



universität
wien

Emmerich Kelih

Institut für Slawistik

How plausible is the hypothesis that the population size is related to the phoneme inventory size?

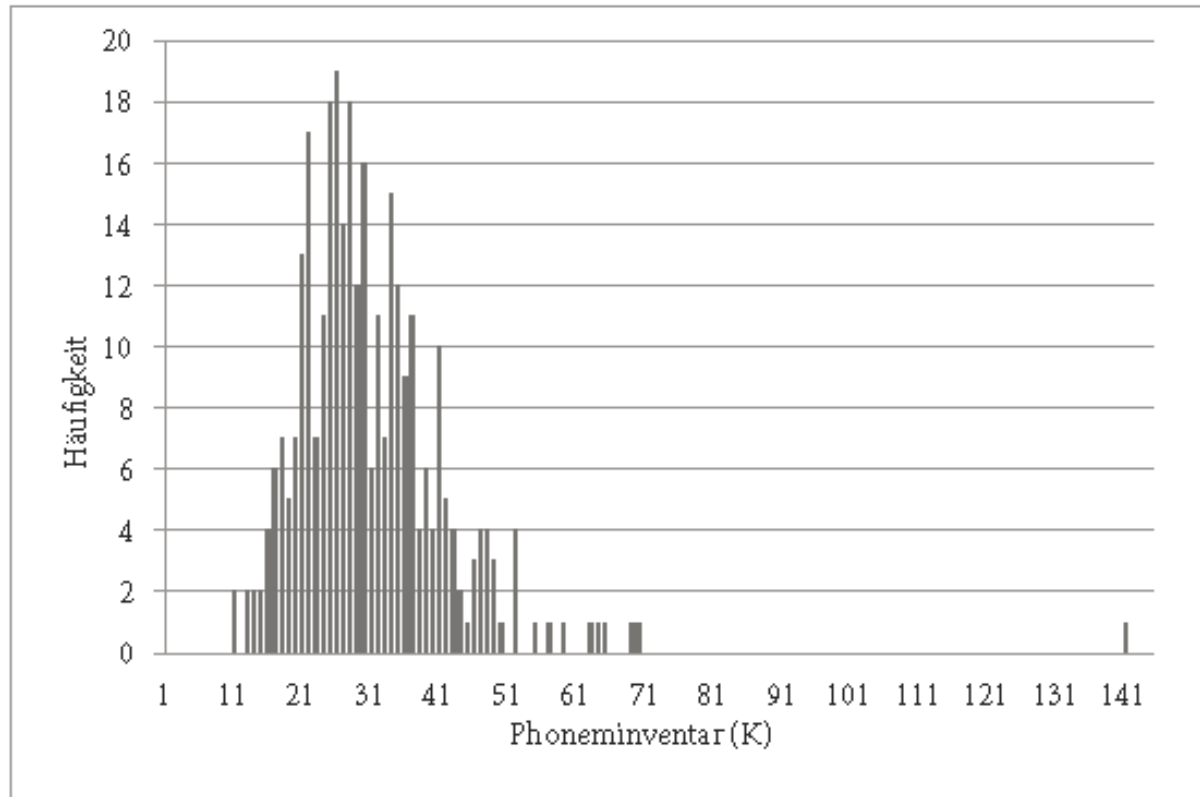
Comment's from a quantitative linguistics point of view

International Quantitative Linguistics
Conference (*QUALICO*) 2016
Trier, 24. – 26th, august 2016

Overview:

1. Recent discussion: Phoneme inventory size – population size
2. Empirical results & counterevidence
3. Phoneme inventory size from a synergetic point of view
4. Phoneme inventory size – repeat rate (phoneme frequencies) – phoneme combinations
5. Phoneme inventory size – word length
6. Discussion of obtained results from Slavic languages
7. Summary – Perspectives

Diversification of the phoneme inventory size



Maddieson (1984)

Analysis of over 320 languages of the world

Mean = 30,99 phonemes

Standard deviation = 11,73

Minimum = 11 phonemes

Maximum = 141 phonemes

- Segmental phonemes
- Other determination of phoneme size inventories possible (what about suprasegmental features?)

1. Recent discussion: Phoneme inventory size – population size

Hay/Bauer (2007), Atkinson (2011) published in *Science* and *Language*

Phoneme inventory size is correlated with the population size:

„The higher the population size, the higher the phoneme inventory size“. (Hay/Bauer 2007: 398)

- authors are “surprised” by the results
- empirical counterevidence found by Pericliev (2004), based on UPSID-database
- Donohue/Nichols (2011), based on 1350 languages: → **no statistical significant correlation between phoneme inventory size and population size;**

Is there a theoretical background for this hypothesis?

1. Recent discussion: Phoneme inventory size – population size

Some suggestions, ideas & problems ...

What are the causes of „extreme“ phoneme inventory sizes (small/big?)

- Social isolation of language communities;
- changes of the phoneme inventory size due to language contact:
 - “The more isolated a language, the lower the phoneme inventory size”.
- High amount of common information and therefore a low amount of redundancy
 - = The role and regulation of „**redundancy**“ in the human communication.

but:

- isolated language communities can also have a rather big inventory size (Caucasus, tribe languages, indigenous languages of the Americas);
- No evidence known for a correlation between areal or genetic affiliation of languages and the phoneme inventory size;
- Changes of phoneme inventory sizes are only partly caused by the “language-external” factors like language contact;
- Loss of phonemes in a language system = loss of their functionality to differentiate meanings;
- „**distinctiveness**“ has to be preserved in a language system.

