Glottometrics 31, 2015, 31-37

A continuous model for polysemy

Emmerich Kelih
Gabriel Altmann

Abstract. In the article it will be shown that modeling polysemy in language can be performed in a rather unified way.

Keywords: diversification, polysemy, Zipf-Alekseev function, exponential function

Polysemy is a special case of diversification which exists in all domains of language and is one of the steady evolution processes of language. According to the given domain and type of language, one can model this phenomenon in quite different ways (cf. e.g. Altmann 1985, Altmann, Best, Kind 1987, Rothe 1991 and newer developments). There are such a number of boundary conditions that different models may be attained by simply changing the condition in the underlying difference or differential equations. However, the aim of the research is to restrict the analysis to a special domain and strive for a unified model – if possible.

It is usual to model the distribution of polysemy by means of a discrete approach, e.g. by means of a Poissonian birth-and-death process, because one may suppose that the individual meanings are discrete, at least in the dictionary; they are born, live and die. The fact that each meaning is rather a multidimensional space with a number of degrees, or still better, a continuous spectrum capturing a sector of the reality, may be ignored in order to show that there is some law behind polysemy. Hence one may consider polysemy also a continuous variable and the number of meanings of the word (x = 1,2,3,…) as averages or lower/upper bounds of an interval. In that case one can propose a continuous function without normalization. Such a step does not change anything in our demands for "truth" because no mathematical model expresses the truth, it is merely a means for capturing a phenomenon in such a way that we understand it, can insert it in a background theory and link it with other phenomena. Besides, there are no discrete or continuous phenomena in the reality and what more, discrete models can be transformed in continuous ones and vice versa (cf. Mačutek, Altmann 2007).

Here we shall restrict ourselves to semantic diversification or polysemy and use only published data. We lean against the unified theory (cf. Wimmer, Altmann 2005), i.e. we start with the simple differential equation

$$\frac{dy}{y} = \frac{g(x)}{h(x)}dx, \tag{1}$$

where $g(x)$ represents a function associated with the topical norm in language and the influence of the speaker, $h(x)$ is the control of the community/hearer striving for equilibrium, y is the number of cases with polysemy x. Since we search for a continuous model, we may omit all polysemies whose number is zero and adapt (1) to

$$\frac{dy}{y-1} = \frac{g(x)}{h(x)}dx. \tag{2}$$

The right hand side of the differential equation can consist of $g(x) = A + B*logX$ where A is the constant of the language, B is the force of the speaker influencing not directly X but only its logarithm, and $h(x) = CX$ is the equilibrating force of the speaker or the community. Solving the differential equation and reparametrizing it we obtain the well know Zipf-Alekseev function used already in different domains of linguistics (cf. e.g. unit length in Popescu, Best, Altmann 2014). The logarithm is probably associated with the human perception. The resulting function is

$$(3) \qquad y = cx^{a + b \log x}$$

or, omitting classes having zero occurrence and striving for theoretical values greater than or equal to 1,

$$(4) \qquad y = 1 + cx^{a + b \log x},$$

which yields always values great than or equal to 1. Function (3) admits computed values smaller than 1 but mostly the fitting is satisfactory. The results of fitting the above functions to classes or the whole dictionary in 6 languages are presented in Table 1. The data may be found in the quoted references.

