The meaning distributions on different levels of granularity

Tsy Yih[1], Haitao Liu[1*]

Tsy Yih is the transliteration of the name of the first author in his mother tongue, Shanghai Wu Chinese. He is also known as ZI YE in Mandarin pinyin.

[1] Zhejiang University

[*] Corresponding author’s email: lhtzju@yeah.net

DOI: https://doi.org/10.53482/2023_54_405

ABSTRACT

The meaning distributions of certain linguistic forms generally follow a Zipfian distribution. However, since the meanings can be observed and classified on different levels of granularity, it is thus interesting to ask whether their distributions on different levels can be fitted by the same model and whether the parameters are the same. In this study, we investigate three quasi-prepositions in Shanghainese, a dialect of Wu Chinese, and test whether the meaning distributions on two levels of granularity can be fitted by the same model and whether the parameters are close. The results first show that the three models proposed by modern quantitative linguists can both achieve a good fit for all cases, while both the exponential (EXP) model and the right-truncated negative binomial (RTBN) models behave better than the modified right-truncated Zipf-Alekseev distribution (MRTZA), in terms of the consistency of the goodness of fit, parameter change, rationality, and simplicity. Second, the parameters of the distributions on the two levels and the curves are not exactly the same or even close to each other. This has supported a weak view of the concept of ‘scaling’ in complex sciences. Finally, differences are found to lie between the distributions on the two levels. The fine-grained meaning distributions are more right-skewed and more non-linear. This is attributed to the openness of the categories of systems. The finer semantic differentiation behaves like systems with open set of categories, while the coarse-grained meaning distribution resembles those having a close set of few categories.

Keywords: semantic diversification, meaning distribution, scaling, self-similarity, level of granularity, the exponential distribution, right-truncated negative binomial

1 Introduction

Semantic diversification, or meaning distribution, is a phenomenon characterizing the differentiation of the meanings of words or other linguistic units[1] (Altmann 1985, 2005, 2018; Köhler, 1991; Wang et al. 2021). If all the meanings of a word in texts are arranged in descending order in terms of their frequencies, meaning distributions generally follow a power-law-like curve (Altmann, 2018)[2]. This law is specifically called Beöthy's Law in memory of Beöthy and Altmann's three classical studies on the Hungarian prefixes (1984a, b; 1991). In quantitative linguistics, there has already been a large number of studies in this respect[3].

However, meanings can be observed or classified on different levels of granularity. It is then natural to ask, whether the meaning distributions on different levels abide by the same law. Yet to our knowledge, no previous research in quantitative linguistics has been investigated in this way. In addition, if meaning distributions can be modelled by the same function, whether the parameters are the same. If not, whether they can be used to differentiate between the distributions on two levels.

Among all linguistic meanings or functions expressed by human language, semantic roles[4] have constituted a suitable topic of linguistic study since the idea to differentiate among different levels of granularity appeared. The most recent one we have found is the theory of three-level roles put forward by Van Valin (2005). To conclude the ideas, the traditionally perceived semantic roles such as agent, patient, instrument, recipient, beneficiary, etc., called meso-roles,

[1] Strictly speaking, the term 'semantic diversification' denotes a dynamic process prima facie, while in practice, it is generally used to describe the equilibrium state in that process. Therefore, it is synonymous with 'meaning distribution' or 'meaning diversification' (Fan & Altmann 2008, Fan et al. 2008) in some contexts. They will be used interchangeably in the present study.

[2] Two variable notations, N and V, respectively representing token size and type size are generally used in word frequency studies. In meaning distribution studies, N is kept while M, the counterpart of V, is used standing for the number of meanings of a linguistic form in texts, which will be employed in the remainder of this paper.

[3] For the collection of numerous case studies on semantic diversification, see Strauss & Altmann (2006) and Altmann (2018, Ch. 5).

[4] Semantics roles are alternatively known as theta roles, thematic roles, or participant roles in different traditions.

can be seen as the clustering of verb-specific or event-specific micro-roles. For instance, HITTER (the one who hits) and HITTEE (the one being hit) are micro-roles in the HIT event[5]. Following this model, Hartmann et al. (2014) have illustrated a semantic space where traditional meso-roles, such as agents and patients, can be seen as the clustering of micro-roles when zooming in from a coarse-grained level to a fine-grained level.

We, therefore, regard semantic roles as a suitable lens for semantic diversification on different levels. Semantic roles are generally encoded by case markers and adpositions[6] formally, on which there have already been abundant quantitative linguistic studies (Fuchs, 1991; Hennern, 1991; Roos, 1991; Rothe, 1991; Sanada & Altmann, 2009; Liu 2012; Kolenčíková & Altmann; 2020) due to their multifunctionality (Croft, 2003; Haspelmath, 2003). Thus, it is appropriate to proceed with the research in this line.

Specifically, in this paper, we intend to answer the following research questions:

1. Can meaning distributions on two different levels of granularity be fitted by the same model? Which model is the best?

2. Do the distributions on the two levels have the same parameters or look similar graphically?

3. What are the major differences between the distributions on the two levels and if there exist any potentially affected factors?

The present paper is structured as follows: Section 2 introduces the corpus and procedure. Section 3 shows the results, based on which we will answer the proposed questions in Section 4. Section 5 concludes the whole paper and points out some limitations.

[5] Note that Van Valin also propounds a third and most coarse-grained level of semantic roles, the macro-roles. There are only two macro-roles, actor and undergoer, similar to Dowty’s (1991) proto-agent and proto-patient, which serve as the poles lying on two ends of the continuum of actness. Yet binary classification hardly makes sense for a distribution. Therefore, in this research, macro-roles are not annotated.

[6] The adposition is a cover term for prepositions and postpositions. In languages like Chinese and English, adpositions are predominantly preposed, while postpositions are found in Japanese, Korean and the like. In the remainder of this study, prepositions will simply be used since the main language under investigation is a Wu Chinese, a Sinitic language.