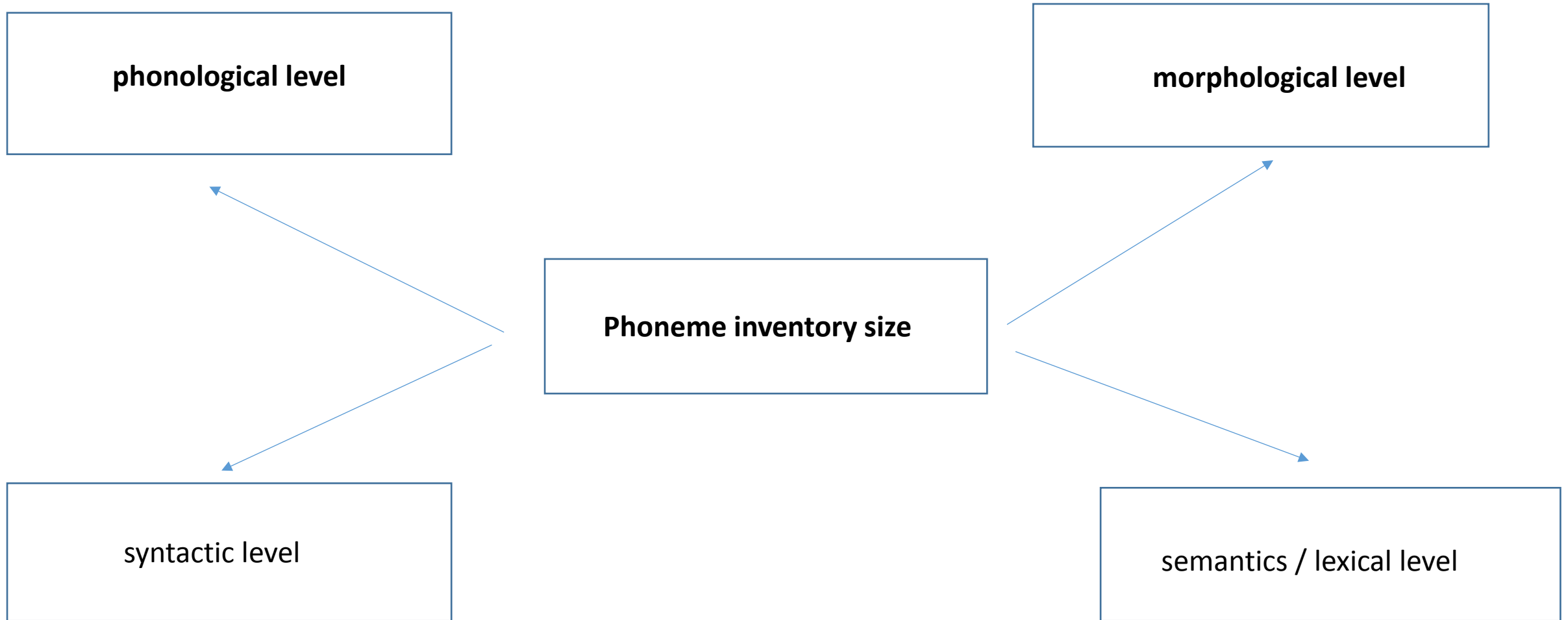
Recent discussion: Phoneme inventory size – population size

Intermediate result:

“Despite the new-found relative ease of testing world-wide hypotheses, a basic requirement of scientific inquiry must not be neglected: hypotheses must have some inherent plausibility in terms of their fit with known causal mechanisms.”
(Bybee 2011: 147)

