Glottometrics 16, 2008, 46-56

Modelling polysemy in different languages: a continuous approach

Emmerich Kelih[1]*, Graz*

Abstract. This paper investigates the frequency of polysemy in six genetically unrelated languages. It can be shown that these distributions can be described by a power model, developed by the Estonian linguist Juhan Tuldava. Furthermore, interrelations between descriptive parameters of the analyzed empirical distributions have been obtained. Special attention has been paid to the behaviour of the parameters of the theoretical models, taking into account different influence factors (language analyzed, sample size, parts of speech).

Keywords: frequency of polysemy, parameter behaviour, interrelations

0. Introduction

Since Zipf (1935, 1949) it is a well known fact that the number of meanings follows certain regularities. These regularities can be explained by two opposing forces: (1) the forces of unification and (2) the forces of diversification; e.g. the most efficient way for the speaker would be one word with many different meanings, and for the hearer, one word with only one meaning. The first "force" (i.e. the speaker's economy) tends to reduce the effort of encoding, while the second "force" (i.e. the hearer's economy) leads to a minimization of decoding effort. In other words, we are concerned with the principle of least effort, which competes with the necessity to communicate efficiently. The interaction of the forces of unification and diversification results – as on every other level in language – in a compromise in the form of self organization, e.g. in a specific shape of probability distributions of the number of meanings of words.

The focus in our contribution is not on a discrete model[2] for this distribution, but rather on an empirical re-analysis of a continuous model, developed by Tuldava (1979, 1998). According to Wimmer/Altmann (2005:792) it is basically irrelevant, whether linguistic regularities are being described by discrete or by continuous models. Both approaches are approximations to linguistic reality and they are – as shown in Mačutek/Altmann (2007) – transformable into one another. However, neither such theoretical problems nor a survey of the state of the art needs to be presented here. Cf. Levickij (2005) and Hoffmann (2001) for a comprehensive overview of quantitative studies of lexical polysemy.

The focus of this paper will be on the following problems:

1. empirical verification of the continuous model, developed by Tuldava (1979, 1998),
2. the integration of his model in a synergetic approach,
3. the interrelations between descriptive parameters of the analyzed empirical distributions,
4. the behaviour of the parameters of the theoretical models, taking into account the following factors of influence, such as the language analysed, the sample size of the examined dictionaries and parts of speech.

[1] Address correspondence to: Inst. für Slawistik, Merangasse 70/1; 8010 Graz. Austria. E-mail: emmerich.kelih@uni-graz.at

[2] Cf. the theoretical deduction of adequate discrete models in Wimmer/Altmann (1999) and the empirical corroboration of these models in Kelih (2007).

1.1 Continuous model for polysemy

One of the well known models[3] for the frequency of polysemy was developed by Tuldava (1979) and Tuldava (1998: 120). He postulates this modified exponential function to be adequate for modelling the number of meanings:

(1) $$. y = ae^{-b\sqrt{x}}$$

In formula (1) *y* denotes the relative frequency of words with a given number of meanings, *x* the number of meanings; *a* and *b* are parameters and *e* is the basis of the natural logarithm. According to Tuldava (1998: 120) the root of the number of meanings is a new measurement unit of the "semantic extent". The sequence of natural numbers 1,2,3 … (i.e. the root of 1,2,3…) marks different degrees of polysemy in languages.

Tuldava (1998:120) calculated, using the above mentioned formula, the theoretically expected relative number of meanings for three languages (Russian, Hungarian, English)[4]. All data are – as is the practice in quantitative analysis of polysemy – based on monolingual explanatory dictionaries. The starting point for the modelling is the relative frequency of words with 1,2,3 … meanings. Further details are given in Table 1. However, Tuldava (1979, 1998:120) did not perform a test or give an indicator which would give deeper information on the goodness of fit of the tested models. Therefore, we re-analyzed the data from Tuldava (1998, 1979), using an iterative approximation of the parameters *a* and *b* and calculating the determination coefficient *D* (cf. Table 1).

