Zipf's Law and Avoidance of Excessive Synonymy

Dmitrii Y. Manin

Abstract

Zipf's law states that if words of language are ranked in the order of decreasing frequency in texts, the frequency of a word is inversely proportional to its rank. It is very reliably observed in the data, but to date it escaped satisfactory theoretical explanation. This article suggests that Zipf's law may result from a hierarchical organization of word meanings over the semantic space, which in turn is generated by the evolution of word semantics dominated by expansion of meanings and competition of synonyms. A study of the frequency of partial synonyms in Russian provides experimental evidence for the hypothesis that word frequency is determined by semantics.

Keywords: Zipf's law; Semantics; Word frequency; Word meaning; Synonymy

Zipf's law (Zipf, 1949) may be one of the most enigmatic and controversial regularities known in linguistics. It has been alternatively billed as the hallmark of complex systems and dismissed as a mere artifact of data presentation.[1] The simplicity of its formulation, its experimental universality, and its robustness starkly contrast with the obscurity of its meaning. In its most straightforward form, it states that if the words of a language are ranked in order of decreasing frequency in texts, the frequency is inversely proportional to the rank,

$$f_k \propto k^{-1} \tag{1}$$

where f_k is the frequency of the word with rank k. As a typical example, consider a log-log plot of frequency vs. rank in Fig. 1. It is calculated from a frequency dictionary of the Russian language compiled by S. Sharoff (n.d., 2002). The dictionary is based on a corpus of 40 million words, with special care (Sharoff, 2002) taken to prevent data skewing by words with high concentrations in particular texts (like the word *hobbit* in a Tolkien novel).

Zipf's law is usually presented in a generalized form where the power law exponent may be different from -1,

$$f_k \propto k^{-B}. \tag{2}$$

Correspondence should be sent to Dmitrii Manin, 3127 Bryant St., Palo Alto, CA 94306. E-mail: manin@pobox.com
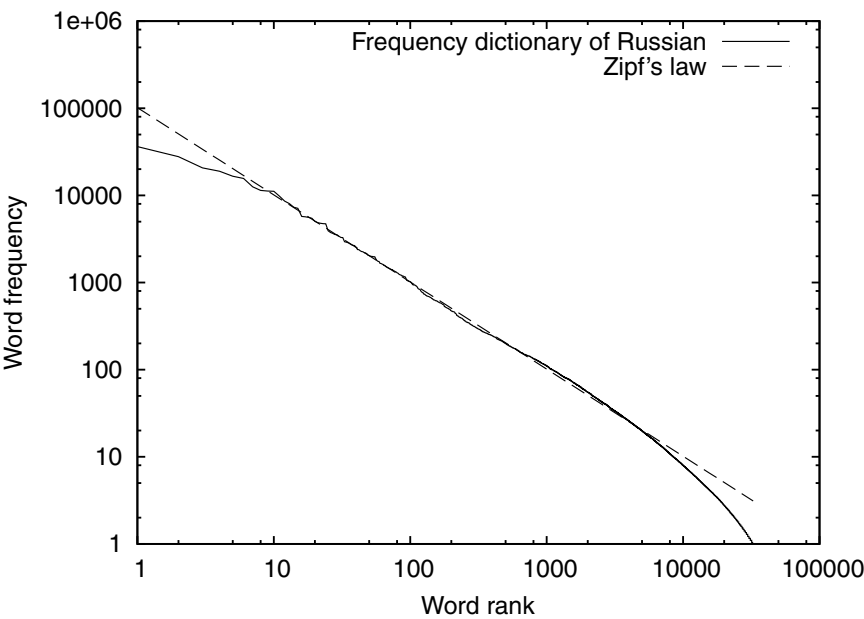
Fig. 1. Zipf's law for the Russian language.

