universität wien

Emmerich Kelih

Institut für Slawistik

Loan Words: A quantitative linguistics perspective

IQLA
International Quantitative Linguistics Association

Aims of the study

- General remarks about linguistic studies of loan words/borrowings
- Quantitative contributions to loan word studies
- Relevance and importance of Piotrowski's law
- Loan words in the World's languages: Frequency studies
- Frequency of loan words in different lexico-semantical groups
- Theoretical modelling
- Summary
- Final ideas about Quantitative Linguistics perspectives in loan-word research

Loan words/ borrowings in the word languages

- Word forms (lexemes, words, parts of words, morphemes etc.) are imported from a donor language into a recipient language

- no „loan“ or “borrowing”, but incorporation/adaption of non-indigenous word forms

- terminologically different kinds of “borrowings” and “loans”: e.g. German tradition: “äußeres” Lehngut: direkte Entlehnungen (Fremdword, Lehnwort), Scheinentlehnungen (lexikalisch, semantisch), “inneres” Lehngut: Lehnbedeutung, Lehnübersetzung, Lehnübertragung etc.)

- different relevance of borrowings and loans in various linguistic traditions (German linguistics, English spoken world, high relevance for Slavic linguistics etc.)

- loan words/borrowings as one central possibility of the enlargement of the lexical stock of a language (additionally to derivation and neologisms)

- requirements – needs: coding-requirement, filling of lexical/semantical gaps, need for innovation, regulation of polysemy and synonymy, need for stylistic effects etc.

Quantitative Linguistics: contribution to loan word studies ?

– increase of loan words within one particular language follows an systematic quantitative trend

– incorporation of new lexical material follows a S-shaped curve (= slow beginning increase – turning point – decrease/stagnancy)

– mathematical interpretation by Russian linguist R.G. Piotrowskij (1922-2009)

– mathematical re-interpretation and systematization by Gabriel Altmann (Piotrowski-Altmann Law) (cf. Altmann 1983)

– **set of logistic equations for modelling different growth processes**

- vocabulary growth in texts
- chronological frequency changes of morphological forms
- language acquisition processes (L1-acquisation)
- dispersion and increase of loan words (diachronic aspects)

Some examples: **Arabic loanwords in German since 14th century**

Tabelle 1
Arabismen im Deutschen

Jhd.	t	beobachtet	kumuliert	berechnet
14.	1	38	38	34.0934
15.	2	14	52	56.3274
16.	3	32	84	83.4508
17.	4	26	110	109.8020
18.	5	21	131	130.3095
19.	6	14	145	143.6839
20.	7	5	150	151.4299
$a = 7.4099$	$b = 0.6964$	$c = 160$	$D = 0.996$	

$$(1) \quad p_t = \frac{c}{1 + ae^{-bt}}$$

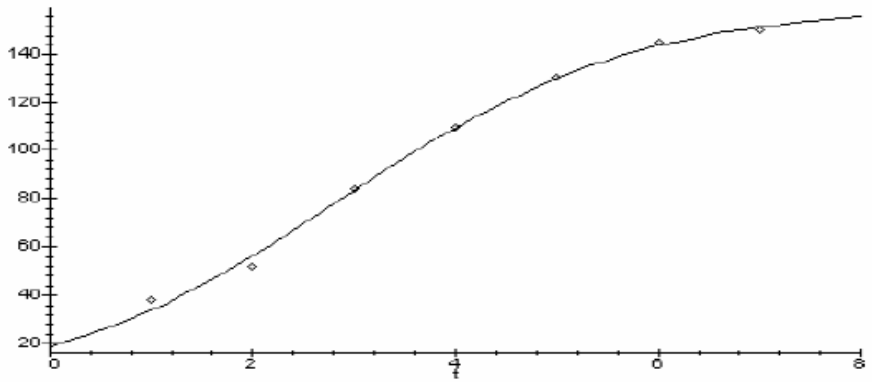


Abb.1 Die Entwicklung der Arabismen im Deutschen. (In dieser Graphik steht $t = 1$ für die Entlehnungen bis zum 14. Jahrhundert einschließlich, $t = 2$ für das 15. Jahrhundert; etc.)

Best, K.-H. (2004): Zur Ausbreitung von Wörtern arabischer Herkunft im Deutschen, *Glottometrics*, 8, 75-78.

