Emmerich Kelih

(Graz)

Phoneme-Inventory (Size) – Word length:

Some theoretical thoughts and empirical findings

- Institut für Slawistik, Universität Graz
- http://www-gewi.uni-graz.at/quanta/ [Graz-project on Quantitative Textanalysis]
- http://www-gewi.uni-graz.at/staff/kelih/ [emmerich.kelih@uni-graz.at]

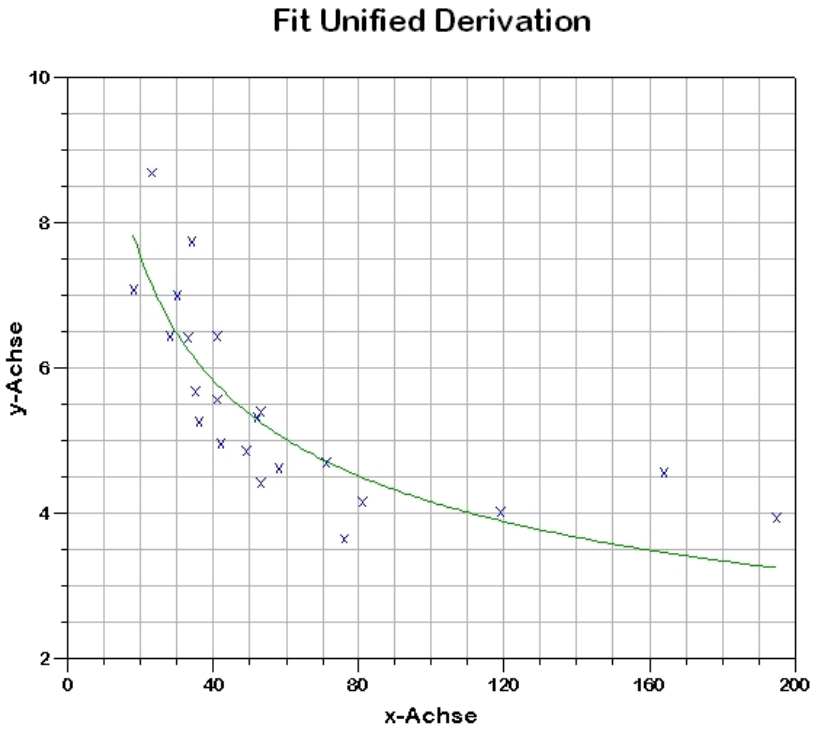
Aims and objectives of the study:

Is there a systematic relation between the phoneme inventory (size) and the mean word lenght?

1. linguistic aspects (theory and empirical problems)
2. system-based considerations
3. step by step: phoneme-inventory and phonotactics
4. empirical validation
5. general discussion

Starting point:

“that as the number of contrastive segments [S] in a language increases, the average length of a word [L] will decrease“. (Nettle 1995: 359).



22 languages (african and others)

$R^2 = 0.69$

... Some open questions!

problems of the relationship phoneme-inventory – word length:

- determination of the phoneme inventory is theory-driven
- segment inventory vs. phoneme inventory?
- determination of the level: word/wordform/lexem
- definition of the word, lexem ...
- measuring unit of the word length (phoneme, grapheme, syllable, morpheme)
- on which level the word length is determined?

→ Developing a new synergetic control cycle:

Missing relations:

1. **Phoneme inventory ←→ Phoneme distribution (phonotactics)**
2. **Phoneme inventory ←→ Syllable length**
3. **Phoneme inventory ←→ Syllable structure**
4. **Phoneme inventory ←→ Morpheme length**
5. **Phoneme Inventory ←→ Word length**

1. **Phoneme inventory (I)** (number of vowels and consonants)
2. **Set of phoneme combinations (R)** (the finite set of phoneme combinations, which are allowed by the language system)

On which level can we obtain phonem-combinations

a. on the paradigmatic level ? = language system

b. on the syntagmatic level ? = the realization in concrete texts

What is the base unit for obtainig phoneme combinations?

unit	example	R	R (abs.)
word	/Preporot/	PR, RE, EP, PO, OR, RO, OT	**7**
morpheme	/Pre-porot/	PR, RE, PO, OR, RO, OT,	**6**
syllable	/Pre-po-rot/	PR, RE, PO, RO, OT	**5**

Example: distribution matrix for Slovene

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	
		a	e	i	o	u	v	j	r	l	m	n	p	b	f	t	d	c	s	z	š	ž	č	dž	k	g	h	
1	a	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	25
2	e	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	25
3	i	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	25
4	o	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	25
5	u	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	25
6	v	x	x	-	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	23
7	j	x	x	-	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	24
8	r	x	x	-	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	23
9	l	x	x	-	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	24
10	m	x	x	-	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	-	22
11	n	x	x	-	x	x	x	x	x	x	x	-	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	24
12	p	x	x	x	x	x	-	x	x	x	-	x	-	-	x	x	-	x	x	x	x	-	x	-	x	-	x	18
13	b	x	x	x	x	x	x	x	x	x	x	x	-	-	-	-	x	-	-	x	-	x	-	-	-	x	-	15
14	f	x	x	x	x	x	-	x	x	x	x	x	-	-	-	x	-	x	x	-	-	-	x	-	x	-	-	15
15	t	x	x	x	x	x	x	x	x	x	x	x	x	-	x	-	-	x	x	-	-	-	x	-	x	-	x	18
16	d	x	x	x	x	x	x	x	x	x	x	x	-	x	-	-	-	-	-	x	x	x	-	-	-	x	-	16
17	c	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	-	-	x	-	-	-	-	-	x	-	x	17
18	s	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	x	x	-	-	x	-	x	-	x	-	x	20
19	z	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	-	x	x	-	-	-	x	-	20
20	š	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	-	x	-	-	-	-	x	-	x	-	x	18
21	ž	x	x	x	x	x	x	x	x	x	x	x	-	x	-	-	x	-	-	-	-	-	-	-	x	x	-	15
22	č	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	-	x	x	-	-	-	-	-	x	-	-	17
23	dž	x	x	x	x	x	x	-	x	-	-	x	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	8
24	k	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	-	x	x	x	x	-	x	-	-	-	x	20
25	g	x	x	x	x	x	x	x	x	x	x	x	x	x	x	-	x	x	-	x	x	-	x	-	-	-	x	20
26	h	x	x	x	x	x	x	x	x	x	x	x	x	-	-	x	x	x	x	-	x	-	x	-	x	-	-	19
																												521