Table 1
Re-analysis: Data by Tuldava (1979, 1998)

x	**English**		**Hungarian**		**Russian (verbs)**	
	y		y		y	
	obs.	exp.	obs.	exp.	obs.	exp.
1	0.427	0.4263	0.504	0.5155	0.615	0.622
2	0.203	0.2049	0.265	0.2251	0.254	0.2159
3	0.117	0.1168	0.118	0.1192	0.071	0.0959
4	0.072	0.0727	0.052	0.0697	0.03	0.0483
5	0.048	0.0479	0.024	0.0435	0.013	0.0265
6	0.035	0.0328	0.013	0.0284	0.007	0.0153
7	0.023	0.0232	0.008	0.0192	0.003	0.0093
8	0.016	0.0168	0.005	0.0133	0.002	0.0058
9	0.013	0.0124	0.003	0.0094	0.002	0.0038
10	0.009	0.0093	0.002	0.0068	0.002	0.0025
11	0.0073	0.0071	0.0014	0.005	0	0.0017
12	0.006	0.0055	0.0012	0.0037	0	0.0011
13	0.0053	0.0042	0.009	0.0028	0	0.0008
14	0.0034	0.0033	0.007	0.0021	0.01	0.0006
15	0.0032	0.0026	0.007	0.0016	0	0.0004
> 15	0.014	0.0021	0.002	0.0013	0	0.0003
parameter	a	b	a	b	a	b
	2.4997	1.7688	3.8117	2.0007	8.0033	2.5546
D	0.9992		0.9891		0.9924	

[3] Further continuous models were developed by Krylov/Jakubovskaja (1977) and Polikarpov (1987).
[4] Tuldava (1979, 1998) took the data for Russian (verbs) from Krylov/Jakubovskaja (1977), for English from Višnjakova (1976) and for Hungarian from Papp (1967).

For all the three languages a determination coefficient $D > 0.98$ is obtained (cf. Table 1). This result must be interpreted as a convincing empirical verification of the model proposed by Tuldava (1979, 1998).

This first positive result is our starting point for a further empirical analysis on a larger data basis: we have based our study on 45 polysemy frequency distributions from Russian, English, German, Maori, Hungarian and Polish (cf. for details see Table 2).

Table 2
Analyzed languages and used resources

No.	Language	Specification[5]	Sample size[6] (N)	Source
1		Dic.-comp.; Ve;	2765	
2		Dic.-comp., No.,	3278	Levickij et al. (1999)
3		Dic.-comp., Adj.;	490	
4		Dic.-comp.; (no. 1-3)	6533	
5	Maori	Dic.-comp.;	7689	Wimmer/Altmann (1999)
6		Dic.-comp.; 11-14. century	2394	
7		Dic.-comp.; 15.-17. century	2953	
8	Russian	Dic.-comp.; 18. century	3420	Andreevskaja (1990)
9		Dic.-comp.; 19. century	4110	
10		Dic.-comp; 20. century	4185	
11		Dic.-comp., (no. 6-10)	17062	
12		Dic.-comp.; Adj.;	7191	
13		Dic.-comp.; Adv.;	287	
14	English	Dic.-comp.; No.;	15673	Višnjakova (1976)
15		Dic.-comp.; Ve.;	2796	
16		Dic.-comp.; (no. 12-16)	25947	
17		Dic.-comp.; SO; Ve.;	9502	
18		Dic-sa.; SO; Ve. (I,K,S);	1329	Krylov/Jakubovskaja (1977)
19		Dic-sa.; SSRLJA; Ve. (I,K,S);	2711	
20		Dic.-comp; SO; Ve.;	10570	
21	Russian	Dic.-comp; SO; No.;	16748	Krylov (1982)
22		Dic.-comp; (no. 20-21)	32559	
23		Dic.-comp; MAS;	82159	
24		Dic.-comp; SO (9. edition)	57003	
25		Dic.-comp; SSRLJA;	120481	Polikarpov (1987)
26	English	Dic.-comp; HO;	44372	
27		Dic.-comp; Sho;	79801	
28	Russian	Dic-sa.; MAS;	3931	Polikarpov/Krjukova (1989)
29		Dic-sa.; MAS; Adj.	431	
30		Dic-sa.; MAS; Adv.;	138	
31		Dic-sa.; MAS; No.;	1716	
32		Dic-sa.; MAS; Ve.;	1613	
33		Dic-sa.; MAS	3203	

[5] The abbreviations are as follows: Dic.-comp.: complete dictionary; Dic-sa.: sample from dictionary; No.: nouns; Ve.: verbs; Adj.: adjectives; Adv.: Adverb; I,K,S: sample of lexemes with initial letters I, K and S.; SO: Slovar Ožegova, SSRLJA: Slovar' sovremennogo russkogo literaturnogo jazyka; MAS Slovar' russkogo jazyka pod. red. A.P. Evgen'evoj; HO: Hornby: Oxford Advanced Learner's Dictionary of Current English; Sho: Shorter Oxford English Dictionary. See the bibliographical references for further details on used issues, edition etc.