Some Examples: **English loanwords in Russian since 1980th – 2004**

Tabelle 2
Anglizismen im Russischen. Berechnete Häufigkeiten

Jahr	t	Beobachtete Häufigkeiten	Berechnete Häufigkeiten
1980	1	10	11,3114787
1981	2	14	14,1547311
1982	3	19	17,6943503
1983	4	27	22,0906757
1984	5	36	27,535347
1985	6	43	34,2543396
1986	7	54	42,5094999
1987	8	74	52,5973802
1988	9	81	64,8437702
1989	10	89	79,5919748
1990	11	102	97,1827997
1991	12	119	117,924646
1992	13	132	142,053442
1993	14	151	169,684625
1994	15	172	200,762915
1995	16	195	235,019532
1996	17	260	271,949128
1997	18	349	310,818074
1998	19	374	350,710428
1999	20	405	390,608407
2000	21	433	429,493655
2001	22	463	466,448639
2002	23	498	500,737434
2003	24	532	531,852102
2004	25	549	559,521372

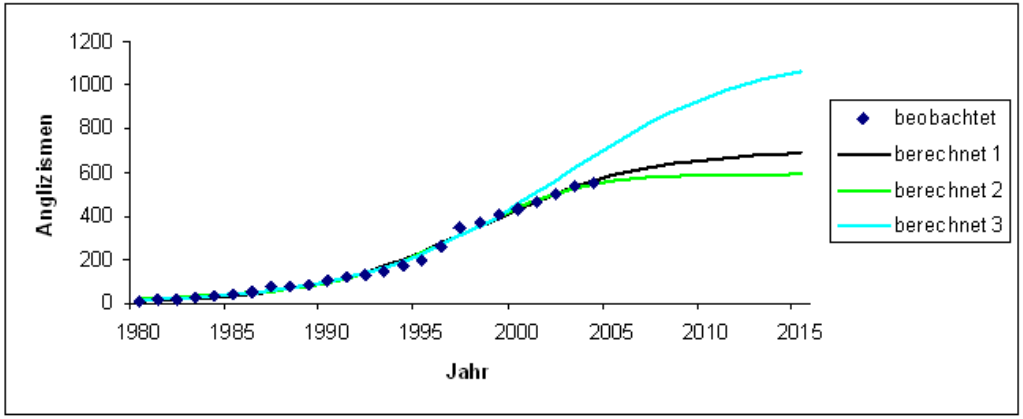


Abbildung 2. Prognose. Drei mögliche Kurven

→ **Many empirical studies confirm (similar) trends regarding the chronological incorporation of loans**

Stuhlpfarrer, Michael (2010): Anglizismen im Russischen, *Glottotheory* 3, 1, 97–109.

General linguistics: loan/borrowings

– Basic data about the frequency of loan words/borrowings (different registers, corpora)

– frequency counts – various aspects (POS – loans, length of loans etc.)

New input and approach by the project of Haspelmath/Tadmor (2009)

The World Loanword Database (WOLD)

It **provides vocabularies (mini-dictionaries of about 1000-2000 entries) of 41 languages from around the world**, with **comprehensive information about the loanword status of each word.** It allows users to find l**oanwords**, **source words** and **donor languages** in each of the **41 languages**, but also makes it easy to **compare loanwords across languages**. Each vocabulary was contributed by an expert on the language and its history. The database can be accessed by language, by meaning or by author.

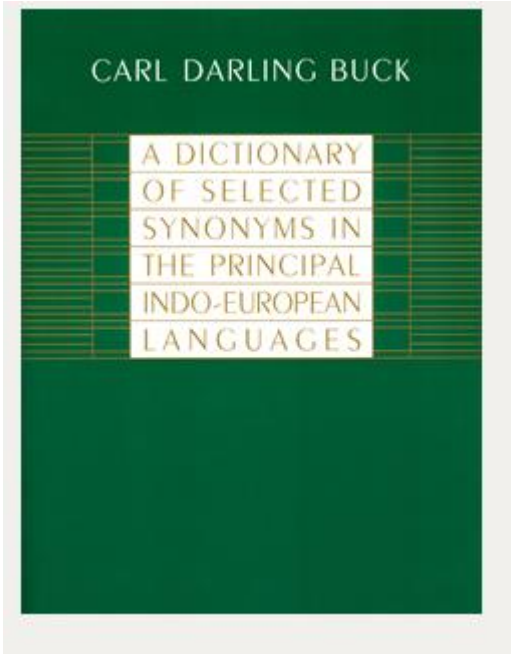
The list of **1460 meanings on which the vocabularies are based is called the Loanword Typology meaning list**, and it is in turn based on the list of the Intercontinental Dictionary Series.