2 Materials and methods

2.1 Materials

The target language under investigation is Wu Chinese, a language of the Sinitic family spoken in Eastern China. Geographically, Wu is distributed in the municipality of Shanghai and in parts of the provinces of Zhejiang, Jiangsu, and Jiangxi. To be more specific, Shanghainese or Shanghai dialect, which is a representative dialect of Wu, was selected. It is the dialect spoken in downtown Shanghai and is the mother tongue of the first author. Since Shanghainese is a dialect officially, it is rarely written down in spite that theoretically, the language experts claim that it can be written in Chinese characters. In recent years, there are folk groups who aim to revive the writing of this language and there have been attempts published in some newspapers such as 新民晚报 *Xinmin Wanbao* "Xinmin Evening News", Wechat pushes, and even Wikipedia entries. Thus, thanks to these resources, we took a corpus-based approach in the present study. Being a dialect also means that Shanghainese lacks an official, authoritative dictionary. There are, nevertheless, two dictionaries of Shanghainese written by scholars, which are 上海话大词典 *Shànghǎihuà Dà-Cídiǎn* 'The Grand Dictionary of Shanghainese' (*SDC*) and 上海方言词典 *Shànghǎi Fāngyán Cídiǎn* 'Shanghai Dialect Dictionary' (*SFC*).

Specifically, we investigated three quasi-prepositions in this study. They are called 'quasi-prepositions' due to the characteristics of the Sinitic languages, where prepositions are generally grammaticalized from verbs or can grammaticalize into conjunctions. Therefore, there are many linguistic forms that stay at the middle stage and possess both the functions of prepositions and verbs or conjunctions. This is reflected in the terms, 'coverbs' and 'prepositional conjunctions' in some classic reference grammars of Chinese (Chao, 1968: 335, 791). The reason we do not exclude the verbal/conjunctional meaning is that there are obvious connections between different uses and these are not cases of homonymy. In other words, from a semasiological perspective, all meanings of the same word forms should be taken into consideration. Yet only the prepositional meanings, or meso-roles, will break down into micro-roles according to our definition of two levels of granularity.

Note that since quasi-prepositions are generally overlooked and less delved into in the traditions of Chinese dialectology, the abovementioned two dictionaries are both sketchy in this respect. A pilot study showed that the forced choice method according to the dictionaries gives poor results. In addition, a compiler of *SFC*, 陶寰 Tao Huan, told us that he deliberately omitted the meanings which are in common with the usage in Mandarin due to the limit of space since it is written in Mandarin and targeted at normal Chinese readers equipped with full lexical competence of Mandarin (p. c.). On the one hand, such background has left us a good chance to have a detailed look at the functions of its prepositions in this language. On the other hand, it calls for manual semantic annotation, which would be somehow subjective. However, due to the fact that on the micro-level, all the micro-roles were verb-specific in the framework adopted by us. We could rely on the verb forms to help discern the prepositional meanings, thereby reducing the degree of subjectivity. As for the meso-role level, we slightly modified the set of well accepted traditional roles according to each case as would be shown below.

The corpus employed was Shanghai Spoken Corpus (SSC) v2.0, compiled by University of Alberta (Han et al., 2017)[7]. In this corpus, all the data were transcribed in Chinese characters. We also transcribed them in Wuyu Pinyin 吴语拼音[8] for the sake of illustration in the remainder of the paper. The whole corpus consisted of six sub-corpora based on genre (conversation, interview, monologue, opera, TV script, song). While it was designed to be a balanced corpus, it was obviously biased towards spoken language. In addition, since in the genres of opera and song, texts usually did not conform to the grammatical pattern of everyday language, they were excluded from the study. For the rest four sub-corpora, the basic information is presented in Table 1.

Table 1: Sizes of sub-corpora in SSC v2.0.

Genre	Number of files	Number of words
conversation	2	28709
interview	5	31251
monologue	21	47663
TV script	23	20942
Sum	51	128565

[7] We appreciate Weifeng Han's help for providing the corpus.

[8] It is a kind of romanization of Wu language proposed by Wu Chinese Society (http://www.wu-chinese.com/).

The corpus querying software used in this study was Wordless v1.3.0 (Ye, 2019). We chose it over common software tools such as WordSmith and AntConc in that the user could choose the sentence rather than a small text within certain spans in all directions around the node word as context. However, truncated sentences were insufficient and confusing in semantics. Thus, a complete context was necessary for semantic annotation with the consideration of our research purpose. After the sentences containing the node words were extracted, they were imported to Microsoft Excel, where we did annotations and basic statistics.

After a simple pilot survey, we selected three representative quasi-prepositions in Shanghainese, 拿 *nau* (and its phonological variant *ne*), 把 *peh* (and its bisyllabic variant *pehla*), 搭 *tah* (and its phonological variants *teh*, *theh*). The basic statistics of the three quasi-prepositions are shown in Table 2. Those hits which were repetitive and unclear were eliminated.

Table 2: Frequencies of all three queries.

Form	Hits in the corpus	Effective hits
拿 nau	357	337
把 peh	386	377
搭 tah	71	63

In our study, we designed two sets of semantic annotations on the basis of dictionaries and the assumed theory of micro-roles. Meanings on two levels of granularity were then annotated manually assuming monosemy. In terms of coarse-grained meanings, we referred to the dictionaries and traditional meso-roles with modifications. As for the fine-grained semantics of prepositions, since micro-roles are verb-specific, the forms of verbs they collocate with are tangible and concrete criteria. Aspectual markers including but not limited to 过 *ku*, 脱 *theh*, 着 *zeh*, 辣海 *lahhe*, 好 *hau*, and directive complements such as 过去 *kuchi*, 进去 *cinchi* were omitted. The complete framework of meaning differentiations is presented in Table 3[9]:

[9] Bold represents the argument introduced by the preposition. ** indicates that the meaning is recorded in both *SDC* and *SFC*, while * in just *SDC*. For *peh*, the dictionary does not distinguish between the verbal usage 'give' and the prepositional usage of dative considering their translational counterparts in Mandarin share identical forms and close relationships on the grammaticalization path. Here we nevertheless make a distinction on both levels.

Table 3: Meaning differentiations of three quasi-prepositions in Shanghainese.