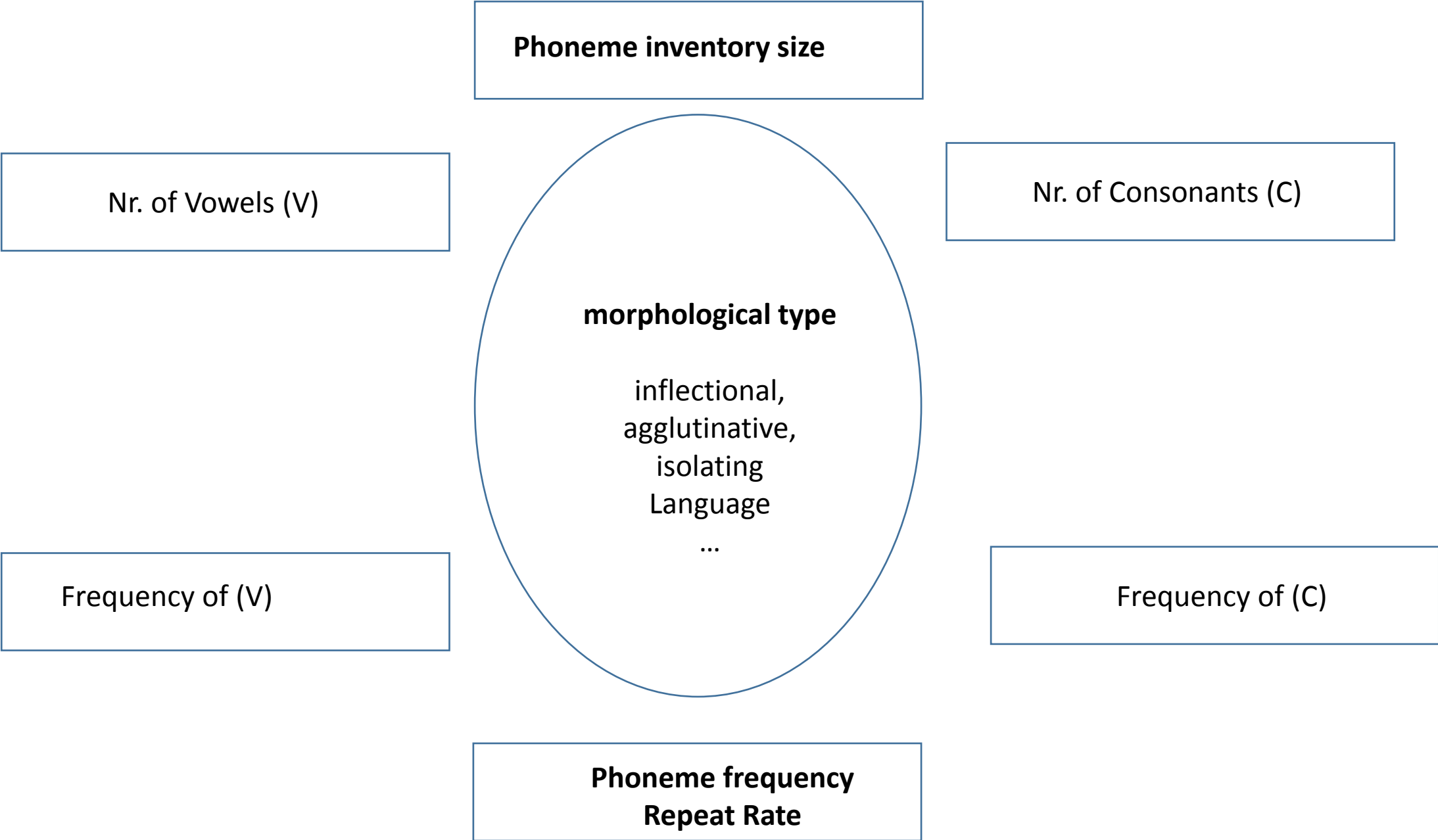
- other “external” factors (climate, density of communication, degree of language contact)
- stimulation for the analysis of the linguistic complexity ...
- What about “intrinsic” factors and the system linguistics value of the phoneme inventory size?

Phoneme inventory size from a synergetic point of view



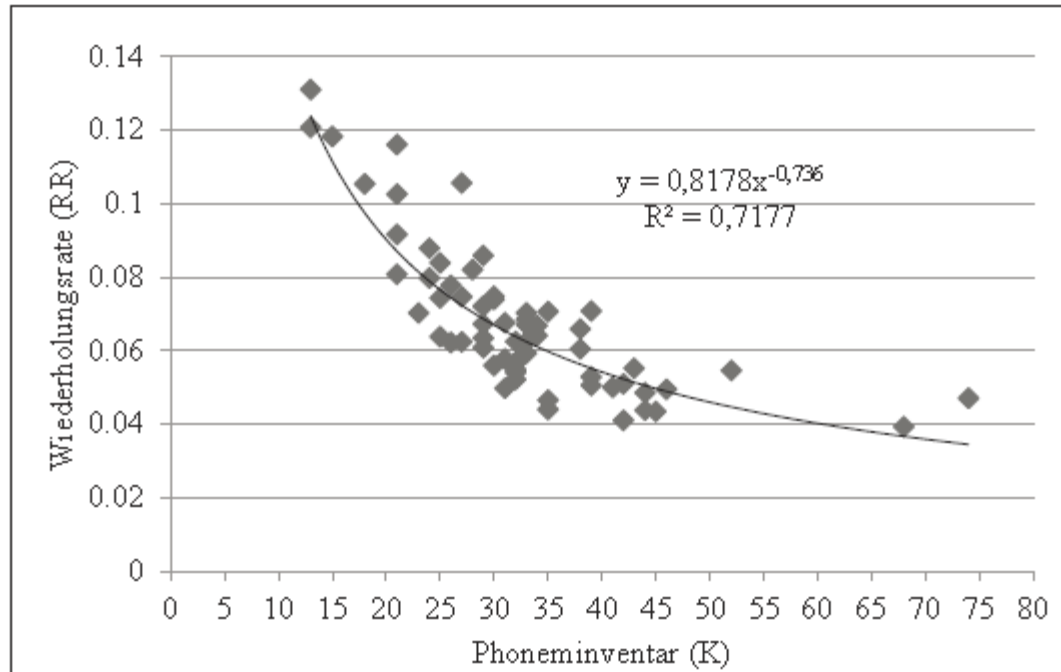
Kelih, Emmerich (2016): *Phonologische Diversität – Wechselbeziehungen zwischen Phonologie, Morphologie und Syntax*. Frankfurt am Main u.a.: Peter Lang.

Phoneme inventory size from a synergetic point of view



Finding (1):

„The higher the phoneme inventory size, the lower the repeat rate (R)“.



$$R = \sum_{r=1}^n p_r^2$$

63 languages analysed by
Altmann/Lehfeldt (1980)

- Languages with a low phoneme inventory size → high exploitation of the available phonemes
- big phoneme inventory size → “uniform” distribution of phoneme frequencies.

Phoneme inventory size from a synergetic point of view

phoneme inventory size

syllable types (C, CV ...)