The World Loanword Database is the result of a collaborative project coordinated by Uri Tadmor and Martin Haspelmath between 2004 and 2008, called the Loanword Typology Project (LWT).

The World Loanword Database (WOLD)

list of approx. 1500 meanings (given in English)

Buck, Carl Darling (1949): *A Dictionary of Selected Synonyms in the Principal Indo-European Languages.* Chicago.



- Special focus: Analysis of **basic vocabulary**
- material is grouped into **22 lexico/semantical subgroups**
- Word list, including meta-information about loan words status for over 40 languages

Some further details

Nr.	lexico-semantical group	Nr. of meangings
1	The physical world	75
2	Kinship	82
3	Animals	96
4	The body	157
5	Food and drink	81
6	Clothing and grooming	59
7	The house	48
8	Agriculture and vegetation	74
9	Basic actions and technology	78
10	Motion	82
11	Possession	46
12	Spatial Relation	75
13	Quantity	38
14	Time	57
15	Sense perception	49
16	Emotions and values	48
17	Cognition	51
18	Speech and Language	41
19	Social and political relations	36
20	Warfare and hunting	40
21	Law	26
22	Religion and belief	26
Sum		1460

Languages analysed:

Swahili, Iraqw, Gawwada, Hausa, Kanuri, Tarifiyt Berber, Seychelles Creole, Romanian, Selice Romani, Lower Sorbian, Old High German, Dutch, English, Kildin Saami, Bezhta, Archi, Manange, Ket, Sakha, Oroquen, Japanese, Mandarin Chinese, Thai, Vietnamese, White Hmong, Ceq Wong, Indonesian, Malagasy, Takia, Hawaiian, Gurindji, Yaqui, Zinacantán Tzotzil, Q'eqchi', Otomi, Saramaccan, Imbabura Quechua, Kali'na, Hup, Wichí, Mapudungun

Extra-category (added by Haspelmath/Tadmor 2009)

- Modern world
- Miscellaneous function words (both are not analysed)

What kind of meta-information information do we get?

	Upper-Sorbian		
meaning	**word form**	**borrowed status**	**source words**
world	swět	5. no evidence for borrowing	land 'land, country' New High German
land	land	1. clearly borrowed	kraj 'country, land' Upper Sorbian
land	kraj	2. probably borrowed	
soil	zemja	5. no evidence for borrowing	
dust	proch	5. no evidence for borrowing	
mud	kuŕ	5. no evidence for borrowing	
sand	pěsk	5. no evidence for borrowing	
mountain or hill	góra	5. no evidence for borrowing	
cliff or precipice	skała	5. no evidence for borrowing	
plain,	rownina	5. no evidence for borrowing	
field	płonina	5. no evidence for borrowing	
valley	doł	5. no evidence for borrowing	
island	kupa	5. no evidence for borrowing	
...	...	...	...

1. Absolute frequency of loan words within one lexico-semantical group
2. total amount of loan words per language

→ **quantitative loan word profile**

„However, different languages display a remarkable degree of consistency which regard to which fields are more or less affected by borrowing. While there are certainly cross-linguistic differences, most languages tend to borrow more words into similar fields, and the same fields turn up again as the ones most resistant to borrowing. (Tadmor 2009: 64)

frequently affected by loans:

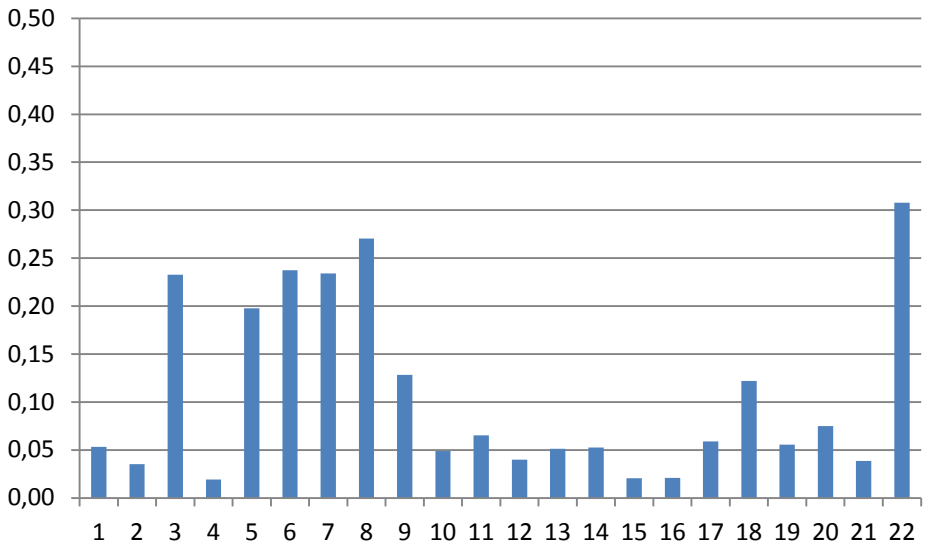
1. Religion and belief
2. Clothing and grooming
3. House/living

less affected by loans:

1. The body
2. Spatial relation
3. Sense perception
4. Function words with deictic function

Case study: Loan Words in Slovene (South Slavic, Indo-European)

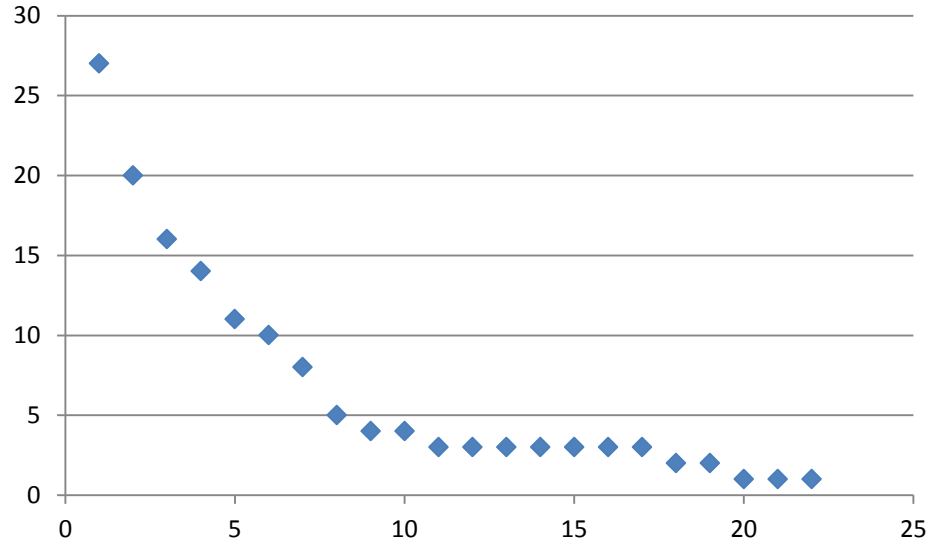
Nr.	lexico-semantical group	Nr. of meangings	abs. f.
1	The physical world	75	4
2	Kinship	82	3
3	Animals	96	27
4	The body	157	3
5	Food and drink	81	16
6	Clothing and grooming	59	14
7	The house	48	11
8	Agriculture and vegetation	74	20
9	Basic actions and technology	78	10
10	Motion	82	4
11	Possession	46	3
12	Spatial Relation	75	3
13	Quantity	38	2
14	Time	57	3
15	Sense perception	49	1
16	Emotions and values	48	1
17	Cognition	51	3
18	Speech and Language	41	5
19	Social and political relations	36	2
20	Warfare and hunting	40	3
21	Law	26	1
22	Religion and belief	26	8
Sum		1460	147



rel. frequency of loanwords in lexical-semantical groups

- language specific behaviour
- 10% loan words in basic vocabulary

Transformation to a rank frequency distribution



– linguistically a birth and death process (Krylov 1982, Altmann 1985, Altmann 1997, Altmann/Köhler 1997, Krlyov 2002)