[6] The sample size N is the number of analyzed words.

34		Dic-sa.; SO;	3971	
35		Dic-sa., SO; Adj.	446	
36		Dic-sa.; SO; Adv.	136	
37		Dic-sa.; SO; No.;	1731	
38		Dic-sa.; SO; Ve.;	1630	
39	German	Dic-sa.; No.;	5919	Schierholz (1991)
40	Hungarian	Dic.-who.;	59574	Papp (1967)
41		Dic-sa.; No.;	13356	
42		Dic-sa.; Ve.;	6053	
43	Polish	Dic-sa.; Adj.;	8777	Hammerl (1991)
44		Dic-sa.; Adv.	1391	
45		Dic.-who.; (no. 41-45)	29577	

Our specific choice of data allows us to analyze whether the discussed Tuldava model is suitable for all the different languages used in this study. Analyzing the 45 data samples by calculating the parameters *a* and *b* (iterative approximation) and the determination coefficient, we get a very clear result: the average determination coefficient $\overline{D}$ = 0.9950, with a minimum of *D* = 0.9704 and a maximum of *D* = 0.9999. In other words, the model proposed by Tuldava (1998, 1979) seems to be suitable and adequate for all six languages.

The calculated determination coefficient *D* and the parameter *a* and *b* for every analyzed sample are in Table 3; furthermore, we have included two additional descriptive parameters, the average polysemy $\overline{x}$ and the relative frequency of words with only one meaning p_1. These two parameters will be used in a further analysis in chapter 2.

Table 3
Descriptive, theoretical parameters and *D*

No.	$\overline{x}$	$\mathbf{p_1}$	**D**	**a**	**b**
1	2.0886	0.5009	0.9904	3.77	-2.00
2	2.0799	0.4793	0.982	3.31	-1.90
3	2.2959	0.4347	0.9876	2.5	-1.72
4	2.0998	0.4851	0.988	3.42	-1.93
5	1.5763	0.6647	0.9997	12.47	-2.93
6	1.6817	0.7026	0.9978	24.47	-3.55
7	1.3356	0.786	0.9999	45.42	-4.06
8	1.2535	0.8228	0.9999	69.53	-4.44
9	1.2545	0.8236	0.9999	72.38	-4.48
10	1.2645	0.8117	0.9999	58.05	-4.27
11	1.3307	0.797	0.9999	54.1	-4.22
12	2.5574	0.4148	0.9918	2.24	-1.66
13	1.4286	0.7108	0.9954	17.64	-3.21
14	2.1437	0.5552	0.9999	5.98	-2.38
15	3.5293	0.2711	0.9712	0.91	-1.13
16	2.3997	0.4821	0.9997	3.55	-2.00
17	1.642	0.6151	0.9913	7.98	-2.55
18	1.6561	0.6185	0.9922	8.27	-2.58
19	2.1498	0.5242	0.9916	4.69	-2.18
20	1.5553	0.6662	0.9981	12.24	-2.91
21	1.372	0.7477	0.9994	26.05	-3.55