1. no. of vowels (V) : 5
2. no. of consonants (C) : 21

Phoneme inventory (I) = 26

theoretical possible number of phoneme combination

$I^2 = 676$

But only **521** phoneme combinations (R) are „allowed“ by the system

Restrictions within language systems

- not all geminates (VV, CC) are possible
- the combination of voiced and voiceless consonants is prohibited
- ...

→ not all cells $(=I)^2$ of the distribution matrix are filled out

How can we model the relation between the phoneme-inventory (I) and phoneme combinations (R) in typologically different languages?

Data base of our study:

Analyzed languages (by Kleinlogel/Lehfeldt 1972)

no.	Language	I	V	C	R-empir.
1	Hethitisch	**21**	4	17	**233**
2	Spanisch (CR)	**23**	5	18	**290**
3	Baskisch (Maya)	**26**	5	21	**277**
4	Altjapanisch	**28**	5	23	**294**
5	Indonesisch	**29**	6	23	**459**
6	Serbokroatisch	**31**	6	25	**690**
7	Maharasti	**32**	5	27	**300**
8	Ungarisch	**37**	14	23	**922**
9	Sanskrit	**41**	10	31	**853**
...	...	...	...	...	...
32	Slovakisch	**46**	17	29	**1066**

New data basis

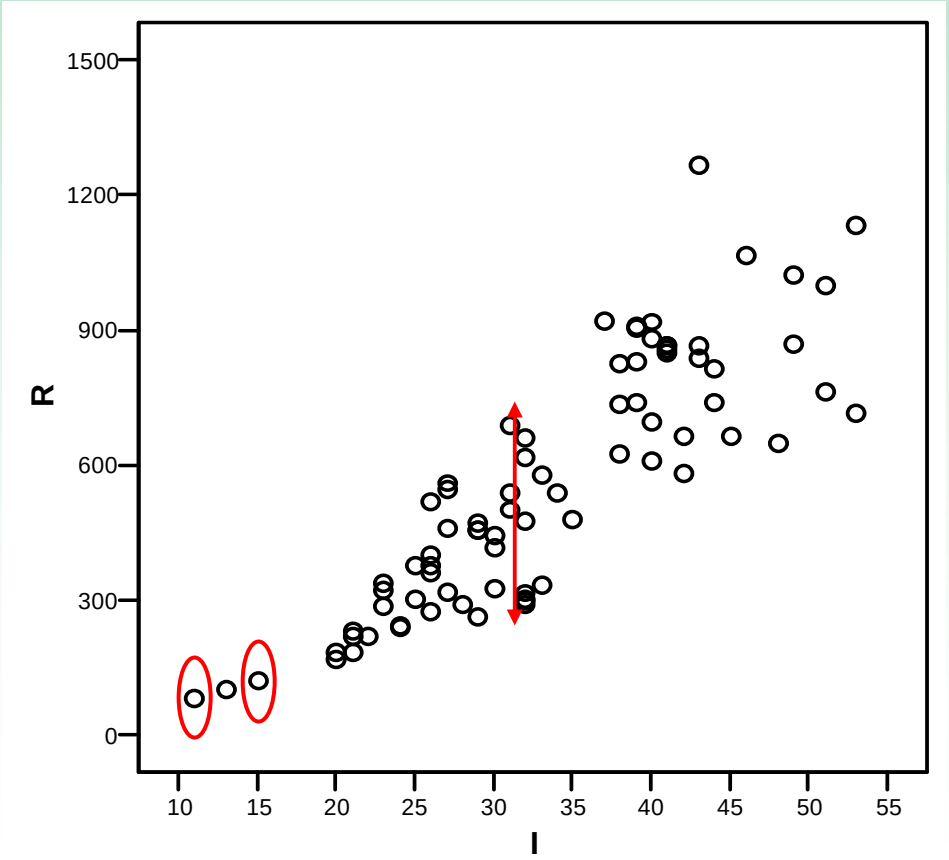
32 languages by Kleinlogel/Lehfeldt (1972)