No. of phoneme combinations

morphological type

inflectional,
agglutinative,
isolating
language

monosyllabism

phonotactical „load“

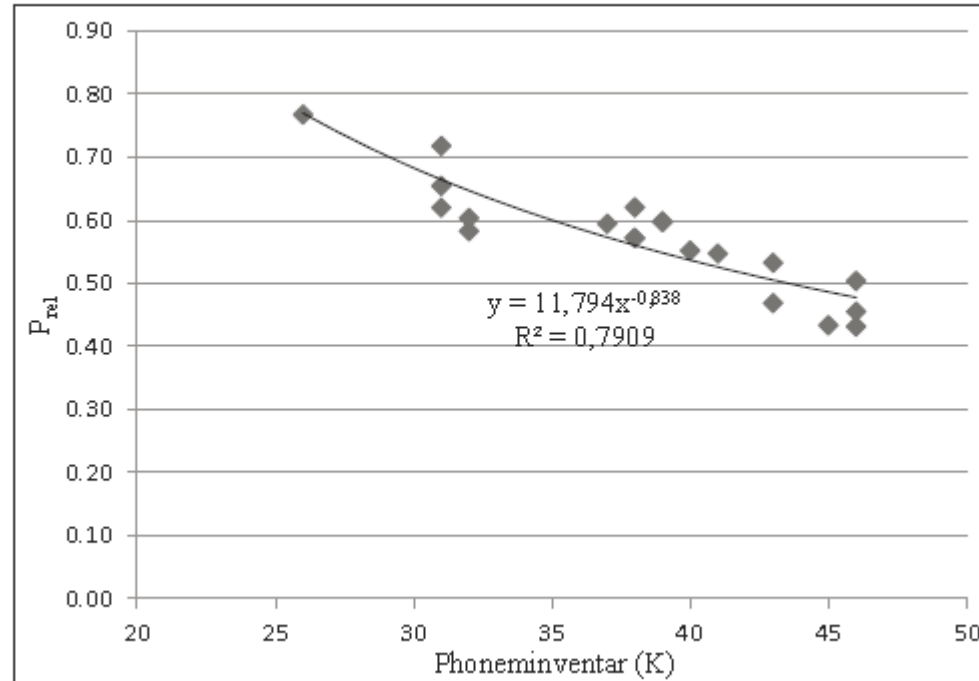
syllable length

Finding (2):

Phoneme inventory size – phonotactical „load“

„The higher the phoneme inventory size, the lower P_{rel} .“

Abb. 6: Phoneminventar (K) und relativer Ausnutzungsgrad P_{rel} in slawischen Sprachen.



$$P_{rel} = RPC/K^2$$

K = Phoneme inventory size

K^2 = theoretical number of phoneme combinations

RPC = „realised“ phoneme combinations

Re-analysis of data for 12 Slavic languages

Phoneme inventory size: **26 – 46 phonemes**

- Languages with small phoneme inventories → high number of phoneme combinations
- Some kind of compensation/trade off, in dependency of the phoneme inventory size

Phoneme inventory size from a synergetic point of view

What about other interrelations?



We might naturally expect that, other things being equal, it would in general require longer strings of segmental phonemes to express given meanings in a language like Hawaiian [= language with 13 segmental phonemes – EK] than on one like Chipewyan [= language with 45 phonemes – EK]. (Hockett 1958: 93).

→ The bigger the phoneme inventory size, the lower the word length.

Cf. Zipf (1949a: 395), Hockett (1958), Skalička (1958: 74), Sigurd (1963), Lomtev (1965), Saporta (1966: 69–70), Milewski (1965), Décsy (1970: 11), Lehfeltdt (1975), Slipčenko (1977), Dressler (1979: 268), Polikarpov (1979), Köhler (1986), Kempgen (1990), Nettle (1995, 1998), Plank (1998), Carstairs-McCarthy (1999: 13), Wurzel (2001a: 379), Weber (2005).

Phoneme inventory size – word length

Empirical findings (Nettle 1995, 1998)

“Functional theories of language structure predict that as the number of contrastive segments [S] in a language increases, the average length of a word [L] will decrease.” (Nettle 1995)

- Analysis of 22 languages
- counts, based on the number of „contrastive segments“
- Vocalic segments are multiplied by the no. of tones
- Word length is measured in 50 word forms from dictionaries (random sampling)