Words	**Part-of-speech**	**Coarse-grained level**	**Fine-grained level**
拿 nau	verb	‘take’**	‘take’
		‘hold’**	‘hold’
		desiderative	‘Give me/I want X.’
		‘use’**	‘use’
	preposition	instrument**	‘do sth. **with X**’
		patient	‘relieve **X**’
			‘process **X**’
			......
		theme	‘tell **X** to Y’
			‘conceive **X** as Y’
			……
		‘taking’	‘taking **X** as an example, VP’
把 peh	verb	permissive**	‘allow’
		causative*	‘cause’
		‘give’**	‘give’
	preposition	recipient**	‘give X **to Y**’
			‘tell X **to Y**’
			……
		patient	‘put **X** Y’
			……
		beneficiary	‘sing X **to Y**’
			‘buy X **for Y**’
			……
		agent (passive)**	‘V-ed **by X**’
		‘according to’	‘according to X, VP’
搭 tah	conjunction	NP conjunction	‘X **and** Y’
	preposition	companion	‘with X’
		recipient	‘tell **X** Y’
			……
		beneficiary	‘do X **for Y’**
			‘build X **for Y’**
			……
		‘same’	‘be the same **as X**’
			‘be different **from X**’
			…
		‘relation’	‘get along **with X**’
			……
		patient	‘meet **X**’
			……
		comparative	‘compared with X’

2.2 Methods

To address the research questions proposed above, we fit five models to the meaning distributions of each quasi-preposition on both coarse-grained and fine-grained levels.

The Zipfian or right-truncated zeta function in (1) is the most common candidate in the literature on rank-frequency distributions. Mandelbrot's formula or the Zipf-Mandelbrot distribution as in (2) introduced a displacement parameter (Mandelbrot 1965). These two fitting models have the advantage of simplicity and are widely used in other scientific disciplines.

(1) $P_x = Cx^{-a} \qquad x = 1, 2, \dots, n$

(2) $P_x = C(x + b)^{-a} \qquad x = 1, 2, \dots, n$

where C denotes an adjusting factor which helps make the sum of whole probabilities one.

Apart from these two, modern quantitative linguists have proposed several models to characterize semantic diversification. Among them, three rival models stand out. First, Altmann (1985) introduced the negative binomial distribution derived from a birth-and-death process. A second model is the Zipf-Alekseev distribution[10] (Hřebíček 1996). In practice, two variants called the right-truncated negative binomial distribution (RTNB) and the modified right-truncated Zipf-Alekseev distribution (MRTZA) are often used. The formula of RTNB and MRTZA are given respectively in (3) and (4):

(3) $P_x = \binom{k + x - 2}{x - 1} p^k (1 - p)^{x-1} \quad x = 1, 2, \dots, n$

where $k > 0$, $0 < p < 1$.

(4) $$P_x = \begin{cases} \alpha & x = 1 \\ \dfrac{(1-\alpha)x^{-(a+b\ln x)}}{T} & x = 2, 3, \dots, n \end{cases}$$

where $T = \sum_{j=2}^{n} j^{-(a+b\ln j)}, a, b \in R, 0 < \alpha < 1.$

Another candidate is the exponential model, also called the stratificational approach, shown in (5), whose assumption is that the relative rate of change of ranked frequencies is constant. This

[10] It is also known as the Zipf-Dolinskij distribution.

distribution has been employed in Fan & Altmann (2008), Popescu et al. (2010), Altmann (2018) and a number of other studies.

(5) $y = 1 + ae^{-bx} \quad x = 1, 2, \ldots, n$

where a, b are parameters, and b stands for the rate of change.

In terms of the nature of models, the first four models are based on probability distributions, which is generally the case, while the last one is indeed a function[11]. The difference between the two cases lies in whether dependent variables add up to one. Popescu et al. (2010) attributed the peculiarity of the last model to the lack of fitting software of exponential distributions. For the rationale or motivation behind each model, interested readers are further referred to the original literature, or to several pieces of work in the handbooks or encyclopedias, such as Altmann (2005), Wimmer & Altmann (2005) and Strauss & Altmann (2006).

The fitting tool used included Altmann Fitter v3.1.0 (Altmann, 2000), which has been frequently employed in quantitative linguistics, and NLREG, which is employed to fit the exponential function. In the next section, we first compare the fitting results of five models and then discuss the research questions in turn.

3 Results

In this section, we present the results of model fitting and the graphical representations of distributions on the two levels. The original data of the observed frequencies can be found in the appendices of this paper.

[11] We thank anonymous reviewers for pointing this out.

Table 4: Parameters of models.

	EXP	**MRTZA**	**RTNB**	**RTZ**	**ZM**
nau coarse	a = 218.2461 b = 0.4731 R^2 = 0.8686	a = 0.0414 b = 1.0705 α = 0.3650 (n = 8) R^2 = 0.9304	k = 2.7438 p = 0.6856 (n = 8) R^2 = 0.9434	a = 1.2058 (n = 8) R^2 = 0.7301	a = 12.0000 b = 18.9835 (n = 8) R^2 = 0.8454
nau fine	a = 1674.6521 b = 2.6199 R^2 = 0.9825	a = 0.3414 b = 0.0482 α = 0.3650 (n = 126) R^2 = 0.9992	k = 0.2482 p = 0.0087 (n = 126) R^2 = 0.9548	NULL	a = 1.1625 b = 0.5499 (n = 126) R^2 = 0.6669
peh coarse	a = 195.6757 b = 0.3964 R^2 = 0.9538	a = 0.0342 b = 0.8091 α = 0.3263 (n = 8) R^2 = 0.9722	k = 2.6833 p = 0.6397 (n = 8) R^2 = 0.9865	a = 1.0761 (n = 8) R^2 = 0.8074	a = 11.9999 b = 23.1314 (n = 8) R^2 = 0.9294
peh fine	a = 295.5034 b = 0.8993 R^2 = 0.9853	a = 0.7630 b = 0.0446 α = 0.3263 (n = 108) R^2 = 0.9776	k = 0.2525 p = 0.0125 (n = 108) R^2 = 0.9842	a = 1.1425 (n = 108) R^2 = 0.9388	a = 1.2146 b = 0.3618 (n = 108) R^2 = 0.8978
tah coarse	a = 22.8774 b = 0.3146 R^2 = 0.8866	a = 0.5333 b = 0.3506 α = 0.2540 (n = 8) R^2 = 0.9125	k = 2.6936 p = 0.5612 (n = 8) R^2 = 0.9247	a = 0.8091 (n = 8) R^2 = 0.7357	a = 11.9998 b = 42.2678 (n = 8) R^2 = 0.8905
tah fine	a = 32.7177 b = 1.0265 R^2 = 0.9380	a = 0.2219 b = 0.0634 α = 0.2063 (n = 38) R^2 = 0.9813	k = 0.5621 p = 0.0405 (n = 38) R^2 = 0.8765	a = 0.7360 (n = 38) R^2 = 0.8641	a = 0.9193 b = 1.5354 (n = 38) R^2 = 0.7358