Table 1a
Semantic diversification of word classes

Source	**a**	**b**	**c**	**R^2**
German verbs (Levickij, Kiiko, Spolnicka (1996) (4)	-2.1665	-1.4556	9878.1011	0.9998
German verbs: Texts (Levickij, Drebet, Kiiko 1999) (4)	1.0655	-1.2826	146.1013	0.9948
German nouns (Levickij, Drebet, Kiiko 1999) (4)	-0.0636	-1.0846	1572.1306	0.9982
German nouns; Texts (Levickij, Drebet, Kiiko 1999) (4)	1.0894	-1.0973	160.0230	0.9874
German nouns (Schierholz 1991) (3)	-0.8306	-0.3526	2599.7945	0.9997
German adjectives (Levickij, Drebet, Kiiko 1999) (4)	-0.1812	-0.8249	212.1173	0.9993
German adjectives: Texts (Levickij, Drebet, Kiiko 1999) (4)	1.9881	-1.3448	20.4406	0.9883
Russian 11-14th (Andreevskaja 1990) (3)	-2.2456	0.0073	1681.2993	0.9996
Russian 18th (Andreevskaja 1990) (4)	-1.8601	-1.1123	2813.0288	1.0000
Russian 19th (Andreevskaja 1990) (4)	-2.1491	-0.7644	3383.9869	1.0000
Russian 20th (Andreevskaja 1990) (4)	-1.8760	-0.9490	3396.9807	1.0000
Russian all (Andreevskaja 1990) (4)	-2.1535	-0.5548	13597.9953	1.0000
English adjectives (Višnjakova 1976) (4)	-0.2816	-0.6839	2982.4149	0.9999
English adverbs (Višnjakova 1976) (3)	-0.6651	-1.6055	204.0278	0.9987
English verbs (Višnjakova 1976) (3)	0.0627	-0.5414	746.3902	0.9937
Russian verbs SO (Krylov, Jakubovskaja 1977, Krylov 1982) (4)	-0.8109	-1.1528	7041.4818	1.0000
Russian nouns SO (Krylov, Jakubovskaja	-1.2201	-1.1964	12521.3205	1.0000

1977, Krylov 1982) (4)				
Polish adjectives (Hammerl 1991) (4)	-2.3388	-0.5356	7206.9255	1.0000
Polish adverbs (Hammerl 1991) (4)	-1.9981	-0.7359	1112.9775	1.0000
Polish nouns (Hammerl 1991) (4)	-0.8879	-1.0597	8800.3470	1.0000
Polish verbs (Hammerl 1991) (4)	-0.2091	-1.0624	3130.4044	0.9999
Polish total (Hammerl 1991) (3)	-1.2500	-0.8025	20353.6346	1.0000
Hungarian all (Papp 1976) (3)	-0.2164	-1.0185	30220.1644	1.0000
Maori words (Wimmer, Altmann 1999) (3)	-1.1965	-0.7329	5109.3108	0.9999

Some data can be better captured by the exponential function in whose differential equation there is no interaction of forces [g(x)/h(x)] but merely a proportionality constant. Setting $dy/y = -bdx$ we obtain

$$(5) \qquad y = a\exp(-bx)$$

or with omitting the zeroes and values smaller than 1

$$(6) \qquad y = 1 + a\exp(-bx).$$

Some of the data could be better captured by (5) or (6). They are presented in Table 1b. This may depend on the source, on severe self-regulation, on the way of preparing the data, on definitions, etc. As a matter of fact, it is preliminarily impossible to give reasons for the choice of the simple or the logarithmic influence.

Table 1b
Exponential fitting

Source	**a**	**b**	**R^2**
Russian 15-17th (Andreevskaja 1990) (6)	12063.6572	-1.6494	0.9994
English nouns (Višnjakova 1976) (6)	21111.8336	-0.8962	0.9961
Japanese postpositions T1 (Sanada, Altmann 2009) (5)	3361.1378	-0.2181	0.9852
Japanese postpositions T2 (Sanada, Altmann 2009) (6)	5354.7916	-0.1670	0.9785
Japanese postpositions T3 (Sanada, Altmann 2009) (6)	4642.9967	-0.1890	0.9592
English all (HO) (Polikarpov 1987) (6)	259405.8320	-1.9698	0.9984
English all (SHO) (Polikarpov 1987) (6)	116312.1870	-0.9364	0.9972

In Russian the complete dictionary has been analyzed by Andreevskaja for individual centuries. Only the data from 15-17th century displayed simple exponential relation. SO: „Slovar Ožegova“; MAS „Slovar' russkogo jazyka" ed. by A.P. Evgen'eva.

In Japanese 3 different texts were used.

