Emmerich Kelih

(Graz)

Quantifying Grapheme-Phoneme Relations from a Paradigmatic Perspective

- **Grazer Projekt zur quantitativen Text- und Sprachanalyse *QuanTA***
- **Forschungsschwerpunkt Text – Korpus – Sprache (*TKS*)**

Overview:

- Historical Introduction: Quantifying Phoneme-Grapheme Relations

- State of the Art: Quantitative Script Theory

- Selected aspects from „Analyses of Script Properties“ (2007)

    1. Phoneme – Grapheme – Letter: Inventories
    2. Orthographic uncertainty
    2. Graphemic Load
    3. Grapheme Size
    4. Relationships
    5. First empirical findings (6 languages)

Pioneers of Quantitative Script Analysis

Jan Baudouin de Courtenay (1845-1929)

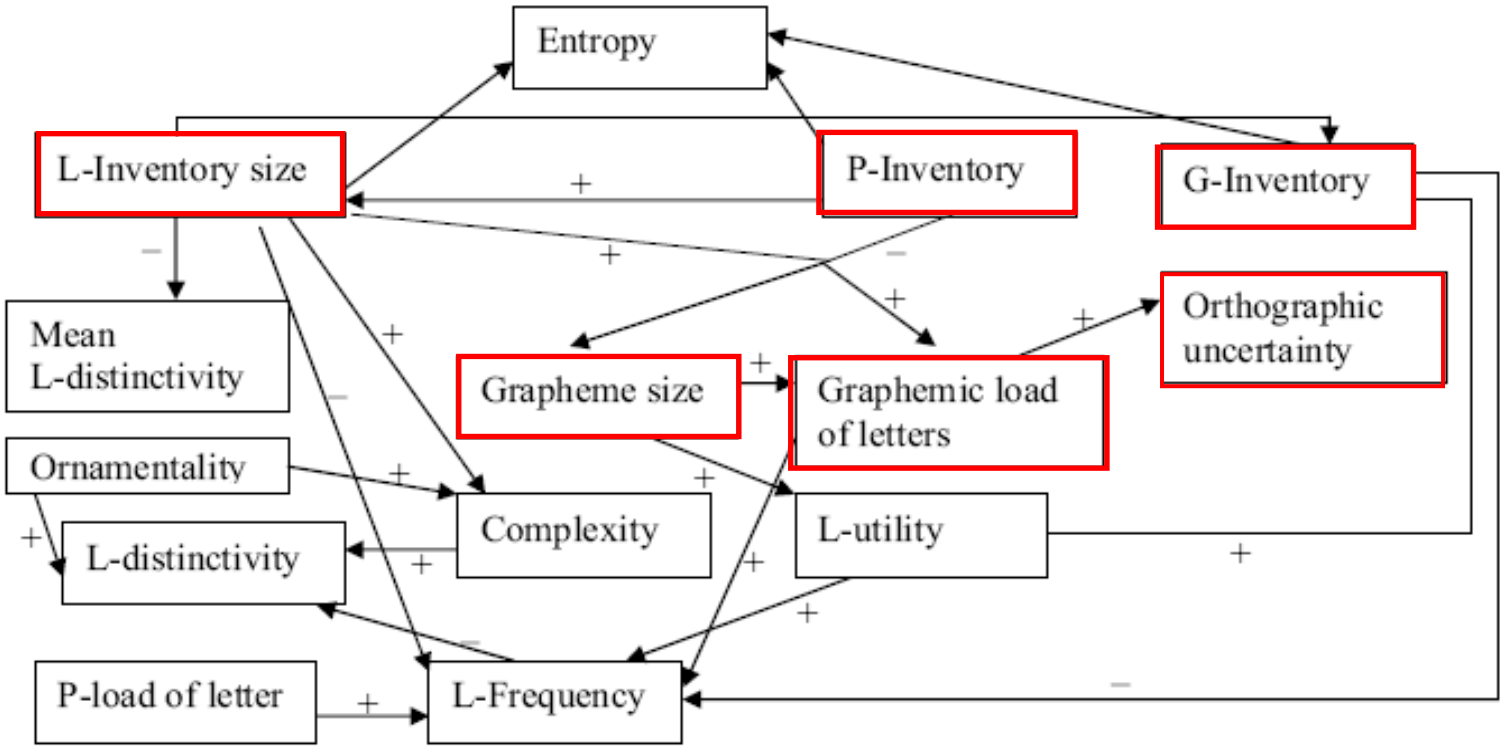
N.F. Jakovlev (1892-1974)

- 1876: *Otčety komandirovannago Ministerstvom Narodnago Prosveščenija za granicu s učenoju celju. I. A. Boduėna-de-Kurtenė o zanjatijach po jazykovedeniju v tečenie 1872 i 1873 gg. Vypusk I. Otčet' za oba polugodija 1872 g.*
- 1912: *Ob otnošenii russkago pis'ma k russkomu jazyku.* Sankt-Peterburg.