Table 4 demonstrates the parameters in the five models we have used. The parameters fall into three groups. The first group is put in parentheses and concerns the boundary conditions, including the maximal value of the domain (*n* in all related cases), and normalization constants (*C*s in RTZ and ZM though not shown in the table[12]). These parameters are case-specific but do not reflect inter-case universals. A second group contains the indicators of goodness-of-fit. In this case, we simply resort to the determination coefficient, R^2 ($R^2 > 0.90$, very good; $R^2 > 0.80$, good; $R^2 > 0.75$, acceptable; $R^2 < 0.75$, unacceptable). What is left constitutes the most important group. The parameters in this group are intrinsic to the models per se. When comparing models, the determination coefficient is a basic criterion.

On the basis of the data, we have the following findings. Assume that we move from a coarse-grained level to a fine-grained one. For MRTZA, in the cases of *nau* and *peh*, *a* increases while *b* decreases when while in the case of *tah*, on the contrary, *a* and *b* both decrease. The results are thus not consistent among the three quasi-prepositions for this model. In the model of RTNB, both *k* and *p* decrease significantly. The same is true for parameters *a* and *b* in the Zipf-Mandelbrot function. As in the exponential model, the parameters *a* and *b* increase significantly in all cases.

We could also note in some cells where the fitting results are bad. For instance, fitting RTZ to the data of *nau* on the fine-grained level fails, which makes the fitting results of this model not comparable for all quasi-prepositions. In addition, in several cases where we fit by means of RTZ and ZM, R^2 is less than 0.75, which indicates unacceptability.

Next, we present the graphical results for comparison between the two levels of meaning granularity.

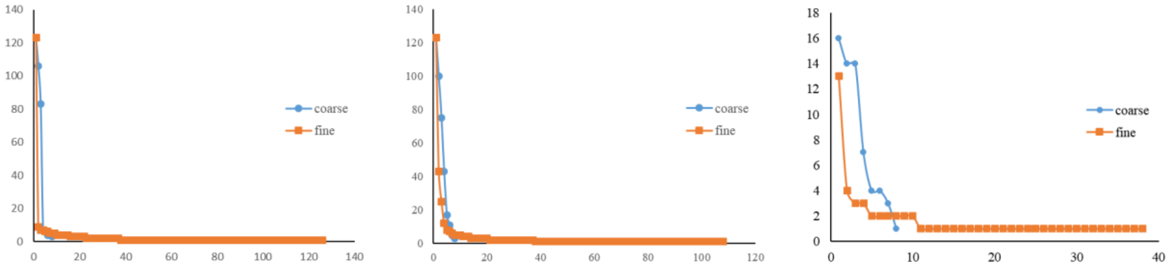

Figure 1. The rank-frequency distributions of meanings on two levels in linear coordinates (left: *nau*; middle: *peh*; right: *tah*).

[12] In fact, as shown in the formula (1–2), the models RTZ and ZM also have such a normalization constant *C*. Yet in the Altmann Fitter, they are regarded as probability distributions rather than functions, such as the exponential model fitted by means of NLREG. Hence, this parameter can be ignored given the additional constraint that the probabilities of all items add up to 1.

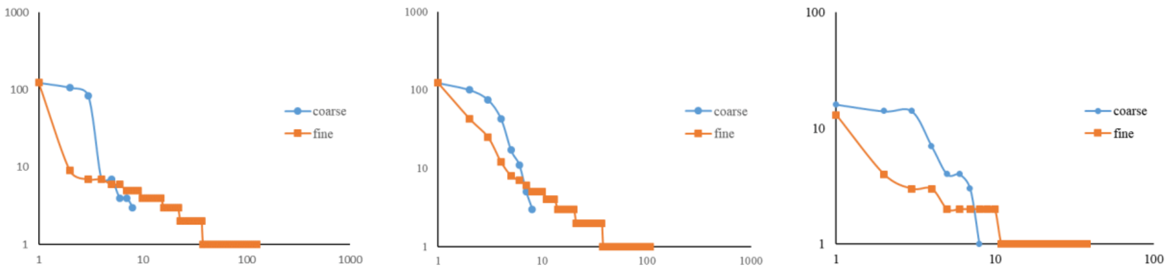

Figure 2. The rank-frequency distributions of meanings on two levels in log-log coordinates (left: *nau*; middle: *peh*; right: *tah*).

Figure 1 shows that the rank-frequency distributions of meanings on fine-grained ones are right-skewed compared with coarse-grained ones. In other words, fine meaning distributions have long tails. At first sight, the shapes of the two distributions are much different. In case there is information hidden by the linear coordinate, we also present them in log-log coordinates (Figure 2). It is shown that in no case are the distributions linear throughout the whole domain, or following a pure power law. Yet there could still be a ‘scaling range’ (Mandelbrot, 1997: 200). On the coarse-grained level, the curves first decrease slowly and then go down straight with a sudden change, while on the fine-grained level, there seem to be two stages. The first stage is linear and the second stage breaks down into steps.