22	1.434	0.7204	0.9993	19.98	-3.32
23	1.503	0.7293	0.9998	26.1	-3.58
24	1.3764	0.7748	0.9999	41.24	-3.97
25	1.6973	0.634	0.9992	9.89	-2.74
26	1.3596	0.8161	0.9992	94.41	-4.75
27	2.0114	0.576	0.9999	6.81	-2.47
28	1.6566	0.6115	0.995	7.81	-2.54
29	1.5128	0.6473	0.9907	9.72	-2.70
30	1.3841	0.7101	0.995	16.07	-3.12
31	1.4953	0.6824	0.9977	13.73	-3.00
32	1.8574	0.5226	0.9868	4.23	-2.07
33	1.3562	0.6912	0.9704	12.61	-2.89
34	1.4077	0.7318	0.9993	22.38	-3.42
35	1.287	0.7848	0.9983	37.4	-3.86
36	1.2059	0.8162	0.9973	47.27	-4.06
37	1.3114	0.777	0.9994	35.38	-3.82
38	1.5595	0.6644	0.9988	12.22	-2.91
39	2.7363	0.4396	0.9998	2.68	-1.81
40	1.9455	0.5073	0.9862	3.85	-2.00
41	1.5419	0.6664	0.9986	12.19	-2.90
42	1.9091	0.5174	0.9888	4.10	-2.05
43	1.2706	0.8212	0.9999	73.88	-4.50
44	1.2919	0.8009	0.9999	52.49	-4.19
45	1.5248	0.6882	0.9998	15.37	-3.11

1.2 Integration of Tuldava's model into the Wimmer/Altmann approach

In addition to the first empirical findings it will be shown that Tuldava's model (1979, 1998: 120) can easily be integrated into the theoretical framework of Wimmer/Altmann (2005). According to Wimmer/Altmann (2005: 795) the model of Tuldava (1979, 1998) is a special case of a more common formula ("unified theory"). Hence this law of polysemy is a special case, which can be deduced from formula:

$$\frac{dy}{y} = \frac{-bc}{x^{1-c}}dx \quad .$$

It results in

(2) $y = Ce^{-bx^c}$, whereas Tuldava (1998: 120) fixes c = ½. So the model gets this final form:

(3) $y = Ce^{-b\sqrt{x}}$.

As shown above, this model describes the behaviour of the polysemy distribution in all six languages and thus the Wimmer/Altmann (2005) approach has been indirectly confirmed.

In the next chapter the attention will be drawn to the interrelation between parameters from the empirical frequency distributions and the statistical behaviour of the parameters *C* and *b*, that is in our case, the parameters *a* and *b*.

2. Empirical findings: Interrelations

2.1. Interrelation between relative frequency of words with one meaning and average polysemy

The frequency distribution of polysemy in the analyzed languages does not only follow a general, theoretically integrated model, but also shows an interesting and systematic picture with respect to the behaviour of the empirical parameters.

In dealing with frequency data of polysemy we are concerned with a *natural rank frequency*, e.g. the occurrence of meanings is a monotone decreasing curve from the first frequency class on. It is very likely that this monotony is responsible for a direct interrelation between the average polysemy $\overline{x}$ and the relative frequency of words with one meaning p_1. A priori we postulate that with a decreasing mean value $\overline{x}$ the frequency of p_1 increases. Interestingly enough we have not obtained a linear interrelation, but a monotonous decreasing power function. Cf. the visualization of this relation in Figure 1.

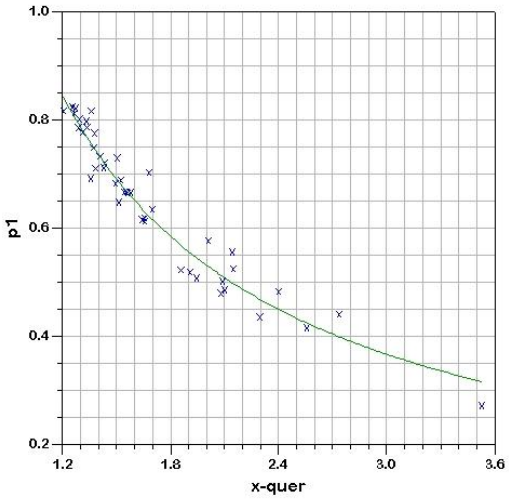

Figure 1: Dependency of $\overline{x}$ on p_1

A simple power function in the form $p_1 = c^d$ suffices to describe this interrelation. With d = -0.9138 a satisfying D = 0.95 can be obtained. This result is a strong empirical evidence for a harmonious relation between $\overline{x}$ and p_1.

Of course, it is certain that adding more data to our analysis the parameter will shift, but nevertheless we propose that the curve will definitely have a similar shape as the above one. So the forces of self organization are observable on the descriptive level already. The relative frequency of words with one meaning is predictable with only the mean value of the distribution.