- 1928: Matematičeskaja formula postroenija alfavita (Opyt praktičeskogo priloženija lingvističeskoj teorii). In: *Kul'tura i pis'-mennost' Vostoka. Sbornik*

State of the Art: Quantitative Script Theory

Gabriel Altmann; Fan Fengxiang (ed.) (2007): *Analyses of Script Properties of Characters and Writing Systems.* Berlin: de Gruyter. To appear in: Quantitative Linguistics, 63. Ed. by Reinhard Köhler Reinhard and Peter Grzybek.



Phoneme – Grapheme – Letter: Inventories (Slovene)

Phonem- Inventory:

V: /i, e, ɛ, ə, a, o, ɔ, u/ 8

C: /p, b, f, v, m, t, d, s, z, n, r, l, š, ž, dž, c, č, j, k, g, h/ 21

= 29 phonemes

Grapheme-Inventory:

Graphem: A grapheme is a **letter**, or a **combination of letters**, or a letter **with additional diacritical marks** (such as those in Slavic languages, French, Spanish, German etc.) used as a whole in a language and **attributable to a phoneme**.

<a,b,c,č,d,e,f,g,h,i, j, k, l, m, n, o, p, r, s, š, t, u, v, z,ž, dž, lj, nj>

= 28 Graphemes

Letter-Inventar:

Letter:

A letter is a single sign adopted from Latin, Greek or other alphabetic scripts; a letter ist not necessarily attributable to a single phoneme.

<< a,b,c,e,f,g,h,i,k,m,o,p,r,s,t,u,v, d, l, n, z, j>>

= 22 letters

Property 1: Orthographic Uncertainty

Number of representations of phonemes by graphemes, e.g. how many graphemes are necessary for the representation of a phoneme?

Example: Slovene

Phoneme	Grapheme	Examples	Number of representing graphemes
/a/	<a>	sam, brat	1
/d/	<d>	dedje	2
	<t>	svatba	
/v/	<v>	siva	4
	<u>	Dachaua	
	<f>	Afgan	
	<l>	bral, bralca	

Property 1: Orthographic Uncertainty (Altmann/Best 2005)

Phoneme	Number of representing graphemes	Uncertainty	Number of phonemes with uncertainty U_x
	n_x	U_x	f_x
a, c, e, ɛ, i, f, h, m, o, □ , r	1	0,00	11
b, d, g, k, j, l, n, p, t, z, ə, s	2	1,00	12
č, u, dž	3	1,58	3
š, v, ž	4	2,00	3

Unweighted orthographic uncertainty:

$$U_{/x/} = \log_2 n_x$$

$$\overline{U} = \frac{1}{N} \sum_x f_x U_x$$

Results for 6 Languages:

Language	Graphemes	Phonemes	mean orthographic uncertainity
German	68	39	**0,9650**
Swedish	57	36	**0,7970**
Slovene	28	29	**0,7841**
Italian	71	59	**0,5641**
Slovak	51	44	**0,7586**
Croatian	31	31	**0,5000**

Property 2: Mean Grapheme size

The size referring to the number of “letters”;

Two methods: (1) with and (2) without additional signs as component

For Slovene: (1)

Size	Grapheme	Number
1	a,b,c,č,d,e,f,g,h,i, j, k, l, m, n, o, p, r, s, š, t, u, v, z,ž	25
2	dž, lj, nj	3

For Slovene: (2)

Size	Grapheme	Number
1	a,b,c,d,e,f,g,h,i, j, k, l, m, n, o, p, r, s, t, u, v, z	22
2	č, š, ž, lj, nj	5
3	dž	1

Results for 6 Languages:

Language	**gra. size (1)**	**gra. size (2)**
German	1,68	1,78
Swedish	1,61	1,67
Slovene	1,11	1,25
Italian	1,65	1,7
Slovak	1,16	
Croatian	1,16	1,35

Property 3: Graphemic load of letters (L-graphemic load)

This is a letter property expressing the participation of letters in building graphemes.