4 Discussion

4.1 Which model is the best?

We have found that the three models of EXP, MRTZA and RTNB all give good results in terms of R^2. The determination coefficients of MRTZA are the largest in most cases. As for the rest two, sometimes the R^2 of the exponential models is larger than that of RTNB while other times the opposite happens. Prima facie, MRTZA is the best choice. However, R^2 is not the only criterion for comparing models. We argue that the exponential function and the RTNB model surpass MRTZA on several other aspects. First, we can see from our results that in terms of the change of parameters, both *a* and *b* in the exponential model, and *k* and *p* in RTNB change in a consistent way between two levels of granularity for each quasi-preposition. Specifically, the first group becomes larger as the semantic granularity goes finer, while the second group decreases. While for the case of MRTZA, the parameters *a* and *b* change in a different way for three quasi-prepositions. Second, the exponential model and RTNB have two intrinsic parameters, while the MRTZA has three. From the perspective of the Occam's Razor Principle, they perform both better than MRTZA. In fact, Altmann (2018: 4) also argued for the exponential function to be the main candidate of a unified model for the diversification phenomena, which is parallel to the status of Zipf-Alekseev function for length distribution. His primary motivation also pertains to simplicity, as the original differential equation and the rationale behind it are simpler than the other models. The exponential function simply follows the assumption that the relative rate of change of ordered frequencies is constant and negative, and the parameter *b* is that constant (Altmann, 2018: 3). On the other hand, before the advent of the exponential model, RTNB has always been among the best models characterizing the meaning diversification phenomena (see Beöthy & Altmann, 1984a, b and a number of papers in Rothe (ed.), 1991). Our findings again support the applicability of the model.

In sum, both the exponential model and RTNB have good rationales for being considered the best fitting models characterizing meaning distributions on both levels of granularity, and there seems to be no reason to argue for a winner between them based on the data provided in this paper. Moreover, the parameters can be used to differentiate between the two levels.

4.2 Parameters, same or different?

In this section, we aim to answer the question of whether the meaning distributions on two levels of granularity are similar. Both the parameters and graphical representations in Section 3 show that the distributions are very different between the two cases. On the one hand, the fitting results indicate that the parameters of the distributions change drastically, whereas on the other, the curves presented in either coordinate do not possess the same or similar shapes.

In what follows, we shall relate our finding to the concept of 'scaling' in complex sciences, i.e., the study of complex systems. 'Scaling' can be roughly understood as that systems observed on different scales manifest similar phenomena or follow the same rules (Kretzschmar, 2009). In several seminal works of Kretzschmar (2009, 2015, 2018, Kretzschmar et al. 2013), for instance, he investigated this issue from the perspective of the sociolinguistics of phonological systems. He found that the frequency distributions of phonemes on different scales in the acoustic space all manifest A-curves graphically[13], though he did not fit certain probability distributions to his data. Therefore, he deemed that he had proven the property of scaling at least in the field of sociophonetics.

Nevertheless, scaling may have two interpretations. The strong version of scaling sees it as the synonymy of 'self-similarity', which holds that meaning distributions on different levels of granularity follow the same fitting model and probably even have the same or similar parameters. This is a standpoint taken in Kretzschmar (2009). On the contrary, a weak or mild version of scaling says that distributions on different levels or scales are not strictly isomorphic. Rather, it is just that they all manifest A-curves in Kretzschmar's term, but do not necessarily have the same distribution functions, or other statistical parameters. This view was proposed in Kretzschmar et al. (2013). Our findings apparently support the weak version of scaling.

We shall next spend some space explaining why the strong version does not hold. In Kretzschmar (2009), he showed a strong favor of the idea that 'the part contains the information of the whole' which is a property of fractals based on his early non-quantitative study. For instance, he quoted the definition of Mandelbrot (1982) in Kretzschmar (p. 198). He also drew on the classical, well-known case of the length of the British coast studied by Mandelbrot (1967) (p. 179). However, a common misconception about the story is that a part of the coastline reflects the shape of the whole. In fact, Mandelbrot has already made it rather clear that it should be understood in a statistical sense. The related property is referred to as 'statistical self-similarity' rather than rigorous self-similarity in the sense of pure maths (as reflected by Koch snowflakes for instance). For real-life objects, a part is not the miniature of the whole generally. In other words, parts do not contain the information of the whole, and one could not deduce the total information about the whole from parts. As for the linguistic cases, it holds as well for

[13] He has named such distributions 'A-curves', mimicking 'S-curves' which are common in the field of language change. However, it seems inappropriate since there is no climbing-up part as in the graph of the letter 'A'. Rather, 'L-curve' appears to be more vivid.

the distributional patterns, and there is no such magic power that guarantees the isomorphism. Kretzschmar's prior understanding of 'scaling' falls into the Individualistic fallacy, the reverse of the Ecological fallacy, which is a classic statistical fallacy in science as pointed out by Horvath & Horvath (2003).

Later in Kretzschmar et al. (2013), there seems to be a change of idea. Kretzschmar has come to a milder conclusion with regard to the scaling property. That is, distributions on different scales are not strictly isomorphic. Rather, it is just that they all manifest in A-curves, but do not have the same distribution function, or the same statistical indicators. He also explicitly cited Horvaths' work and publicly support their standpoint. However, his attitude was still vacillating as reflected in his later monographs (Kretzschmar, 2015, 2018) which might be rather confusing to the reader. Therefore, it seems that Kretzschmar is not that certain about the interpretation of scaling, which is thus worth testing with real data. Based on our research, we agree with this moderate view of scaling, although this weak version itself seems to be a less significant claim than the strong version. Yet in other words, it also means that the parameters of models do have the ability to differentiate between levels of meaning granularity.

In the next section, we proceed to discuss the differences between the distributions and the primary factors.

4.3 Differences between the distributions on the two levels

Kretzschmar et al. (2013) claimed that a distribution with larger set of types tends to be more non-linear, and vice versa. This is supported by our results as shown in Figure 1. Therefore, we supplemented their conclusion that the fine-grained meaning distributions are more right-skewed.

In fact, this phenomenon can be explicated by the following proof. Remember that for this specific situation, we have constant N (number of tokens) and variant M (number of meaning types). Assume a distribution denoted as $\{f_r(x)\}$, $r = 1, 2, \ldots M$, where $\sum_{r=1}^{M} f_r(x) = N$. Since the total number of tokens N remains the same, once the group annotated as rank m is given a more fine-grained annotation, this class with frequency $f(m)$ will be broken down into several items with lower frequencies, thereby increasing the area of tail[14]. One extreme case is that if a person is able to identify different meanings in any different context, then $M = N$ and the absolute frequency of any item will become 1. Alternatively, if in all contexts is the word recognized as sharing the same meaning, then one meaning item takes all the frequencies.