2.2 Interrelations between empirical parameters and parameter *a*

In addition to the described relations above on the empirical level some more dependencies between the parameter *a*, the mean value $\overline{x}$ and the relative frequency of words with one meaning p_1 have been noticed. A priori we postulate a direct dependency between the parameter *a* and p_1, because the parameter *a* controls the "shift" of the curve on the *y*-axis. Hence the frequency of words with *one* meaning is controlled by *a*. Therefore it should hold true that with a decrease of p_1 the parameter *a* also decreases and because of the known dependency of p_1 on $\overline{x}$ (cf. Figure 1) the parameter *a* increases with a decreasing mean value $\overline{x}$. These two assumptions are already confirmed in a visualization of the mentioned dependencies (cf. Figure 2a and 2b).

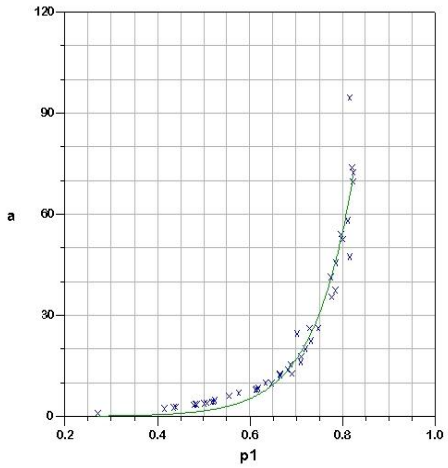

Figure 2a
Interrelation between p_1 and parameter *a*

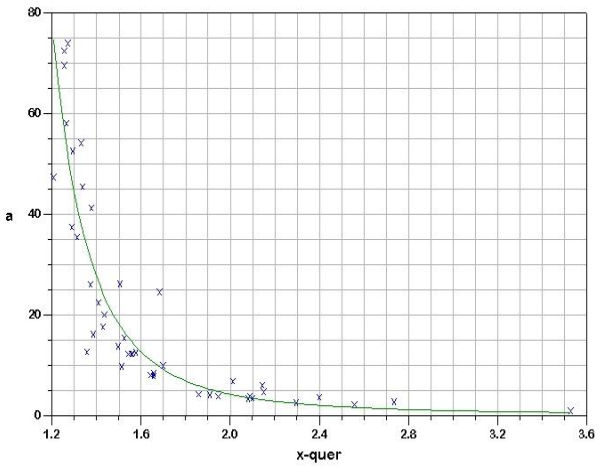

Figure 2b
Interrelation between $\overline{x}$ and parameter *a*

The first interrelation between p_1 and the parameter *a* can be captured by the simple formula: $a = c \exp(dp_1)$ with $D = 0.94$ (parameter $c = 0.0048$ and $d = 11.68$). For the interrelation between $\overline{x}$ and the parameter *a* (Table 2) the model $a = g \exp(-h / \overline{x})$ is suitable: With the parameters $g = 0.0553$ and $h = -8.6909$ a reliable[7] $D = 0.82$ is obtainable (cf. Figure 2b).

From Figure 2a it can be seen that from approximately $p_1 > 0.80$ the parameter *a* rises sharply. This observation is explainable by the fact that at this point a minimum of polysemy is reached. Above this point a "normal" and efficient communication is probably no longer possible. A similar behaviour is shown by the mean polysemy $\overline{x}$, which may never equal 1, since in this case a language would have no polysemy at all, e.g. this would lead to a severe complication and inefficiency of the communication act. Therefore the self-regulated behaviour of *a*, p_1 and $\overline{x}$ is a necessary precondition of the language system.

3.3. Parameter *a* and *b:* language specificity

The specific behaviour of the parameter *a* is the starting point for further analysis of this parameter. In chapter 1 a general cross linguistic valid model, based on Tuldava's approach, has been found. Even if the existence of polysemy is supposed to be a "linguistic universal" (cf. Levickij 2006: 161f.; Croft 2003), the question of the language specificity of polysemy-distributions must be raised. In other words, are the parameters of our model specific for a certain language or not? In case of such specificity the conceptual power of our approach would be confirmed, since the general and the specific behaviour of the frequency distribution can be described at the same time.

To get an impression about this behaviour, we have calculated the mean values of the parameters $\overline{a}, \overline{b}$ from Tuldava's model and the mean value of the average polysemy $\overline{x}(1)$ for Russian, German, English and Polish. For Maori and Hungarian the single values were taken as the basis for our interpretation (cf. Table 4). Because of an unbalanced number of sources per language the following comments and interpretations are preliminary and should be understood as a first attempt to a parameter interpretation of polysemy distributions.