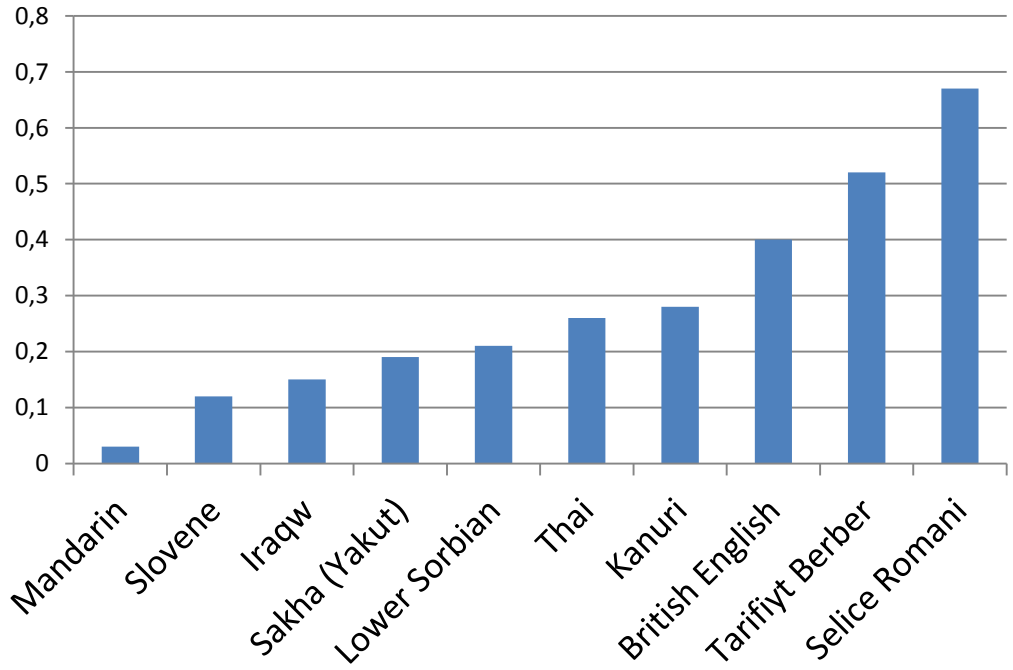
– integration/adaption of new loan words

– elimination/replacement of loan words (puristic movements, lost coding requirement)

– appropriate model: **negative binomial distribution**

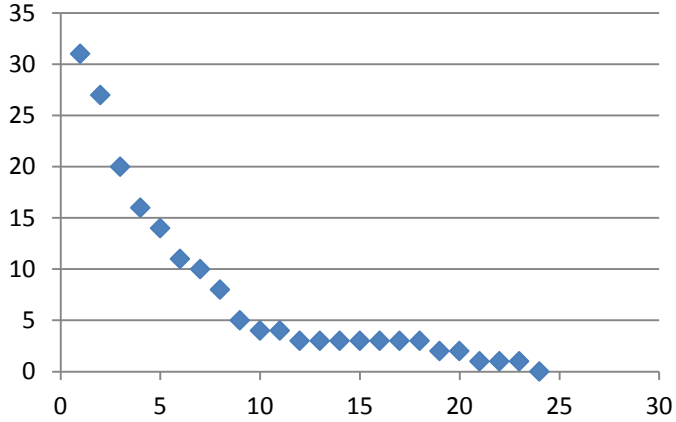
Case study: Loan words in the world languages (selected languages)

Nr.	Language	% Loan words	Nr.	Language	% Loans
1	Mandarin (Chinese)	3.00	6	Thai	26.10
2	Slovene	10.00	7	Kanuri	28.00
3	Iraqw	15.00	8	British English	40.00
4	Sakha (Yakut)	19.00	9	Tarifiyt Berber	52.00
5	Upper-Sorbian	21.00	10	Selice Romani	67.00

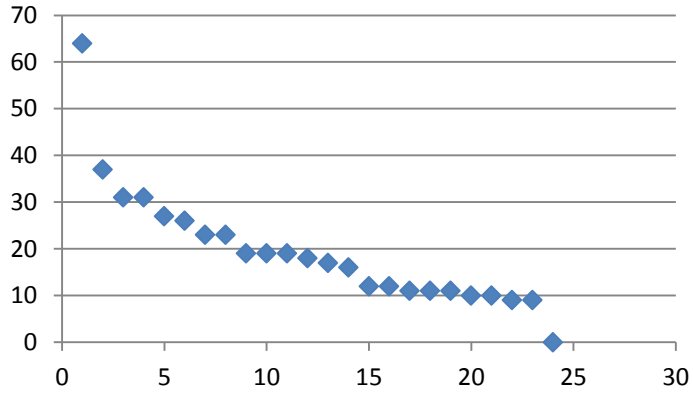


One common model? Or different models required?