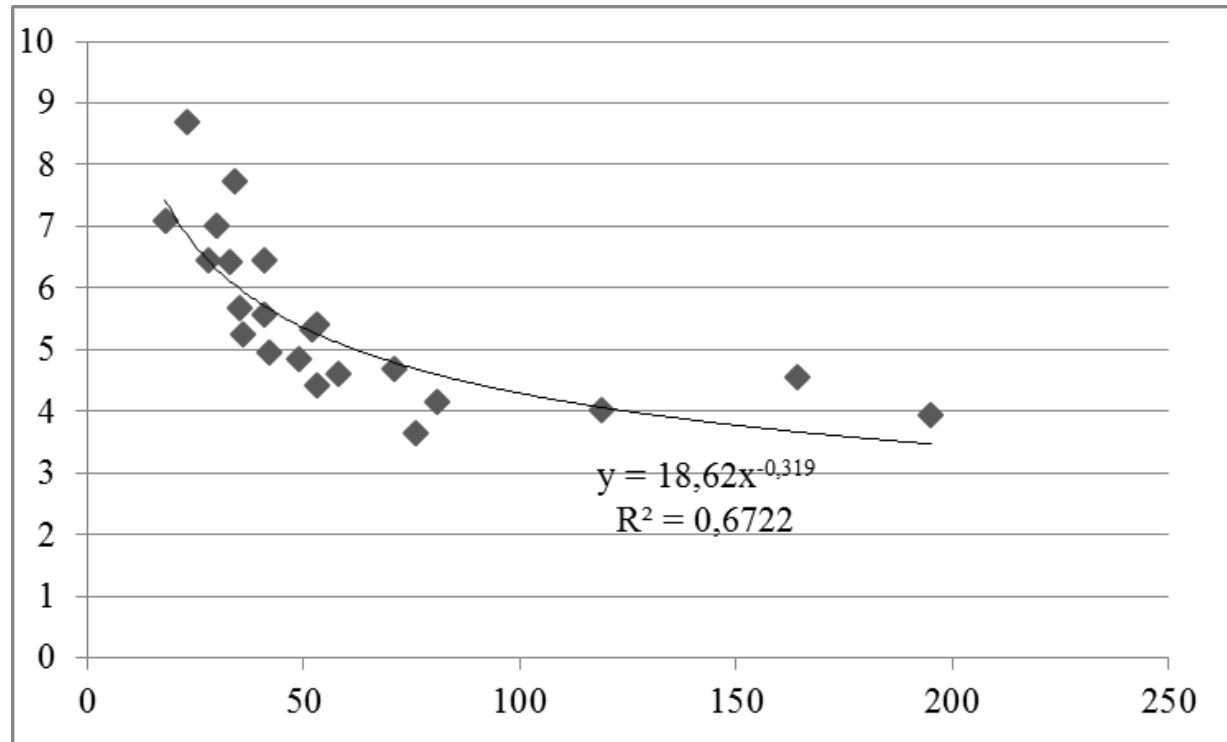
Phoneme inventory size – word length

Empirical findings (Nettle 1995, 1998)

Language	Nettle (1995)		Nettle (1998)		
	segments S (No.)	Average word length	Language	segments S (No.)	Average word length
Thai	76	3,65	Fula	33	6,42
Italian	30	7,00	Hausa	35	5,68
Hindi	41	5,57	Tamasheq	36	5,26
Hawaiian	18	7,08	Songhai	42	4,96
!Xū	119	4,02	Bambara	49	4,86
Turkish	28	6,44	Ngizim	52	5,32
Nahuatl	23	8,69	Edo	53	4,42
German	41	6,44	Igbo	58	4,62
Georgian	34	7,74	Mende	71	4,7
Mandarin	53	5,40	Ewe	81	4,16
			Vata	164	4,56
			Vute	195	3,94

Phoneme inventory size – word length

Empirical findings (Nettle 1995, 1998)



→ What happens when we analyse further languages: confirmation or rejection?

How the achieve homogeneity in the determination of the word length?

- Analysis of the **word length in the basic vocabulary** (Swadesh list – 200 word forms)
- Slavic languages only (Russian, Ukrainian, Belarusian, Polish, Upper Sorbian, Slovak, Czech, Bulgarian, Macedonian, Croatian, Serbian, Slovene)
- Phonological transcription
- Segmental and suprasegmental/prosodic features considered

COMMONLY USED ADJECTIVES

<i>PSl</i>	<i>OCS</i>	<i>Ukr</i>	<i>R</i>	<i>Br</i>	<i>Blg</i>	<i>M</i>	<i>SC</i>
bělъ(ь)	бѣлъ	білий	белый	белы	бял	бел	бѣо
cělъ(ь)	цѣлъ	цілий	целый	цѣлы	цял	цел	цѣо
čistъ(ь)	чистъ	чистий	чистый	чысты	чист	чист	чѣст
čьrnъ(ь)	чрьнъ	чорний	черный	чорны	черн	црн	црн
čьrvenъ(ь)	чрьвенъ	червоний	—	чырвоны	червен	црвен	црвен
dьlgъ(ь)	дльгъ	довгий	долгий	доугі	дльг	долг	дѣг
dobъrъ(ь)	добра	добрий	добрый	добры	добър	добар	дѣбар
gorъkъ(ь)	горькъ	гіркий	горький	горкі	горък	горок	гѣрак
kyslъ(ь)	кьслъ	кислий	кислый	кіслы	кисел	кисел	кѣсео
krasъnъ(ь)	красьнъ	красний	красный	красны	красн	красен	крѣсан
l'ubъ(ь)	любѣ	любий	люб	любы	либе (им.)	љуба (им.)	љуба (им.)
malъ(ь)	малъ	малий	малый	малы	мальк	мал	мѣо
mekъkъ(ь)	мянъкъ	м'який	мягкий	мяккі	мек	мек	мѣк

<i>Sln</i>	<i>P</i>	<i>Cz</i>	<i>Slk</i>	<i>UL</i>	<i>LL</i>	<i>Polab.</i>
běl	biały	bilý	biely	běly	běly	b'olě
cěl	cały	celý	celý	cyły	cely	c'ol
čist	czysty	čistý	čistý	čisty	cysty	caistě
čřn	czarny	černý	čierny	čorny	carny	corně
čřljen	czzerwony	červený	červený	čerwjeny	cerwjeny	carveně
dōlg	długi	dlouhý	dihý	dothi	dŕujki	dāud'ě
dōber	dobry	dobrý	dobrý	dobry	dobry	dūbrě
gōrek	gorzki	hořký	horký	hórki	gorki	d'ört'ě
kisel	kisly	kyselý	kyslý	kisały	kisały	—
krāsēn	krasny	krásný	krásny	krasny	kčasny	krosně
ljūb	luby	libý	l'úby	luby	luby	l'aibě
māli	mały	malý	malý	mały	mały	molě
mehāk	miękki	měkký	mākký	mjhcki	měki	maq'ě

Studies by Kelih (2008, 2009, 2012a, b, 2016)

- Analysis of the word length in **parallel texts of Slavic languages** (chapter 1 of **How the steel was tempered by N.A. Ostrovskij**)
- Slavic languages = translations only (Ukrainian, Belarusian, Polish, Upper Sorbian, Slovak, Czech, Bulgarian, Macedonian, Croatian, Serbian, Slovene)
- Phonological transcription of the word forms
- Segmental and suprasegmental/prosodic features considered

RU: Кто из вас перед праздником приходил ко мне домой отвечать урок - встаньте!