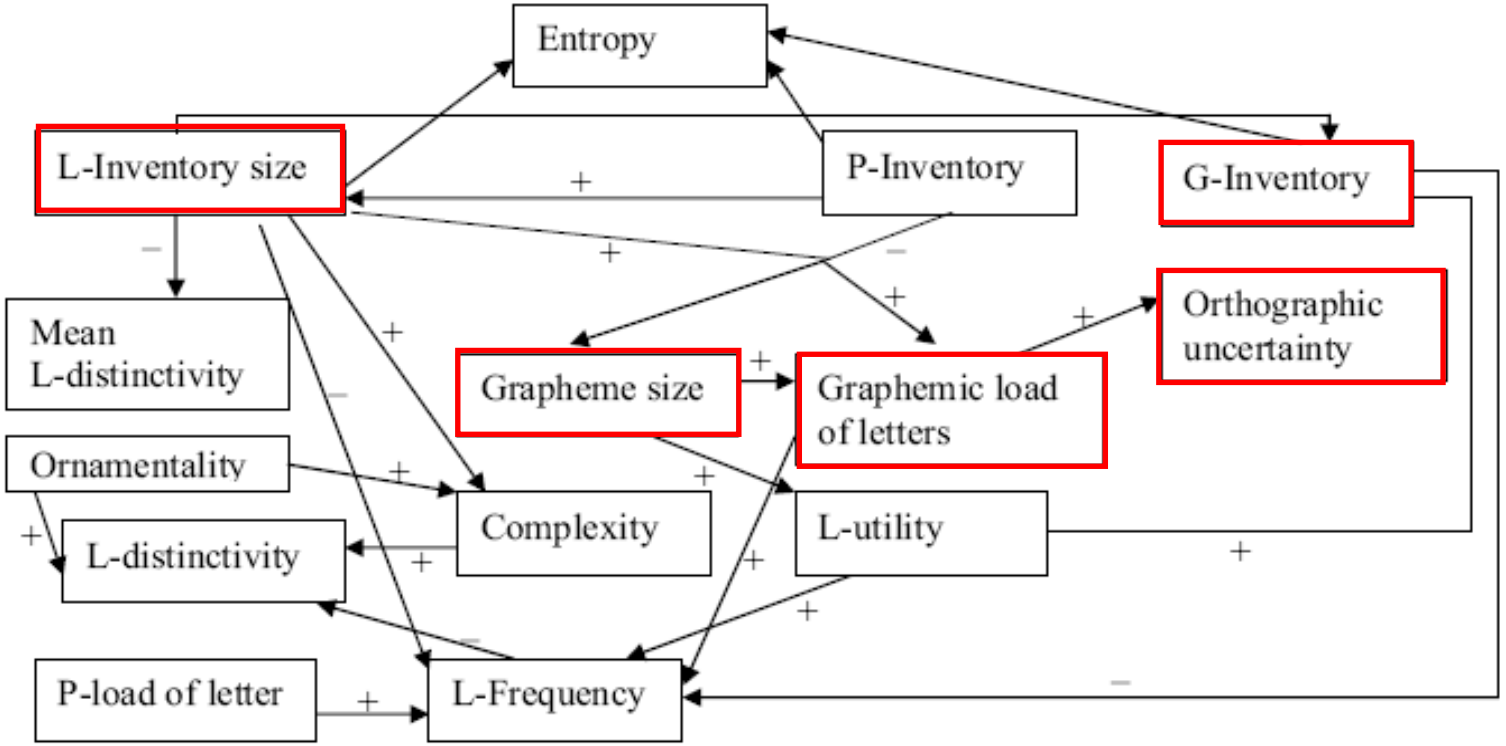
For Slovene:

Component in x graphemes	Latin letter	Number of letters
1	a,b,c,e,f,g,h,i,k,m,o,p,r,s,t,u,v	17
2	d, l, n, z	4
3	j	1

Results for 6 languages:

Language	mean graphemic load
German	4,12
Swedish	3,36
Slovene	1,27
Italian	3,92
Slovak	2,23
Croatian	1,32

Altmann‘s control cycle (2007):