In addition, we draw a key distinction between the two cases. Overall, the meaning distribution is similar to the case of rank-frequency distribution of various constituents. Yet a fine-grained distribution of meanings resembles that of words, whereas a coarse-grained one is alike that of letters or phonemes. The major difference lies in the openness of set of types. In the case of letters, phonemes and coarse-grained meanings, the set of all types M is closed, whereas for words or fine-grained meanings[15] here, it is an open set and grows with the number of tokens. It has been known in the literature that the distribution of words possesses a longer tail and has more hapaxes than that of letters or phonemes (Best & Rottmann, 2017, ch. 9), as well as being more non-linear. Thus in a similar vein, the same applies to the fine-grained meanings.

In sum, based on the graphical representations, we have pointed out the major difference between the two levels of meaning granularity, and attributed it to the openness of categories of the system.

4.4 Other general issues

In this final subsection, other factors that might influence our results are discussed.

First concern the genre of the corpus. Roos (1991) conducted a survey on the semantic diversification of Japanese *ni* and considered four homogenous texts and a mixed corpus. He found that the heterogeneity of the text does not play a crucial role. This has guaranteed the effectiveness of our research which also adopts a speech-biased corpus with several genres.

Second, we have only discussed the effect of the openness and numbers of the categories or

[14] A key condition here is that the set of fine-grained types must be the strict refinement of that of coarse-grained ones. Otherwise, this proof might not hold.

[15] Based on the approach taken in this study, fine-grained meanings are form-dependent, and thus form an open set. In other approaches, if one sets his own fine-grained level with the help of a dictionary or other sources, it will also be a closed set then. However, in real texts, it is common to find a meaning encoded by a word that is not predefined or recorded in the dictionary, a phenomenon caused by innovation in language use.

types, while Kretzschmar et al. (2013) mentioned that the shape of the distribution is also subject to the number of tokens. That is, a size effect might exist. Specifically, he deemed that only a sample with a large token size will manifest a non-linear distribution. In terms of our meaning distributions here, although the whole corpus is large enough with 130k tokens, the amount of the extracted form-meaning pairs might still be small, which is consequently expected to be expanded in the future.

One last facet concerns the identification of meaning-carrying units and the subjectivity of categorization. In all three cases, there are special items (*nau* in 拿……来讲 *nau ... lekaon* 'taking … as an example', *peh* in 把……讲起来 *peh ... kaonchile* 'according to' and *tah* in 搭……比起来 *tah ... pichile* 'compared with') for which the whole constructions rather than single words seem to be more appropriate meaning carriers. As far as we know, we have found no literature discussing the effect of unit identification on distributions so far. On the other hand, in Kretzschmar et al. (2013)'s study, speech as his scope of the study can be measured with accuracy, whereas in our case, we do not have a real semantic space as our foundation and the meaning annotation is more or less subject to subjectivity. The criterion of counting the number of meanings is inevitably vague (see Guiter, 1974 for a thorough discussion). In some studies, dictionaries were resorted to, which can serve as a golden standard. In most of the others, nevertheless, the methods were not clearly reported. However, even if one applies the dictionary approach, the actual use in texts might not be contained in the dictionaries, which leaves us only two remedies. The first is the forced choice method, which means to choose the closest meaning in the dictionary. The second is to go beyond the dictionary and add new meanings based on annotators' intuitive judgment. It is a probable guess that there have long been such moves in that some researchers apparently annotate the word meanings subjectively. For example, in Rothe's survey of the French word *et*, 72 different meanings are counted, which is usually too large a number of meanings for an entry in a dictionary to contain (Rothe, 1986, reported in Altmann, 2018: 41). Either way taken, this issue should hopefully have a better solution in the future studies.

5 Conclusions

In conclusion, this article attempts to investigate the features of semantic diversification on different levels of granularity. By way of extracting three quasi-prepositions from a corpus of the Shanghai dialect of Wu Chinese and annotating them semantically on two levels of granularity, we have answered three research questions.

First, several models are compared and those proposed by quantitative linguists show better performance than simple power functions. The exponential model and the right-truncated negative binomial model are found to be the best two considering the goodness of fit, consistency of parameter change, rationality, and simplicity. Second, our findings support the weak view of 'scaling' in complex sciences, that is, the meaning distributions on different levels of granularity all manifest the so-called A-curves by Kretzschmar in a rough sense. However, the parameters and shapes of models are different. In other words, the interpretation of scaling as self-similarity in a rigorous mathematical sense does not hold. Finally, there are several differences between the distributions on the two levels. The meaning distributions on a fine-grained level are found to be more right-skewed and more non-linear as compared with those on a coarse-grained one. This can also be proven mathematically given constant N (number of tokens) and variant M (number of types). The primary reason for the difference is attributed to the openness of the categories of systems.

The present study also adds to our understanding of the quantitative aspects of syntax-semantics interface or form-meaning mappings. Since the complex nature of 'multiple-forms-to-multiple-meanings' in natural language is widely acknowledged, in practice linguists start from the perspectives of synonymy ('one-meaning-to-multiple-forms') and polysemy ('one-form-to-multiple-meanings') in traditional terms, or onomasiological and semasiological approaches in usage-based, cognitive linguistic terms (Geeraerts, 2010). Köhler (1991) has made a similar distinction between two kinds of diversification from the perspective of quantitative linguistics. There has been such research into the former (Zhu & Liu, 2018) and we have contributed to the latter. On the macro level, the related distributional phenomena are attributed to the metaphorical language forces (Altmann, 1985; Altmann & Köhler, 1996), while on the micro level, they are reflections of several synergetic principles such as the minimization of efforts during language use or of inventories in language users' mind (Köhler, 2005, 2012).

Without doubt, this study also has its limitations. In the first place, it is still inevitable as we have pointed out that the differentiation of meanings is subjective. We have tried to minimize the degree of subjectivity such as resorting to dictionaries or basing the judgments on more concrete forms. Future studies might call for better measurements of meanings. Second, the size effect of the corpus is not tested in this survey, and we acknowledge that the size of hits may be criticized for being too small (up to a few hundreds). Third, we have only distinguished between two levels of granularity of meanings, while there is still a dearth of accurate measures of the hierarchical nature of meaning. Further investigations of these questions, along with a better distributional model or descriptive tool, are in need.

References

Altmann, G. (1985). Semantische Diversifikation. *Folia Linguistica*, 19(1–2), pp. 177–200.