UKR: Хто з вас перед святом приходив до мене додому відповідати урок - устаньте!

BRU: Хто з вас перад святам прыходзіў да мяне на дом адказваць урок - устаньце!

P: Kto z was był u mnie w domu przed świętem, żeby odpowiadać z lekcji — niech wstanie!

Sorbian: Štóž z was je do swjatych dnjow poła mnje doma był, zo by swoje nadawki přednjesł, njech stanje!

SK: Ktorí z vás boli pred sviatkami odpovedať u mňa na byte, nech vstanú!

CZ: Kdo vás byl před svátky u mne v bytě odříkat úlohu, ať vstane.

Bulg. Тия от вас, които идваха преди празника у дома да ги изпитвам — да станат!

Mac. Тие што доаѓаа пред празникот да ги испрашувам - нека станат!

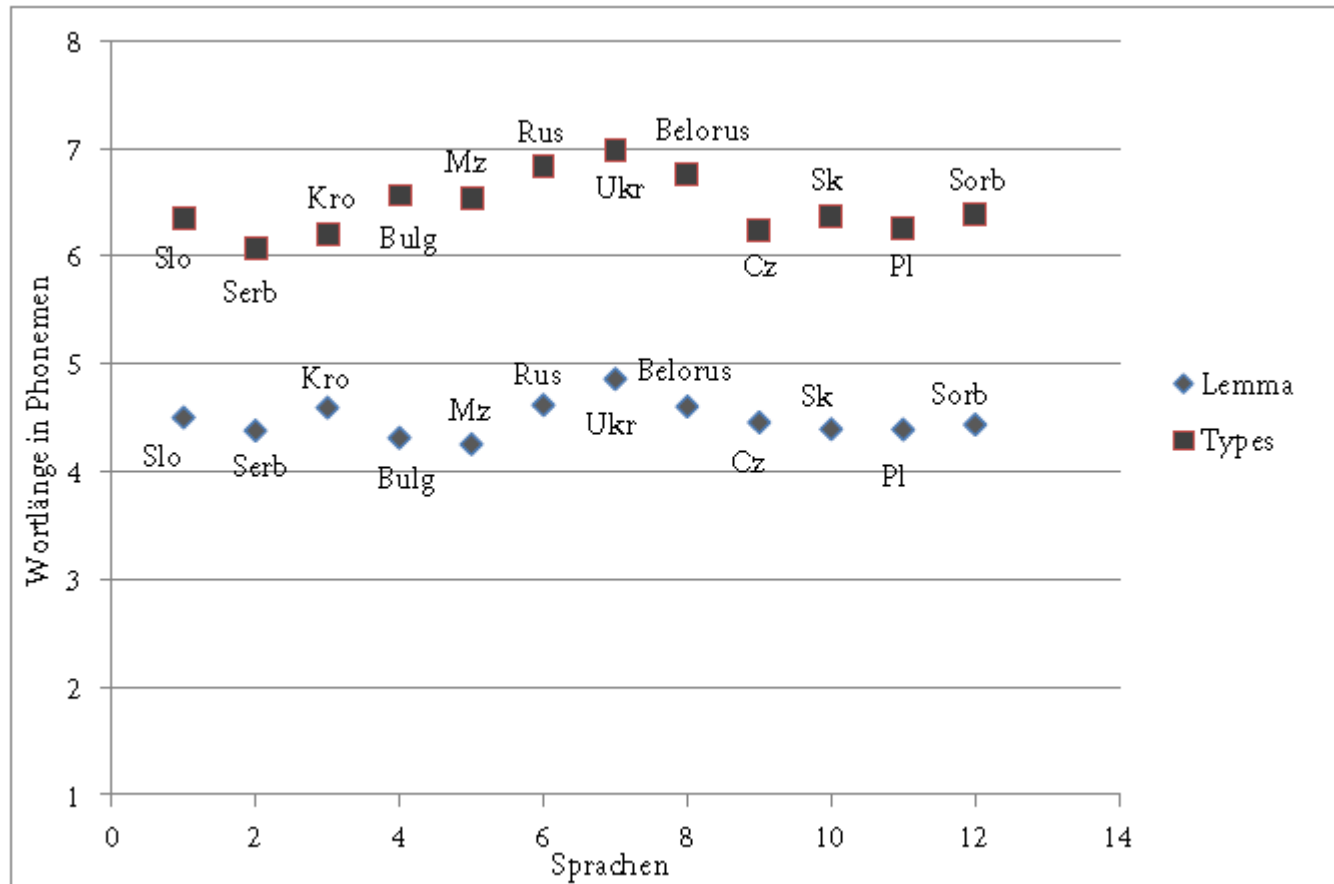
Cro. Tko je od vas prije praznika dolazio mojoj kući da odgovara lekciju, neka ustane.

Serb. Ko je od vas pred praznik dolazio mojoj kući da odgovara lekciju, neka ustane.

Slov. Tisti izmed vas, ki so pred praznikom prihajali k meni ponavljat krščanski nauk, naj vstanejo!