+ 46 languages added

No.	Languages	I	R	No.	Languages	I	R
1	Rotokas	**11**	85	25	Amuesh	**32**	477
2	Hawaiinisch	**13**	104	26	Vedisch	**39**	740
3	Maori	**15**	125	27	Russisch-Kempgen	**39**	831
4	Huichol	**20**	172	28	Russisch (ph1)	**39**	910
5	Kaiwa (Guarani)	**20**	185	29	Polnisch-Dialekt (1)	**40**	610
6	Kurija	**21**	187	30	Polnisch-Dialekt (2)	**40**	696
7	Hethitisch	**21**	221	31	Polnisch	**40**	883
8	Kigongo	**22**	222	32	Sanskrit (3)	**41**	860
9	Sierra Nahuat	**23**	339	33	Sanskrit (2)	**41**	868
10	Luba	**24**	241	34	Sanskrit (1) Ivanov	**41**	869
11	Lomongo	**25**	303	35	Litauisch-Dialekt (1)	**42**	584
12	Mvera	**26**	363	36	Litauisch Dialekt (2)	**42**	666
13	Cuicateco	**26**	402	37	Kashmiri (1) oN	**43**	840
14	Slowenisch (sek.)	**26**	519	38	Weissrussisch-Dialekt (1)	**44**	742
15	Duala	**27**	319	39	Weissrussisch-Dialekt (2)	**44**	817
16	Totonako	**27**	462	40	Ganda	**45**	668
17	Lingala (oT)	**29**	266	41	Bambara	**48**	649
18	Indonesisch (1)	**29**	459	42	Khmer (1)	**49**	1025
19	Indonesisch (2)	**29**	473	43	Khmer (2)	**49**	870
20	Songe	**30**	420	44	Malayalam	**51**	763
21	Tojolabal	**30**	447	45	Punjabi	**51**	1001
22	Kikujno	**31**	503	46	Litauisch	**53**	719
23	Serbokroatisch 2	**31**	542				
24	Ardchamagadchi	**32**	292				

Developing a new model:

1. **dependent** variable: phoneme combinations (R)
2. **independent** variable: the complete phoneme inventory (I)



Heterogeneity of the data

1. high variance of R within I

I	R
32	292
	300
	305
	318
	477
	618
	661

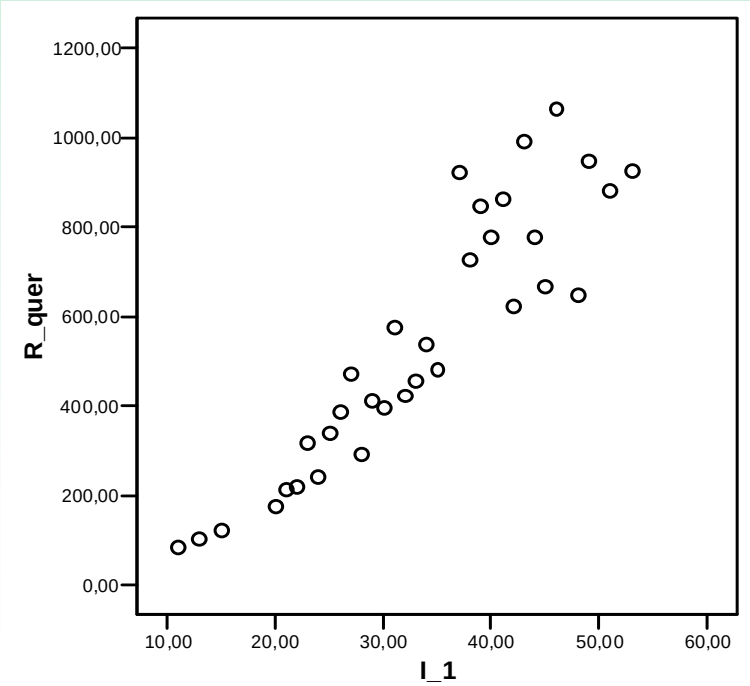
2. sometimes „only“ one language for a given I

→ data-pooling

Data-pooling

(1) average R per I

I	R	mean R
32	292	424,43
	300	
	305	
	318	
	477	
	618	
	661	



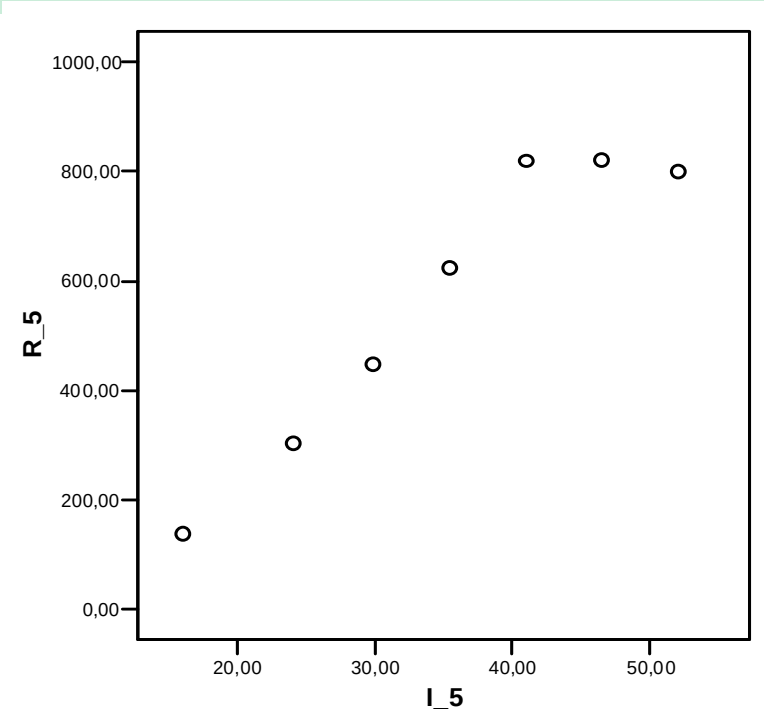
(2) pooling by I and R

intervall of 2 {11,12} {13,14}...