English: HO: „Hornby: Oxford Advanced Learner's Dictionary of Current English“; SHO: „Shorter Oxford English Dictionary“.

As can be seen, both texts and dictionaries can be analyzed. In dictionaries one relies on the completeness of the source, in texts one takes only the locally given meanings and

sums them. But one can use a special class of units to distinguish text sorts or to study the development of a writer.

Individual words and morphemes

The second aspect is the polysemy analysis of an individual unit, e.g. a special word or morpheme. It is well known that synsemantics have many meanings while autosemantics are semantically rather restricted. If one considers the occurrence of individual meanings of a unit in texts, one obtains the usual rank-frequency distribution. This can be characterized in many different ways both synchronically and historically, in individual texts and in corpus, in individual writers, text-sorts and languages. Here we have to do with the auxiliary variable represented by ranks which allows us to observe some regularity. The results of the analysis are displayed in Table 2.

In Polish, the data have been taken from two dictionaries: SPP = Słownik poprawnej polszczyny ed. by W. Doroszewski, H. Kurkowska. Warszawa 1973, and MS = Mały słownik języka polskiego ed. by S. Skorupka, H. Auderska, Z. Łempicka. Warszawa 1968.

Table 2
Rank-frequency dependence in the polysemy of individual units

Source	**a**	**b**	**c**	**R^2**
German “von” as particle (Best 1991) (4)	-0.5042	-0.1129	52.4838	0.9794
German “von” as preposition with adverbial meaning (Best 1991) (3)	-0.2850	-0.2449	9.7582	0.9542
German “von” as preposition in prepositional object (Best 1991) (3)	-0.6487	-0.2694	37.8958	0.9866
German “von” as preposition in attributive phrase (Best 1991) (3)	0.0560	-0.3324	21.1295	0.9836
German “auf” in Text 1 (Fuchs 1991) (4)	-0.6643	-0.2037	22.6680	0.9756
German „auf“ as preposition in attribute Text 1 (Fuchs 1991) (3)	-0.5787	-0.2214	23.6084	0.9644
German “auf” as preposition in Attribute Text 2 (Fuchs 1991) (3)	-0.9508	-0.2561	29.8151	0.9834
German “auf” in fixed phrases T2 (Fuchs 1991) (4)	0.5507	-0.8377	24.9964	0.9824
Polish „w“ SPP (Hammerl, Sambor 1991) (3)	-0.8759	-0.2234	199.3127	0.9967
Polish “w” MS (Hammerl, Sambor 1991) (3)	-2.3687	0.3164	298.1104	0.9913
English “in” (Hennern 1991) (4)	-0.1864	-0.4540	52.1876	0.9304
Slovak “do-“ (Nemcová 1991) (3)	1.4539	-2.3136	22.0245	0.9943
Slovak “na-“ (Nemcová 1991) (3)	-0.1080	-0.3380	29.8423	0.9738
Slovak “o-“ (Nemcová 1991) (3)	0.4163	-0.6163	16.1645	0.8528
Slovak “od-“ (Nemcová 1991) (3)	-2.0774	0.5776	25.0148	0.9996
Slovak “po-“ (Nemcová 1991) (3)	0.2424	-0.6729	59.5296	0.9651
Slovak “pre-“ (Nemcová 1991) (3)	-0.9338	-0.0503	32.9225	0.9848
Slovak “roz-“ (Nemcová 1991) (3)	0.1549	-0.1905	25.9347	0.9502