Studies by Kelih (2008, 2009, 2012a, b, 2016)

- step: determination of word length on different levels:
Basic vocabulary = lemma
parallel text analysis = word form types



Studies by Kelih (2008, 2009, 2012a, b, 2016)

II. step: determination of the phoneme inventory size

No.	Language	V	C	I
1	Russian	5	37	42
2	Ukrainian	5	32	38
3	Belarusian	5	39	44
4	Polish	7	35	42
5	Upper Sorbian	7	30	37
6	Slovak	15	27	42
7	Czech	12	25	37
8	Bulgarian	6	39	45
9	Macedonian	6	26	32
10	Croatian	25	25	50
11	Serbian	24	25	49
12	Slovene	19	21	40

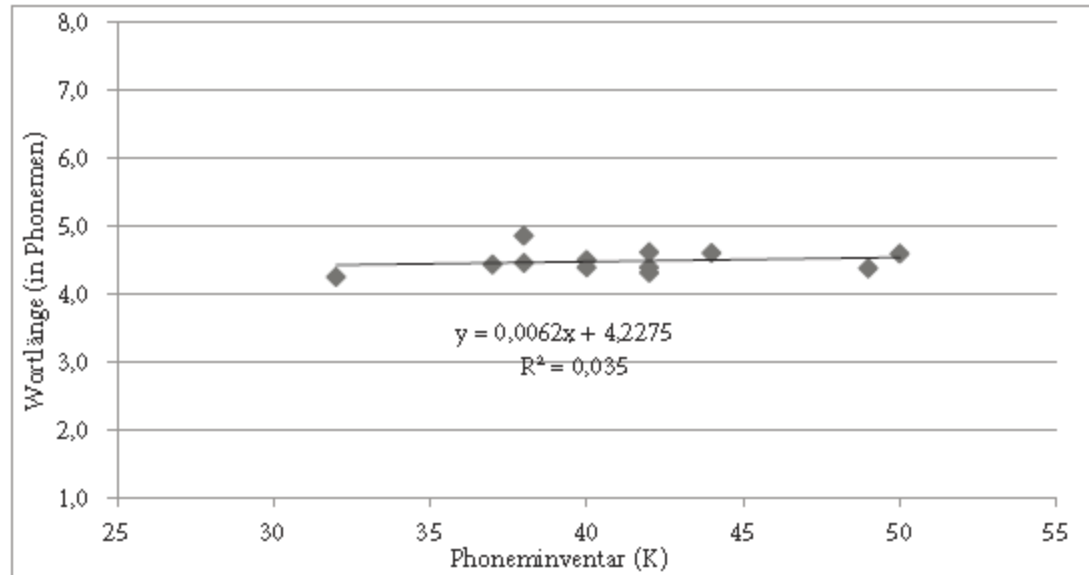
Principles of the determination

- Focus on standard languages, no dialects!
- Uniform handling of the palatalization
- Affricates and diphthongs are treated monophonemically
- uniform treatment of syllabic consonants (/r/, /l/)
- Long vowels are considered separately
- Pitch accent in Serbian, Croatian and Slovene is considered

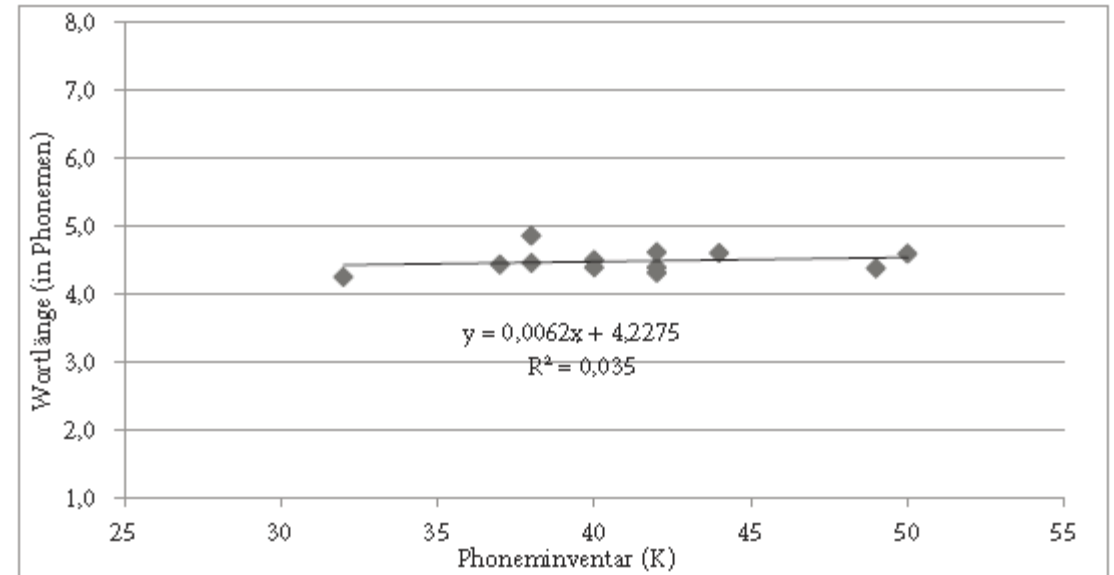
Studies by Kelih (2008, 2009, 2012a, b, 2016)