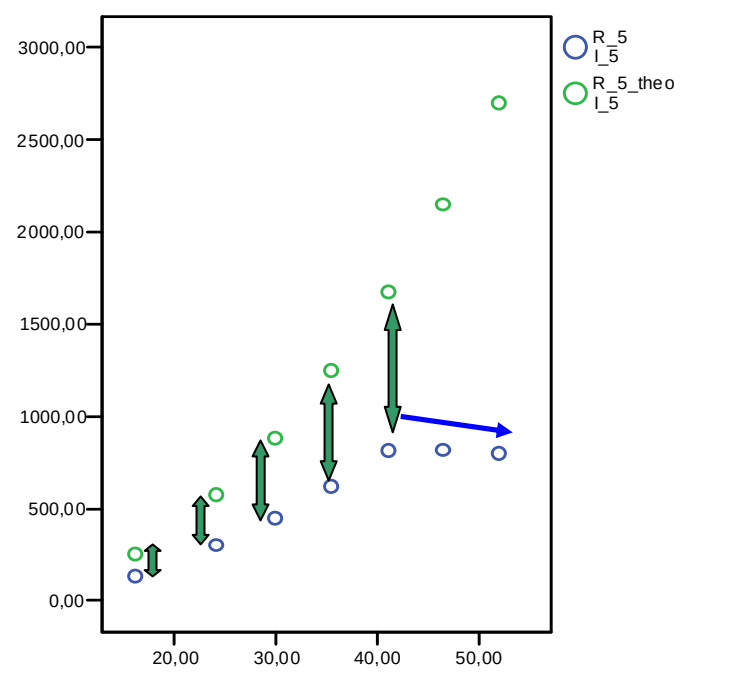
intervall of 3 {11, 12, 13} {14,15,16} ...

intervall of 4 {11,12,13,14} {15...19} ....

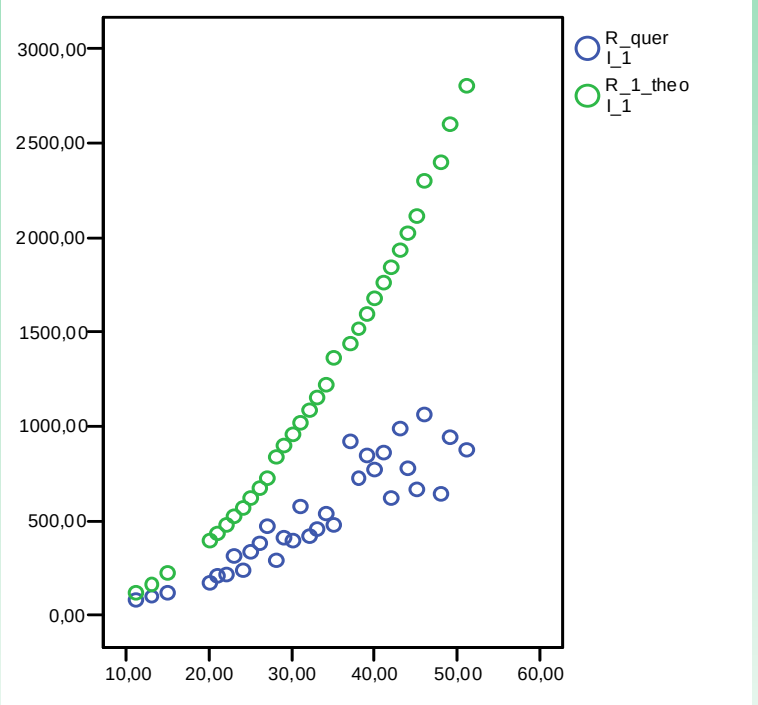
intervall of 5 {11 ...15}; {16 ... 20} ...



Detailed insight into the relation between $(I)^2$, (I) and (R)



Data-pooling (2)



Data-pooling (1)

1. the higher the phoneme inventory (I), the lower the increase of (R)

2. at a certain inventory level the number of (R) even decreases!

Building a new model

1. „breaking-component“ $R = sI^a$

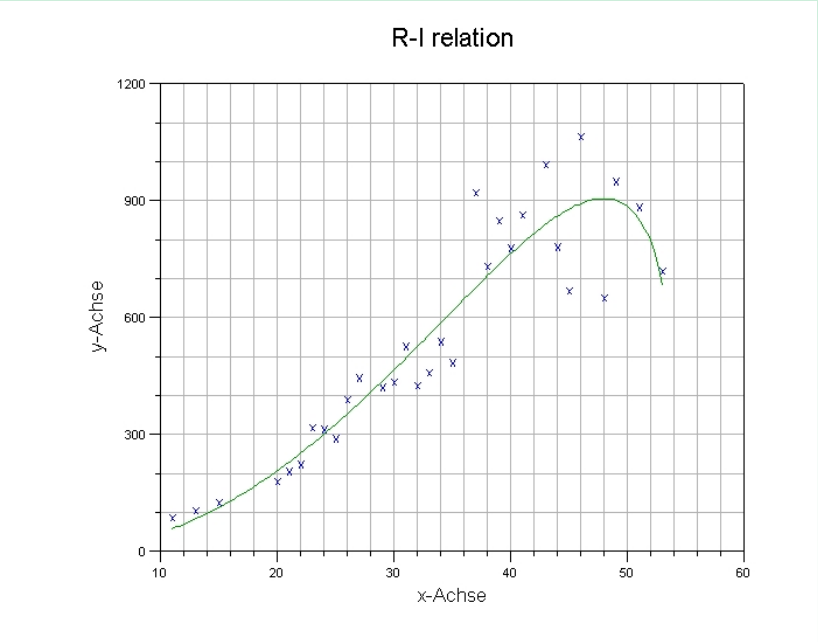
s, a: estimated parameters

2. „reduction-competent“ $(Z - I)^b$

b: estimated parameter

Z: ‘hypothetic’ maximum of the phoneme inventory

$$R = sK^a(Z - I)^b$$



Data-pooling (1)