[7] Dataset no. 26 has been excluded from the analysis, because of its unusual behaviour of the parameter *a* (outlier). Qualitatively (i.e. concerning the homogeneity of the data, type of the dictionary etc.) this decision cannot be justified for the time being.

Table 4
Number of sources, parameter *a* und *b* and mean values

Language	**Number of sources (*n*)**	**Parameter $\bar{a}$**	**Parameter $\bar{b}$**	$\bar{x}(1)$
Polish	5	31.6094	-3.3480	1.5077
Russian	26	26.8922	-3.3372	1.4823
English	7	18.7896	-2.5132	2.2042
Maori	1	12.4700	-2.9300	1.58
Hungarian	1	3.8460	-2.0000	1.95
German	5	3.1360	-1.8699	2.2601

For the time being only a simple qualitative interpretation of the parameters can be offered: The parameter $\bar{a}$ shows clear language specificity, because the values from all languages differ widely. We get the following "order" of languages: Polish, Russian, English, Maori, Hungarian and German (cf. Table 4). Due to the unbalanced number of sources (n) no deeper statistical analyzes are possible. Nevertheless, it is noticeable that the range of the parameter $\bar{b}$ is shorter than the range of parameter *a*. Furthermore, a direct, but statistically not significant, dependency between the parameter $\bar{a}$ and $\bar{b}$ is observable. Thus we postulate that both parameters contain some information about the languages examined. This assumption is supported by the fact that due to the different morphological structures of the languages (and presumably in dependency of the word length) polysemy is adopted in different ways. So it is very likely that morphology does have a significant influence on the specific shape of the distribution of polysemy. See also the considerations by Polikarpov (1979) on polysemy in dependency on the language type (analytic vs. synthetic).

3.4. Parameter *a* and *b*: sample size

The next step deals with the question: to what extent does the sample size of the analyzed dictionaries influence the parameters. We hypothesize that polysemy increases with an increasing lexicon size, since a larger dictionary should contain more meanings than a smaller one. To analyze this assumed relation only data from complete dictionaries will be used (data no. 5-10, 16, 23-27, 40, 45).

In fact empirically neither between the sample size N and the parameter *a*, nor between N and parameter *b* a dependency has been observed (cf. Figure 3a and 3b). One reason for the missing dependency is the high variation of the parameters. Another factor could be the small number of analyzed dictionaries.

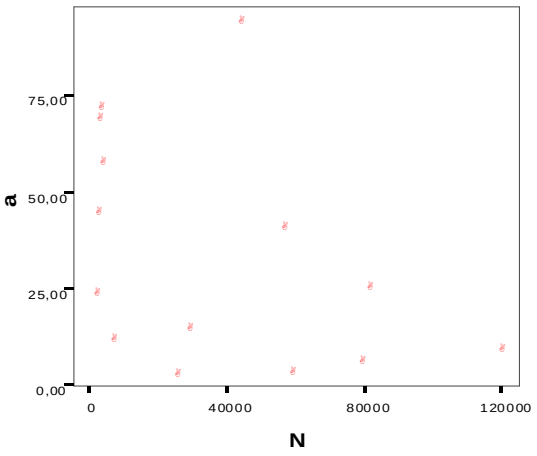

Figure 3a. Interrelation between N and a

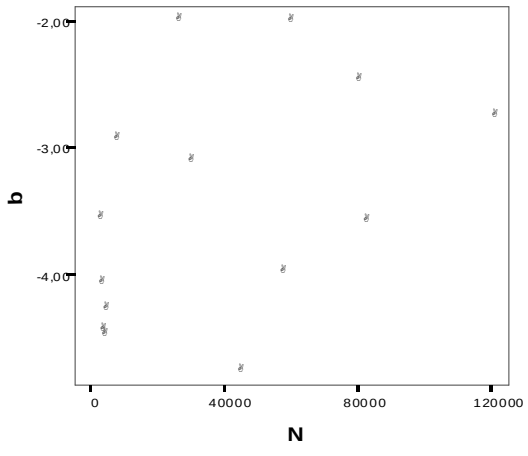

Figure 3b. Interrelation between N and b