III. step: Phoneme inventory size – Word length in Slavic languages

Basic vocabulary – Swadesh list



Parallel texts

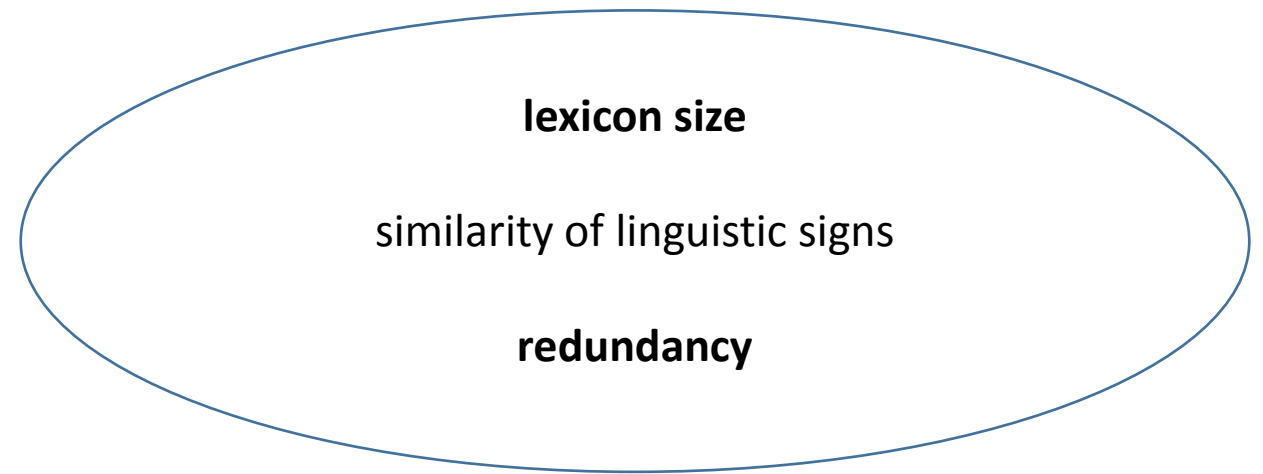


What are the causes for this (negative) result?

- homogeneity of the material analysed (types, lemmas) too high?
- Similarity/relatedness of the Slavic languages too high?
- No “need” for a morphological and phonological diversification
- Unreliable counting principles?

Summary – Perspectives

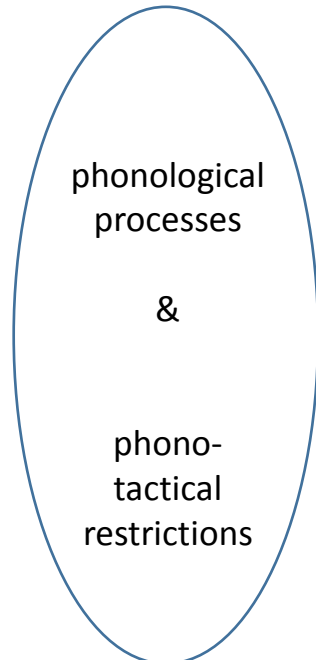
Köhler, Reinhard (1986): *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer (Quantitative Linguistics, 31).



stress/accent

No. of suprasegmental features

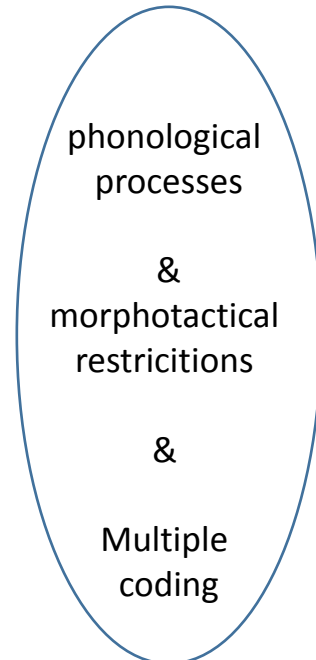
phoneme inventory size



syllable length

syllable types

No. of phoneme combinations



morpheme length

morpheme types

morpheme inventory

word length

References:

Altmann, Gabriel; Lehfeldt, Werner (1980): *Einführung in die quantitative Phonologie*. Bochum: Brockmeyer (Quantitative Linguistics, 7).

Bybee, Joan (2011): How plausible is the hypothesis that population size and dispersal are related to phoneme inventory size? Introducing and commenting on a debate. In: *Linguistic Typology* 15 (2), S. 147–153.

Donohue, Mark; Nichols, Johanna (2011): Does phoneme inventory size correlate with population size? In: *Linguistic Typology* 15 (2), S. 161–170.

Fenk-Oczlon, Gertraud; Fenk, August (2008): Complexity trade-offs between the subsystems of language. In: Matti Miestamo, Kaius Sinnemäki und Fred Karlsson (Hg.): *Language complexity. Typology, contact, change*. Amsterdam / Philadelphia: Benjamins (Studies in Language Companion Series, 94), S. 43–65.

Hay, Jennifer; Bauer, Laurie (2007): Phoneme inventory size and population size. In: *Language* 83 (2), S. 388–400.

Kelih, Emmerich (2016): *Phonologische Diversität – Wechselbeziehungen zwischen Phonologie, Morphologie und Syntax*. Frankfurt am Main u.a.: Peter Lang.

Köhler, Reinhard (1986): *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer (Quantitative Linguistics, 31).

Maddieson, Ian (1984): *Patterns of sounds*. Cambridge: Cambridge University Press.

Nettle, Daniel (1995): Segmental inventory size, word length, and communicative efficiency. In: *Linguistics* 33 (2), S. 359–367.

Nettle, Daniel (1998): Coevolution of Phonology and the Lexicon in Twelve Languages of West Africa. In: *Journal of Quantitative Linguistics* 5 (3), S. 240–245.

Pericliev, Vladimir (2004): There Is No Correlation between the Size of a Community Speaking a Language and the Size of the Phonological Inventory of That Language. In: *Linguistic Typology* 8 (3), S. 376–383.