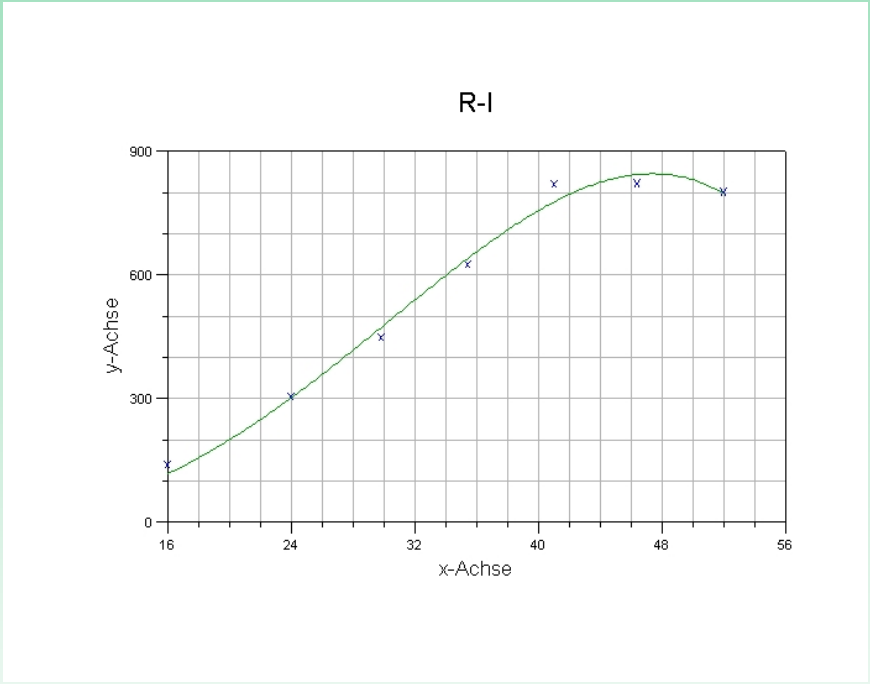
s = 0.00051679395

a = 2.94682218

b = 1.04970599

Z = 64

D = 0.8638

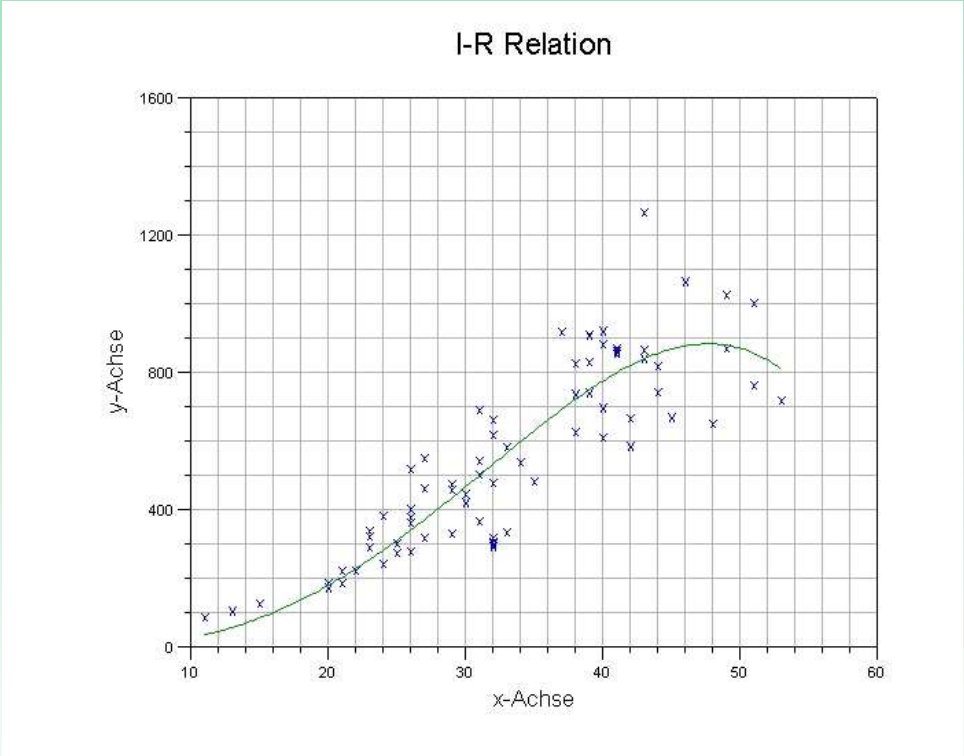


Data-pooling (2)

s = 0.00130445083
a = 2.76326041
b = 0.968273103
Z = 64

D = 0.9920

Without pooling



s = 0.000466404878
a = 2.99866285
b = 1.02676557
Z = 64

D = 0.7777 (!)

... a suitable model ...