Slovak "s/z-" (Nemcová 1991) (3)	-1.2505	-0.0358	70.8579	0.9937
Slovak "u-" (Nemcová 1991) (3)	0.3750	-1.0889	55.4833	0.8596
Slovak "vy-" (Nemcová 1991) (3)	-0.2845	-0.1828	61.1131	0.9781
Slovak "za-" (Nemcová 1991) (3)	-0.7437	-0.5129	76.9220	0.9951
Japanese "ni" (Roos 1991) (3)	0.3600	-0.5806	38.4638	0.9565
Hungarian "föl-" (Beöthy, Altmann 1984) (3)	-0.3716	-0.0950	10.8001	0.9602
Hungarian "el-" (Beöthy, Altmann 1984) (4)	-3.4680	0.1526	82.0003	0.9999
Hungarian "be-" (Beöthy, Altmann 1984) (3)	-0.4796	-0.2510	19.7244	0.9822
Hungarian "ki-" (Beöthy, Altmann 1984a) (3)	0.0749	-0.3637	11.8491	0.9760
Hungarian "meg-" (Beöthy, Altmann 1991) (3)	-4.0127	1.0718	106.9028	0.9956

Table 2b
Exponential y = a*exp(b*x)

Source	**a**	**b**	**R^2**
German „auf" in Text 2 (Fuchs 1991) (6)	497.9016	-0.5134	0.9493
French „et" (Le Petit Prince) (Rothe 1986) (6)	18.4498	-0.1515	0.9455

As can be seen, one obtains good fittings in all cases. Now, we may conjecture that self-regulation which is present in the model of the Zipf-Alekseev formula leads to some dependence between the parameters *a* and *b*. There are two possibilities of finding it: (1) One unites all data or (2) one considers languages/units separately. Though we conjecture that the greater is *a* representing the norms in the language, the smaller should be *b*, hence $b = f(a)$ must be a decreasing function, either linear or non-linear. Unfortunately, we do not have enough cases for all aspects, for example, not all parts of speech, not enough texts for a certain unit (e.g. a preposition), etc., hence we may merely observe that Slovak or Hungarian prefixes, Polish parts of speech, etc. have the above mentioned decreasing tendency. Uniting all results, we obtain a decreasing tendency but the oscillation is too great. Many further systematic analyses of individual phenomena must be performed to find the $b = f(a)$.

In cases where there are more *homogeneous* data, one can easily state that $b = f(a)$ is a function of the type $y = c + d*exp(x)$ but this can be shown only if there will be sufficient amount of data.

The problem whether other diversification phenomena can be modeled in the above mentioned way remains a task for the future.

References

Altmann, G. (1985). Semantische Diversifikation. *Folia Linguistica 19, 177-200.*

Altmann, G., Best, K.-H., Kind, B. (1987). Eine Verallgemeinerung des Gesetzes der semantischen Diversifikation. In: Fickermann, I. (ed.), *Glottometrika 8, 130-139.* Bochum: Brockmeyer.

Altmann, G. (1991). Modelling diversification phenomena in language. In: Rothe, U. (ed.), *Diversification processes in language: Grammar: 33-46.* Hagen: Rottmann.

Andreevskaja, A.V. (1990). Kvantitativnoe issledovanie polisemii kornevych slov russkogo jazyka XI-XX vekov. In: *Kvantitativnaja lingvistika i avtomatičeskij analiz tekstov 6,* 3-11. [= Učenye zapiski tartuskogo gosudarstvennogo univer siteta, 912]

Beöthy, E., Altmann, G. (1984). Semantic diversification of Hungarian prefixes "föl", "el-", "be-". In: Rothe, U. (ed.), *Glottometrika 7, 45-56.* Bochum: Brockmeyer.

Beöthy, E., Altmann, G. (1984a). The diversification of meaning of Hungarian verbal prefixes. II. ki-. *Finnisch-Ugrische Mitteilungen 8, 29-37*

Beöthy, E., Altmann, G. (1991). The diversification of meaning of Hungarian verbal prefixes 1. "meg-". In: Rothe, U. (ed.), *Diversification processes in language: Grammar: 60-66.* Hagen: Rottmann.

Best, K.-H. (1991). „*Von":* Zur Diversifikation einer Partikel des Deutschen. In: Rothe, U. (ed.), *Diversification processes in language: Grammar: 94-104.* Hagen: Rottmann.

Fuchs, A. (1991). Diversifikation der Präposition *auf*. In: Rothe, U. (ed.), *Diversification processes in language: Grammar: 105-115.* Hagen: Rottmann.