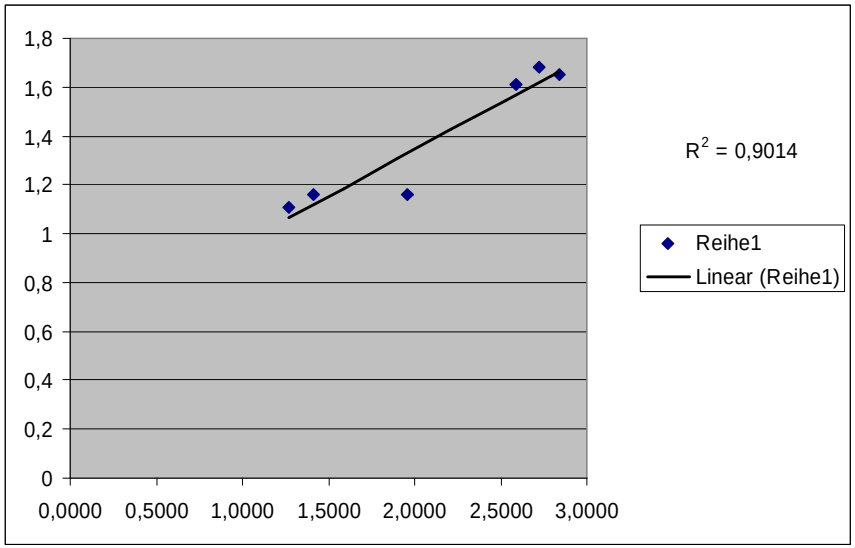


→ selected hypothesis: first empirical findings!

Hypothesis (1):

The relationship between G-inventory and L-inventory influences the mean size of graphemes.

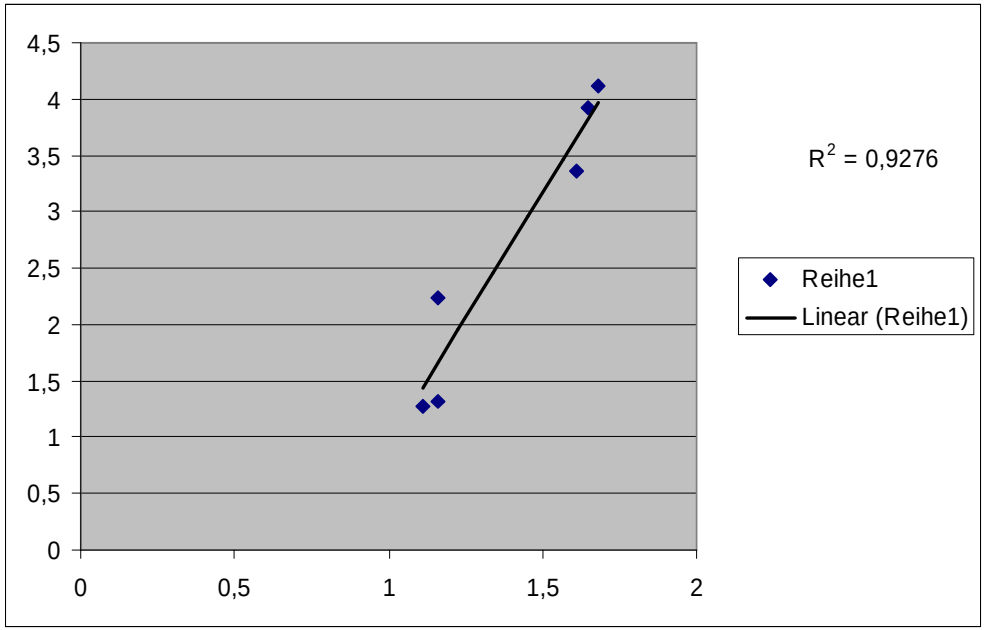
Language	Graphemes	Letter	G/L	gra. size (1)
German	68	**25**	**2,7200**	**1,68**
Swedish	57	**22**	**2,5909**	**1,61**
Slovene	28	**22**	**1,2727**	**1,11**
Italian	71	**25**	**2,8400**	**1,65**
Slovak	51	**26**	**1,9615**	**1,16**
Croatian	31	**22**	**1,4091**	**1,16**



R^2=0,90

Hypothesis (2):
The greater the size of graphemes, the greater the mean graphemic load of letters.