Results and Outlook

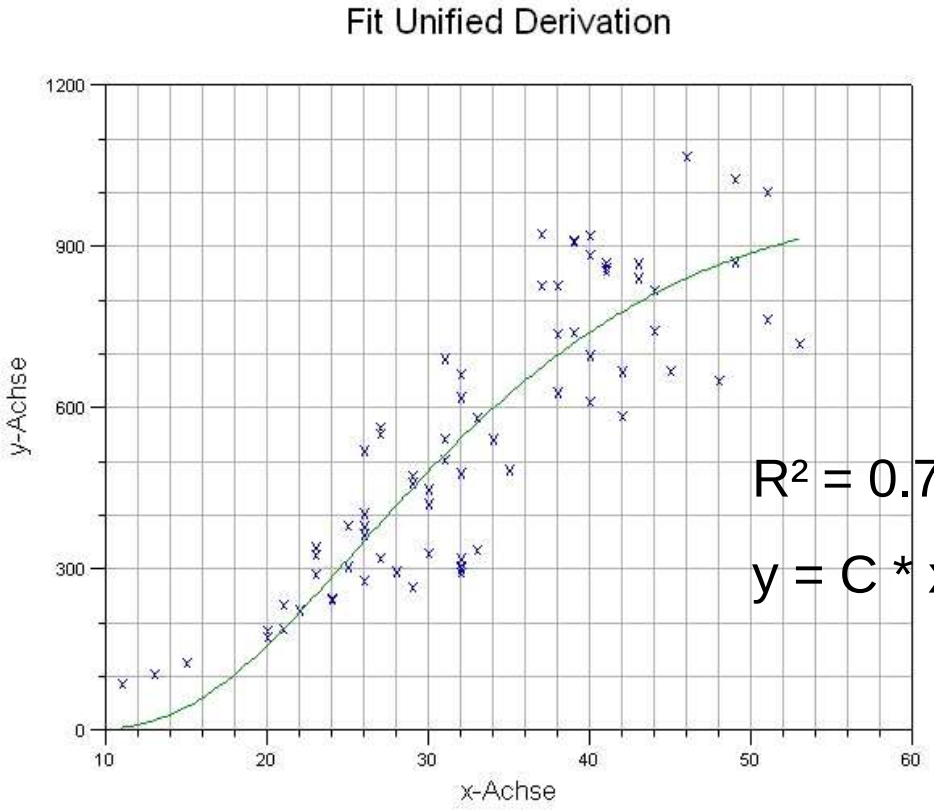
- systematic behaviour of phoneme-inventory size and phonotactics

New model:

- includes only phoneme inventory
- global braking-component
- global reduction-component
- empirically proven

- no generalizations in regard to syllable structure/length and morpheme length
- further stepwise analyzes necessary

Alternative model: no reduction (!)

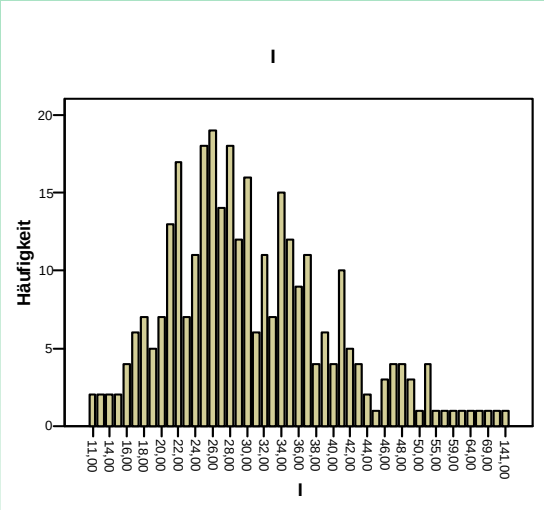


$R^2 = 0.75$

$y = C * x^{a1} * \exp(1)^{(-a2/x)}$

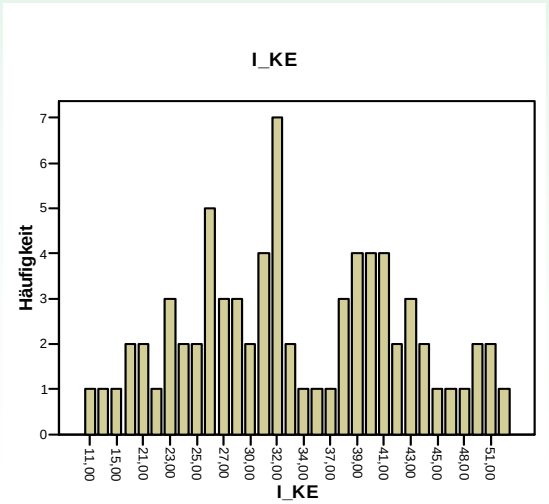
Frequency distribution of inventory size in 317 languages

(data from Maddieson 1984)