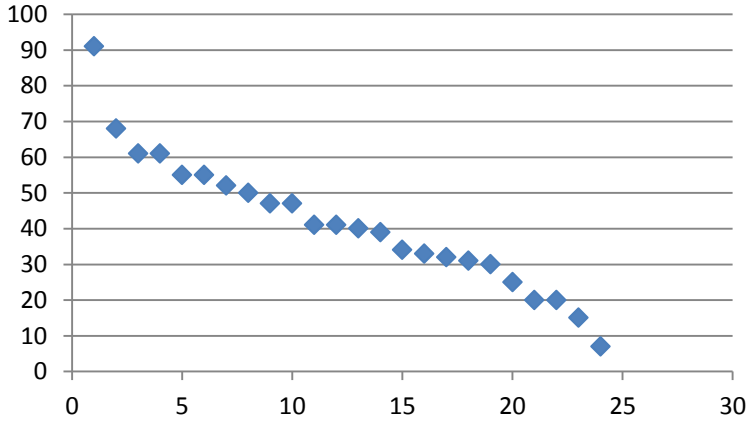
Case study: Quite different rank-frequency profiles



Slovene

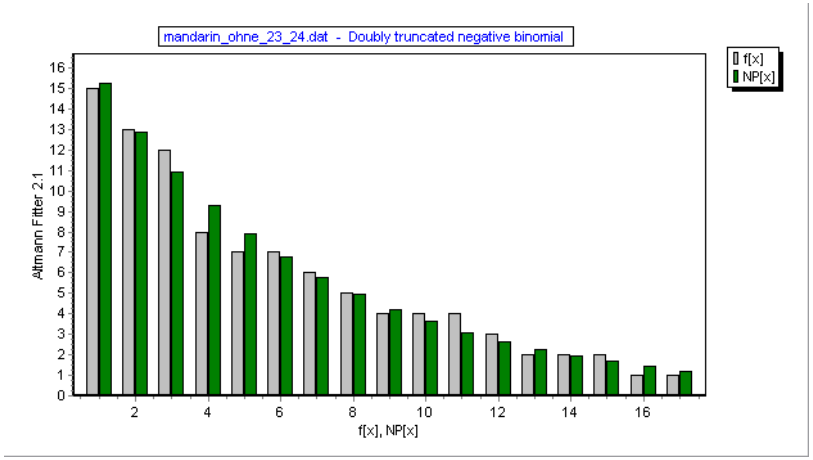


Thai

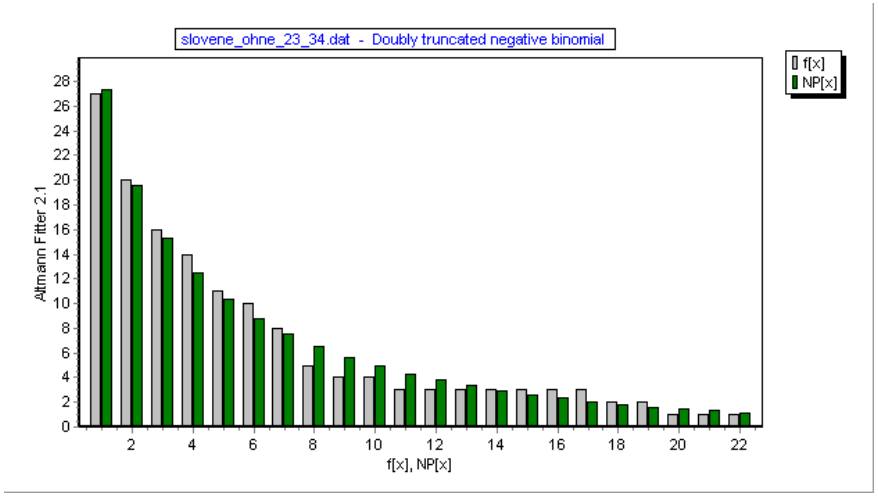


Selice Romani

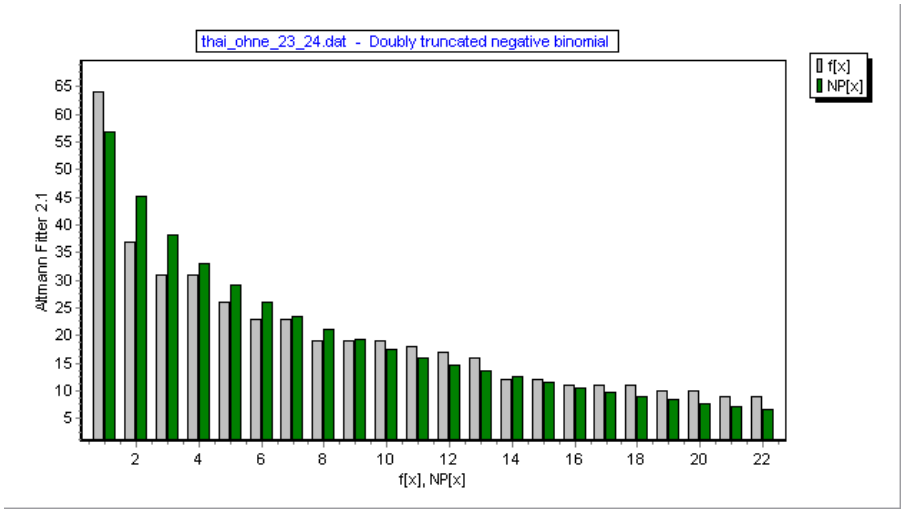
But ONE MODEL: Negative Binomial-Distribution (doubly truncated)



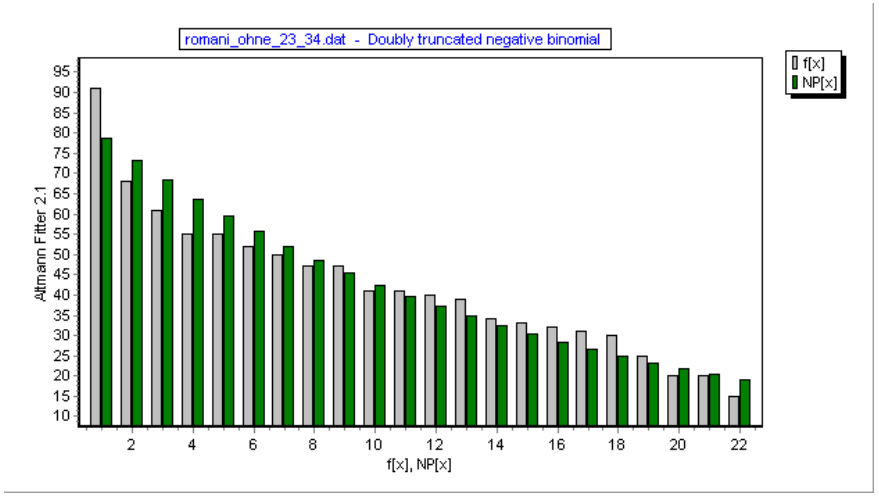
Mandarin P = 0.99



Slovene P = 0.99



Thai P = 0.94



Selice Romani P = 0.92

But ONE MODEL: Negative Binomial-Distribution (doubly truncated)

Nr.	Language	k	p	R	X^2	DF	P
1	Slovene	0.5734	0.0888	22	3.07	17	**0.99**
2	Upper-Sorbian	0.7740	0.0551	22	12.08	17	**0.79**
3	Kanuri	1.0177	0.0765	22	7.60	17	**0.97**
4	Britisch English	0.8607	0.0417	22	11.49	17	**0.82**
5	Sakha (Yakut)	0.7240	0.0404	22	7.22	17	**0.98**
6	Selice Romani	0.9870	0.0637	22	9.50	17	**0.92**
7	Tarifiyt Berber	0.9783	0.0365	22	2.65	17	**0.99**
8	Mandarin	0.9663	0.1416	17	1.04	12	**0.99**
9	Iraqw	0.7013	0.0927	18	1.77	13	**0.99**
10	Thai	0.689	0.06	22	8.71	17	**0.94**

→ in all cases P > 0.79

→ 8/10 P > 0.90

→ one overall model for all analysed languages

Summary of results

- from raw data to modelling
- from frequency-based approaches to Quantitative Linguistics
- going beyond the Piotrowski-law
- systematic regulation of loan words within lexico-semantical groups
- proposed birth/death process seems to be appropriate
- negative binomial distributions fits well

Perspectives

– more languages has to be analysed

– etymological spectrum has to be analysed (number of donor languages)

– synergetic approach has to be implemented

– developing further hypotheses regarding loan words: frequency, age, degree of integration/adaption, survival rate of loans, disintegration of loans …