Altmann, G. (2005). Diversification processes. In: Köhler, R., Altmann, G., Piotrowski, R.G. (Eds.), *Quantitative Linguistics. An International Handbook*. Berlin: de Gruyter, pp. 646–658.

Altmann, G. (2018). *Unified Modeling of Diversification in Language*. Lüdenscheid: RAM-Verlag.

Altmann, G., Köhler, R. (1996). "Language forces" and synergetic modelling of language phenomena. In: Schmidt, P. (Ed.), *Glottometrika 15*. Trier: WVT, pp. 62–76.

Beöthy, E., Altmann, G. (1984a). Semantic diversification of Hungarian verbal prefixes II. ki-. *Finnisch-ugrische Mitteilungen*, *8*, pp. 29–37.

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

Beöthy, E., Altmann, G. (1991). The diversification of meaning of Hungarian verbal prefixes I. "meg-". In: Rothe, U. (Ed.). *Diversification Processes in Language: Grammar*. Hagen: Rottmann, pp. 60–66.

Best, K.-H., Rottmann, O. (2017). *Quantitative Linguistics, an Invitation*. Lüdenscheid: RAM-Verlag.

Chao, Y.-R. (1968). *A Grammar of Spoken Chinese*. Berkeley and Los Angeles: University of California Press.

Cong, J., Liu, H. (2014). Approaching human language with complex networks. *Physics of Life Reviews, 11*, pp. 598–618.

Croft, W. (2003). *Typology and Universals (2nd ed.).* Cambridge: Cambridge University Press.

Dowty, D. (1991). Thematic Proto-Roles and Argument Selection. *Language*, 67(3), pp. 547–619.

Fan, F., Altmann, G. (2008). On meaning diversification in English. *Glottometrics*, 17, pp. 66–78.

Fan, F., Popescu, I.-I., Altmann, G. (2008). Arc length and meaning diversification in English. *Glottometrics*, 17, pp. 79–96.

Fuchs, R. (1991). Diversifikation der Präposition *auf*. In: Rothe, U. (Ed.), *Diversification Process in Language: Grammar*, pp. 105–115. Hague: Rottmann.

Geeraerts, D. (2010). *Theories of Lexical Semantics*. Oxford: Oxford University Press.

Guiter, H. (1974). Les relations/Frequence-longueur-sens/Des mots (Langues Romanes et Anglais). In: Varvaro A. (Ed.). *XIV Congresso Internationale di Linguistica e Filologia Romanza: Napoli, 15–20 Aprile 1974. ATTI*, pp. 373–381. Amsterdam: John Benjamins.

Han, W., Arppe, A., Newman, J. (2017). Topic marking in a Shanghainese corpus: From observation to prediction. *Corpus Linguistics and Linguistic Theory,* 13(2), pp. 291–319. doi:10.1515/cllt-2013-0014

Hartmann, I., Haspelmath, M., Cysouw, M. (2014). Identifying semantic role clusters and alignment types via microrole coexpression tendencies. *Studies in Language*, 38(3), pp. 463–484.

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. In: Tomasello, M. (Ed.). *The New Psychology of Language: Cognitive and functional approaches to language structure* (Vol. 2). Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 211–242.

Hennern, A. (1991). Zur semantischen Diversifikation von *in* im Englischen. In Rothe, U. (Ed.), *Diversification Process in Language: Grammar*. Hague: Rottmann, pp. 116–126.

Horvath, B. M., Horvath, R. J. (2003). A closer look at the constraint hierarchy: Order, contrast, and geographical scale. *Language Variation and Change*, 15, pp. 143–170.

Hřebíček, L. (1996). Word associations and text. *Glottometrika*, 15, pp. 96–101.

Köhler, R. (1991). Diversification of coding methods in grammar. In: Rothe, U. (Ed.). *Diversification Process* *in Language: Grammar*, pp. 47–51. Hague: Rottmann.

Köhler, R. (2005). Synergetic linguistics. In: Köhler, R., Altmann, G., Piotrowski, R. G. (Eds.). *Quantitative Linguistics: An International Handbook*, pp. 760–774. Berlin/New York: de Gruyter.

Köhler, R. (2012). *Quantitative Syntax Analysis*. Berlin: de Gruyter.

Kolenčíková, N., Altmann, G. (2020). Analysis of Prepositions in Marína (Slovak Romantic Poem). *Glottometrics*, 48, pp. 88–107.

Kretzschmar, W. A. Jr. (2009). *The Linguistics of Speech*. Cambridge: Cambridge University Press.

Kretzschmar, W. A. Jr. (2015). *Language and Complex Systems*. Cambridge: Cambridge University Press.

Kretzschmar, W. A. Jr. (2018). *The Emergence and Development of English: An Introduction*. Cambridge: Cambridge University Press.

Kretzschmar, W. A. Jr., Kretzschmar, B. A., Brockman, I. M. (2013). Scaled measurement of geographic and social speech data. *Literary and Linguistic Computing*, 28, pp. 173–187.

Liu, H. (2012). Probability distribution of semantic roles in a Chinese treebank annotated with semantic roles. In: Naumann, S., Grzybek, P., Vulanović, R. Altmann, G. (Eds.). *Synergetic Linguistics. Text and Language as Dynamic Systems*, pp. 101–107. Vienna: Praesens.

Mandelbrot, B. (1965). Information Theory and Psycholinguistics. In: Wolman, B. B., Nagel, E. (Eds.). *Scientific Psychology*. Basic Books.

Mandelbrot, B. (1967). How long is the coast of Britain? Statistical self-similarity and fractional dimension. *Science*, *156*, pp. 636–638.

Mandelbrot, B. (1982). *The Fractal Geometry of Nature*. New York: W. H. Freeman and Company.

Mandelbrot, B. (1997). *Fractals and Scaling in Finance: Discontinuity, concentration, risk*. New York: Springer.

Popescu, I.-I., Altmann, G., Köhler, R. (2010). Zipf's law—another view. *Quality & Quantity*, *44*(4), pp. 713–731.

Roos, U. (1991). Diversifikation der japanischen Postposition "-ni". In: Rothe, U. (Ed.). *Diversification Process in Language: Grammar*, pp. 75–82. Hague: Rottmann.