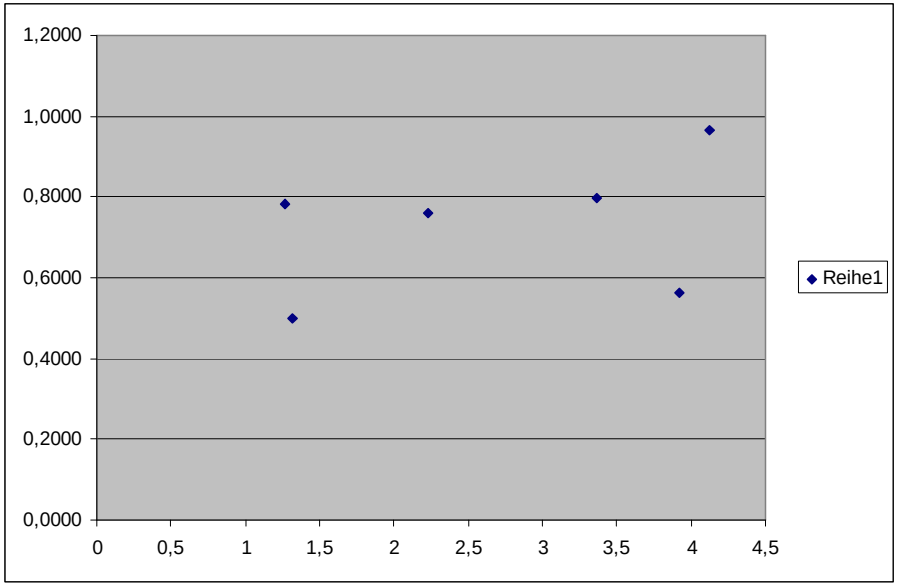
Language	Graphemes	Letter	gra. size (1)	mean graphemic load
German	68	25	1,68	4,12
Swedish	57	22	1,61	3,36
Slovene	28	22	1,11	1,27
Italian	71	25	1,65	3,92
Slovak	51	26	1,16	2,23
Croatian	31	22	1,16	1,32



R^2=0,92

Hypothesis (3):
The greater the mean G-load of letters, the greater the orthographic uncertainty.

Language	mean graphemic load	orthographic uncertainity
German	4,12	0,9650
Swedish	3,36	0,7970
Slovene	1,27	0,7841
Italian	3,92	0,5641
Slovak	2,23	0,7586
Croatian	1,32	0,5000



... For the time being no empirical evidence!

Conclusion:

- selected properties show a regular behaviour

- relationships are linear

- orthographic uncertainty is not part of the control cycle

- more languages should be analysed

- first steps towards an empirical analysis of writing systems ...

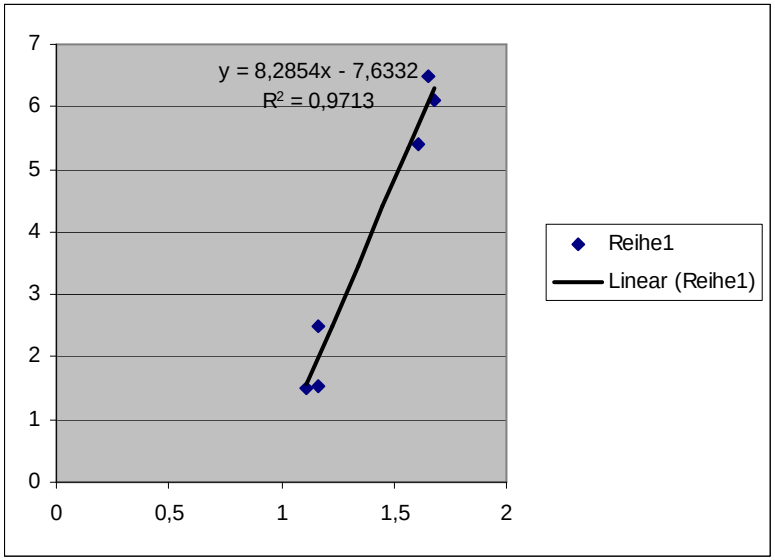
Graphemic usefulness of a letter (L-utility)

This is a kind of weighting for the previous properties. The weight can be ascribed in different ways, e.g. depending on the position in the grapheme.

For example the letter <t> in the grapheme <ght> can obtain the weight 3, or one can count the position in reverse order.

The greater the size of graphemes, the greater the L-utility.

Language	Graphemes	Letter	G/L	gra. size (1)	mean graphemic load	L-utility
German	68	**25**	**2,7200**	**1,68**	**4,12**	**6,12**
Swedish	57	**22**	**2,5909**	**1,61**	**3,36**	**5,41**
Slovene	28	**22**	**1,2727**	**1,11**	**1,27**	**1,5**
Italian	71	**25**	**2,8400**	**1,65**	**3,92**	**6,48**
Slovak	51	**26**	**1,9615**	**1,16**	**2,23**	**2,5**
Croatian	31	**22**	**1,4091**	**1,16**	**1,32**	**1,54**



R^2=0,97 (!)

Some more hypothesis:

- The greater the P-inventory and the smaller the L -inventory, the greater the mean graphemic load of letters.

- The relationship between P-inventory and L -inventory influences the mean size of graphemes.

- The greater the size of graphemes, the greater the graphemic load of letters.

- The greater the size of the graphemes, the greater the L-utility.

- The greater the inventory of graphemes, the greater the L-utility.

- The greater the G-load of a letter, the more frequently it occurs.

- The greater the graphemic utility of a letter, the more frequently it occurs.