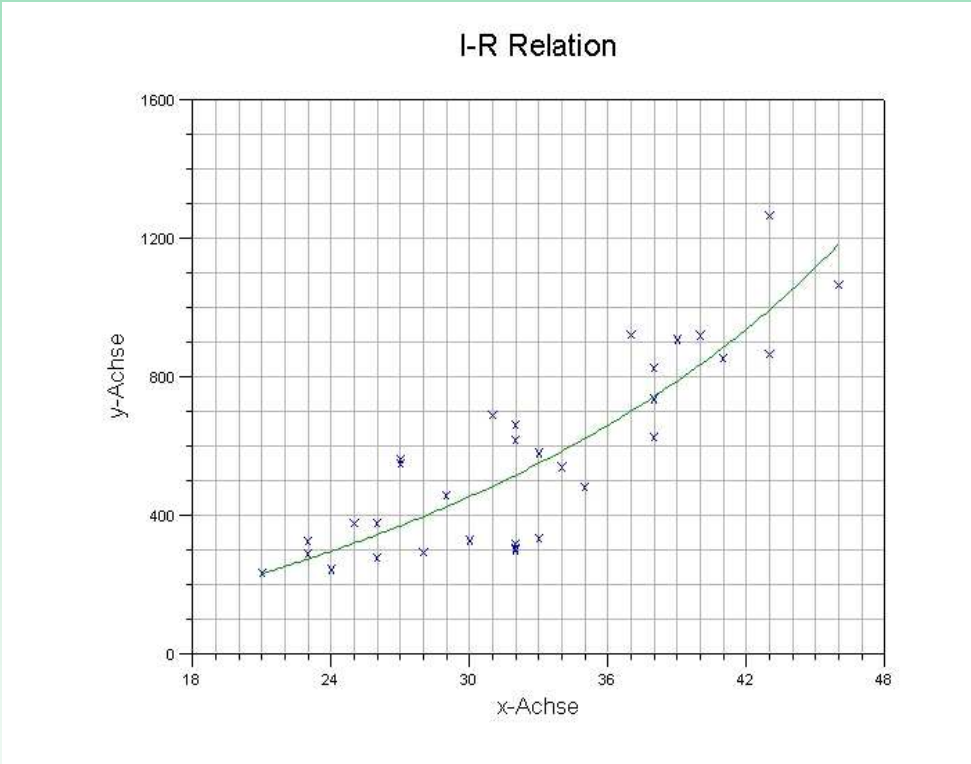
Mittelwert		**30,99684543**
95% Mittelwerts	**Untergrenze**	29,69987285
	Obergrenze	**32,293818**
Median		**29**
Varianz		**137,74999**
Standardabweichung		**11,73669417**
Minimum		**11**
Maximum		**141**
Spannweite		**130**
Schiefe		**3,167140619**
Kurtosis		**24,2032828**

in our data:



Mittelwert		**33,3164557**
95% Mittelwerts	**Untergrenze**	31,17958447
	Obergrenze	**35,45332692**
Median		**32**
Varianz		**91,01395651**
Standardabweichung		**9,540123506**
Minimum		**11**
Maximum		**53**
Spannweite		**42**
Schiefe		**0,065486303**
Kurtosis		**-0,51044176**

Fitting the data from Kleinlogel/Lehfeldt 1972 to the new model

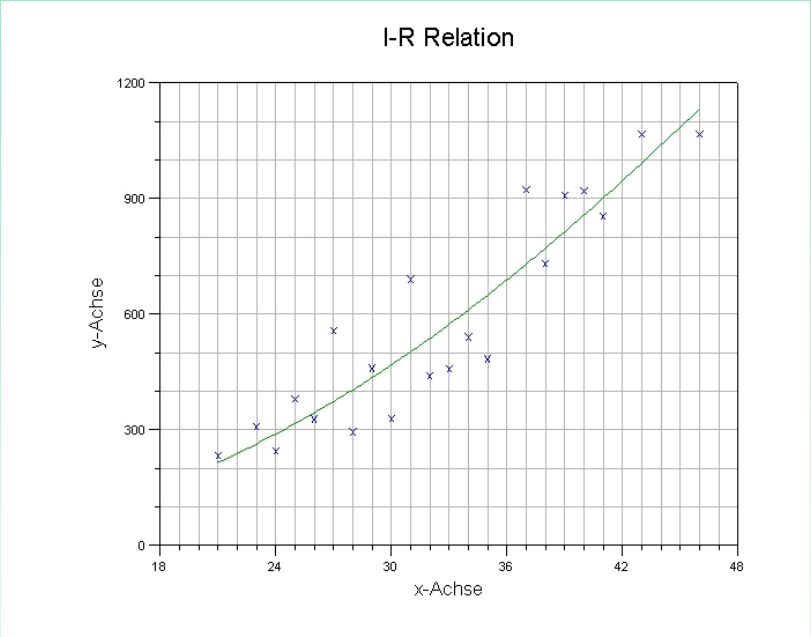


s = 9.90229697
a = 1.58292242
b = **-0.441354357**
Z = 64

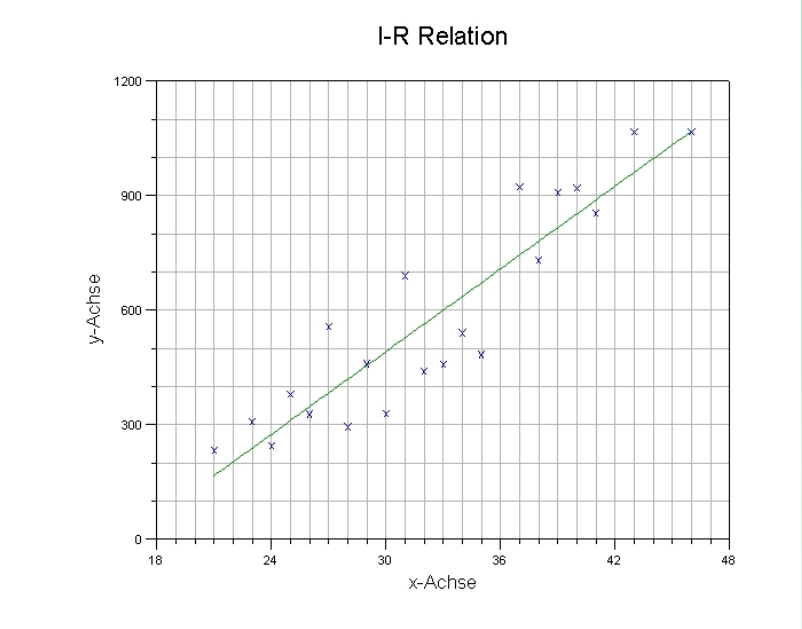
D = 0.7493

Without Pooling

New model



Simple linear model



Pooling (1)