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

Rothe, U. (1991). The diversification of the case: genitive. In Rothe, U. (Ed.). *Diversification Process in Language: Grammar*, pp. 140–156. Hague: Rottmann.

Rothe, U. (Ed.). (1991). *Diversification Process in Language: Grammar*. Hague: Rottmann.

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

Strauss, U., Altmann, G. (2006). *Diversification. In the Encyclopedia of Linguistic Laws and the Laws in Quantitative Linguistics.* Retrieved from http://lql.uni-trier.de/index.php/Diversification

Van Valin, R. D., Jr. (2005). *Exploring the Syntax-Semantics Interface*. Cambridge: Cambridge University Press. doi:10.1017/CBO9780511610578

Wang, L., Guo, Y., Ren, C. (2021). A Quantitative Study on English Polyfunctional Words. *Glottometrics*, 50, pp. 42–56.

Wimmer, G., Altmann, G. (2005). Unified derivation of some linguistic laws. In: Köhler, R., Altmann, G., Piotrowski, R. (Eds.). *Quantitative Linguistics: An International Handbook*, pp. 791–807. Berlin/New York: Walter de Gruyter.

Zhu, J., Liu, H. (2018). The distribution of synonymous variants in Wenzhounese. *Glottometrics*, *41*, pp. 24–39.

Software

Altmann, G. (2000). *Altmann-Fitter 3.1.0* [Computer software]. Lüdenscheid: RAM-Verlag. Retrieved from http://www.ram-verlag.biz/altmann-fitter/

Ye, L. (2019). Wordless (Version 1.3.0) [Computer software]. Retrieved from https://github.com/BLKSerene/Wordless

Appendix I

The meaning distributions of three quasi-prepositions on the coarse level

Wordform	Meaning	x[i]	F[i]	NP[i][16]				
				EXP	MRTZA	RTNB	RTZ	ZM
nau	‘take’	1	123	137	123	124	145	140
	theme	2	106	86	116	102	63	81
	patient	3	83	54	52	59	39	48
	‘hold’	4	7	34	24	29	27	29
	instrument	5	7	21	12	13	21	18
	‘use’	6	4	14	6	6	17	11
	desiderative	7	4	9	3	2	14	7
	‘taking’	8	3	6	2	1	12	5
peh	‘give’	1	123	133	123	114	147	154
	recipient	2	100	90	113	110	70	88
	passive	3	75	61	62	73	45	53
	permissive	4	43	41	34	41	33	33
	beneficiary	5	17	28	20	21	26	21
	patient	6	11	19	12	10	21	14
	causative	7	5	13	8	5	18	9
	‘according to’	8	3	9	5	2	16	6
tah	recipient	1	16	18	16	14	20	17
	beneficiary	2	14	13	17	16	11	13
	companion	3	14	10	10	13	8	10
	‘relation’	4	7	8	7	9	6	7
	NP conjunction	5	4	6	5	6	5	6
	patient	6	4	4	4	3	5	4
	‘same’	7	3	4	3	2	4	4
	comparative	8	1	3	2	1	4	3

[16] The theoretical values are rounded here, as the frequencies are all integers.

Appendix II

The meaning distributions of *nau* on the fine level

Rank	Frequencies	Rank	Frequencies	Rank	Frequencies
1	123	43	1	85	1
2	9	44	1	86	1
3	7	45	1	87	1
4	7	46	1	88	1
5	6	47	1	89	1
6	6	48	1	90	1
7	5	49	1	91	1
8	5	50	1	92	1
9	5	51	1	93	1
10	4	52	1	94	1
11	4	53	1	95	1
12	4	54	1	96	1
13	4	55	1	97	1
14	4	56	1	98	1
15	4	57	1	99	1
16	3	58	1	100	1
17	3	59	1	101	1
18	3	60	1	102	1
19	3	61	1	103	1
20	3	62	1	104	1
21	3	63	1	105	1
22	3	64	1	106	1
23	2	65	1	107	1
24	2	66	1	108	1
25	2	67	1	109	1
26	2	68	1	110	1
27	2	69	1	111	1
28	2	70	1	112	1
29	2	71	1	113	1
30	2	72	1	114	1
31	2	73	1	115	1
32	2	74	1	116	1
33	2	75	1	117	1
34	2	76	1	118	1
35	2	77	1	119	1
36	2	78	1	120	1
37	2	79	1	121	1
38	1	80	1	122	1
39	1	81	1	123	1
40	1	82	1	124	1
41	1	83	1	125	1
42	1	84	1	126	1

Appendix III

The meaning distributions of *peh* on the fine level

Rank	Frequencies	Rank	Frequencies	Rank	Frequencies
1	123	37	2	73	1
2	43	38	1	74	1
3	25	39	1	75	1
4	12	40	1	76	1
5	8	41	1	77	1
6	7	42	1	78	1
7	6	43	1	79	1
8	5	44	1	80	1
9	5	45	1	81	1
10	5	46	1	82	1
11	4	47	1	83	1
12	4	48	1	84	1
13	4	49	1	85	1
14	3	50	1	86	1
15	3	51	1	87	1
16	3	52	1	88	1
17	3	53	1	89	1
18	3	54	1	90	1
19	3	55	1	91	1
20	3	56	1	92	1
21	2	57	1	93	1
22	2	58	1	94	1
23	2	59	1	95	1
24	2	60	1	96	1
25	2	61	1	97	1
26	2	62	1	98	1
27	2	63	1	99	1
28	2	64	1	100	1
29	2	65	1	101	1
30	2	66	1	102	1
31	2	67	1	103	1
32	2	68	1	104	1
33	2	69	1	105	1
34	2	70	1	106	1
35	2	71	1	107	1
36	2	72	1	108	1

Appendix IV

The meaning distributions of *tah* on the fine level

Rank	Frequencies	Rank	Frequencies	Rank	Frequencies
1	13	14	1	27	1
2	4	15	1	28	1
3	3	16	1	29	1
4	3	17	1	30	1
5	2	18	1	31	1
6	2	19	1	32	1
7	2	20	1	33	1
8	2	21	1	34	1
9	2	22	1	35	1
10	2	23	1	36	1
11	1	24	1	37	1
12	1	25	1	38	1
13	1	26	1		