouns and other parts of speech in texts by the following functions". (Ito 2005: 86)

> „Word class frequencies depend on their position in a rank-frequencey of lexemes/word" (Tuldava 1998: 112ff, Liang/Liu 2013)