~~s =0.145660787~~
~~a = 0.145660787~~
~~b = 2.24855856~~
~~Z = 64~~

D = 0.8512

No complex model is necessary for data from Kleinlogel/Lehfeldt 1972!

Pooling (1)

R (2b) = a0*I+W

a0 = 36.09
a1 = -591.15

D = 0.8352

→ the data does not allow to obtain our found relation!

Two discussed models for the I-R relation

(proposed by Kleinlogel/Lehfeldt 1972)

(1) Power model

„global“ reduction-factor

„lokal“ reduction-factor

minimum of all possible VC and CV

$$R = 0.5\,(I-2)^2 + 2\,CV$$

(2) Linear model

constant regression-coefficients

constant reduction parameter

$$R = 33*I + 16*V - 518$$

33 = average I

16 = 2*average V

518 = average R

in 32 analysed languages!

Re-analysing Kleinlogel/Lehfeldt (1972)

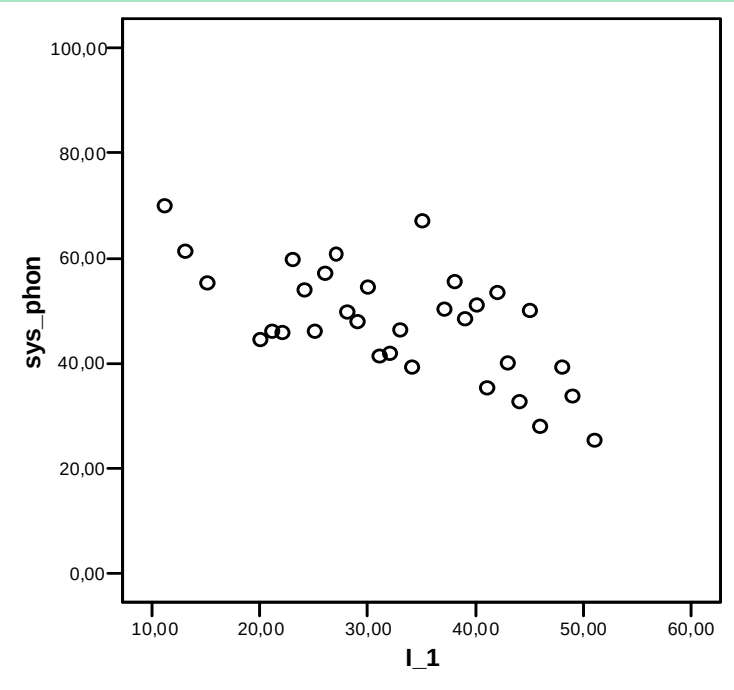
$R(1) = 0{,}5\,(I-2)^2 + 2\,CV$ → D = 0.61

$R(2) = 33*I + 16*V - 518$ → D = 0.63

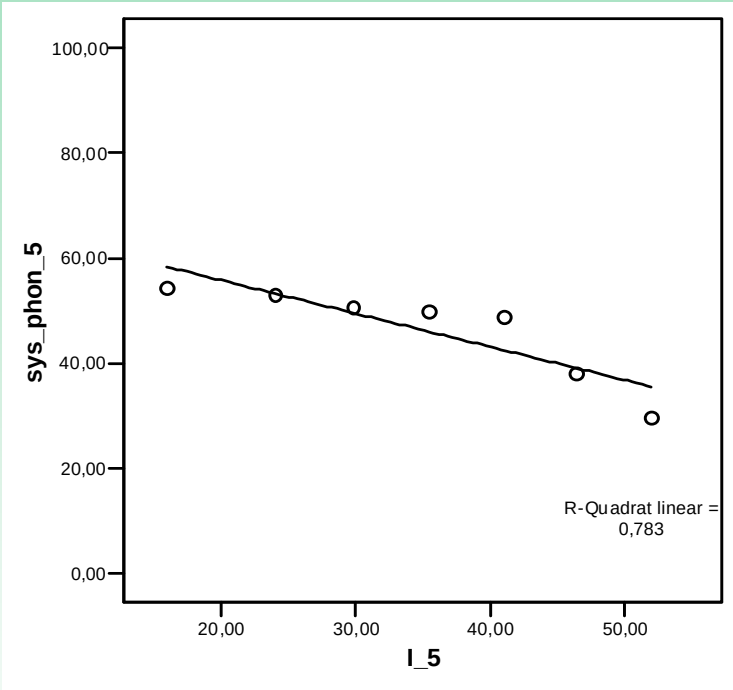
→ relative low values of D !

- is the data basis too small ?
- are the models not suitable ?

Prozentuelle Systemauslastung: Sys_phon = $(R/l^2)*100$



Data-pooling (1)



Data-pooling (2)

1. mit zunehmenden Inventarumfang sinkt die prozentuelle Auslastung des Phonemsystems hinsichtlich der Bildung von Phonemkombinationen