Hammerl, R. (1991). *Untersuchungen zur Struktur der Lexik: Aufbau eines lexikalischen Basismodells.* Trier: Wissenschaftlicher Verlag Trier.

Hammerl, R., Sambor, J. (1991). Untersuchungen zum Beöthy-Gesetz im Polnischen. In: Rothe, U. (ed.), *Diversification processes in language: Grammar: 127-137.* Hagen: Rottmann.

Hennern, A. (1991). Zur semantischen Diversifikation von *in* im Englischen. In: Rothe, U. (ed.), *Diversification processes in language: Grammar: 116-126.* Hagen: Rottmann.

Krylov, Ju.K. (1982). Ob odnoj paradigme lingvističeskich raspredelenij. In: *Trudy po lingvostatistike 8: Lingvostatistika i vyčisliteč'naja lingvistika,* 80-102. [= Učenye zapiski Tartuskogo Gosudarstvennogo Universiteta, 628]

Krylov, Ju.K., Jakubovskaja, M.D. (1977). Statističeskij analiz polisemii kak jazykovoj universalii i problema semantičeskogo toždestva slova. *Naučno-techničeskaja informacija, Serija 2,* 3, 1-6.

Levickij, V.V., Drebet, V.V., Kiiko, S.V. (1999). Some Quantitative Characteristics of Polysemy of Verbs, Nouns and Adjectives in the German Language. *Journal of Quantitative Linguistics 6(2), 172–187.*

Levickij, V.V., Kiiko, J.J., Spolnicka, S.V. (1996). Quantitative analysis of verb polysemy in modern German. *Journal of Quantitative Linguistics 3(2), 132- 135.*

Mačutek, J., Altmann, G. (2007). Discrete and continuous modelling in quantitative linguistics. *Journal of Quantitative Linguistics 14(1), 81-94.*

Nemcová, E. (1991). Semantic diversification of Slovak verbal prefixes. In: Rothe, U. (ed.), *Diversification processes in language: Grammar: 67-74.* Hagen: Rottmann.

Papp, F. (1967). O nektorych količestvennych charakteristikach slovarnogo sostava jazyka, *Slavica 7, 51-58.*

Polikarpov, A.A. (1987). Polisemija: sistemno-kvantitativnye aspekty. *Kvantita tivnaja lingvistika i avtomatičeskij analiz tekstov 3, 135-154.* [= Učenye zapiski tartuskogo gosudarstvennogo universiteta, 774]

Popescu, I.-I., Altmann, G. (2008). On the regularity of diversification in language. *Glottometrics 17, 94-108.*

Popescu, I.-I., Best, K.-H., Altmann, G. (2014). *Unified modeling of length in language.* Lüdenscheid: RAM-Verlag.

Rothe, U. (1986). *Die Semantik des textuellen et.* Frankfurt: Lang.

Rothe, U. (ed.) (1991). *Diversification processes in language: Grammar*. Hagen: Rottmann.

Sanada, H., Altmann, G. (2009). Diversification of postpositions in Japanese. *Glottometrics 19, 70-79.*

Schierholz, S. (1991). *Abstraktheit, Häufigkeit und Polysemie deutscher Substantive*. Tübingen: Niemeyer.

Višnjakova, S.M. (1976). Opyt statističeskogo issledovanija mnogoznačnosti slov v anglijskom jazyke. In: Guseva, E.K., Andrjuščenko, V.M., Revzin, I.I. (eds.), *Vyčislitel'naja lingvistika: 168-178.* Moskva: Nauka.

Wimmer, G., Altmann, G. (1999). Rozdelenie polysémie v maorijčine. In: Genzor, J., Ondrejovič, S. (eds.), *Pange lingua. Zborník na počesť Viktora Krupu: 17-24.* Bratislava: Veda.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R.G. (eds.), *Quantitative Linguistics. An International Handbook: 791-807.* Berlin: de Gruyter.