According to Ferrer i Cancho (2005c), where an extensive bibliography is presented, various subsets of the language obey the generalized Zipf's law (Equation 2). Thus, while the value of $B \approx 1$ is typical for single author samples and balanced frequency dictionaries, different values, both greater and less than 1, characterize the speech of schizophrenics and very young children, military communications, or subsamples consisting of nouns only.

Here we concentrate on the "normal" language approximated by balanced corpora and do not consider the above-mentioned special cases and subsets. Neither do we attempt to generalize our treatment to other power laws found in various domains such as computer science or economics (a comprehensive bibliography compiled by Wentian Li is available at http://www.nslij-genetics.org/wli/zipf/). The purpose of this work is to demonstrate that inverse proportionality (Equation 1) can be explained on purely linguistic grounds. Likewise, we do not pay special attention to the systematic deviations from the inverse proportionality at the low-rank and high-rank ends, considering them second-order effects.

It is not possible to review the vast literature related to the Zipf's law. However, it appears that the bulk of it is devoted to experimental results and phenomenological models. There are not very many models that would aim at explaining the underlying cause of the power law and predicting the exponent. We briefly review models of this type in the first section. In section 2, we discuss the role in the language of words/meanings having different degrees of generality. In section 3, we show that Zipf's law can be generated by a particular kind of arrangements of word meanings over the semantic space. In Section 4, we discuss the evolution of word meanings and demonstrate that it can lead to such arrangements. Section 5 is devoted to numerical modeling of this process. Discussion and prospects for further studies constitute

section 6. In the Appendix we present some evidence to support the assumption that a word's frequency is proportional to the extent of its meaning.

1. Some previous models

1.1. Statistical models of Mandelbrot and Simon

The two best-known models for explaining Zipf's law in the linguistic domain are due to two prominent figures in 20th-century science: Benoît Mandelbrot, of fractals fame, and Herbert A. Simon, one of the founding fathers of AI and complex systems theory.[2]

One of the simplest models exhibiting Zipfian distribution is due to Mandelbrot (1966) and is widely known as *random typing* or *intermittent silence* model. It is just a generator of random character sequences where each symbol of an arbitrary alphabet has the same constant probability and one of the symbols is arbitrarily designated as a word-delimiting "space." The reason why "words" in such a sequence have a power-law frequency distribution is very simple, as noted by Li (1992). Indeed, the number of possible words of a given length is exponential in length (since all characters are equiprobable), and the probability of any given word is also exponential in its length. Hence, the dependency of each word's frequency on its frequency rank is asymptotically given by a power law. In fact, the characters needn't even be equiprobable for this result to hold (Li, 1992).

Based on this observation, it is commonly held that Zipf's law is "linguistically shallow" (Mandelbrot, 1982) and does not reveal anything interesting about natural language. However, the number of distinct words of the same length in real language is not exponential in length and is not even monotonic, as can be seen in Fig. 2, where this distribution is calculated from a frequency dictionary of the Russian language (Sharoff, n.d.) and from Leo Tolstoy's novel *War and Peace*.

It follows that the random typing model is not applicable to natural language, and there has to be a different explanation, possibly, a "deeper" one.

Another purely statistical model for Zipf's law applicable in various domains, including language, was proposed by Simon (1955, 1957). It is based on a much earlier work by Yule (1925), who introduced his model in the context of evolutionary biology (distribution of species among genera) as early as 1925. Currently, this and related models are known as *preferential attachment* or *cumulative advantage* models, since they describe processes where the growth rate of an object is proportional to its current size. Namely, define an *n-word* as a word that has occurred exactly *n* times in the preceding text. Assume that a text is generated sequentially word by word, and that the probability for the next word to be one of the *n*-words is equal to the fraction of all *n*-word tokens in the preceding sequence. Simon showed that this process leads to the Zipfian distribution. He also emphasized that it is compatible with the notion that people select words according to the current topic, rather than completely randomly. For more details on both Simon's and Mandelbrot's models see, e.g., Mitzenmacher, 2003.
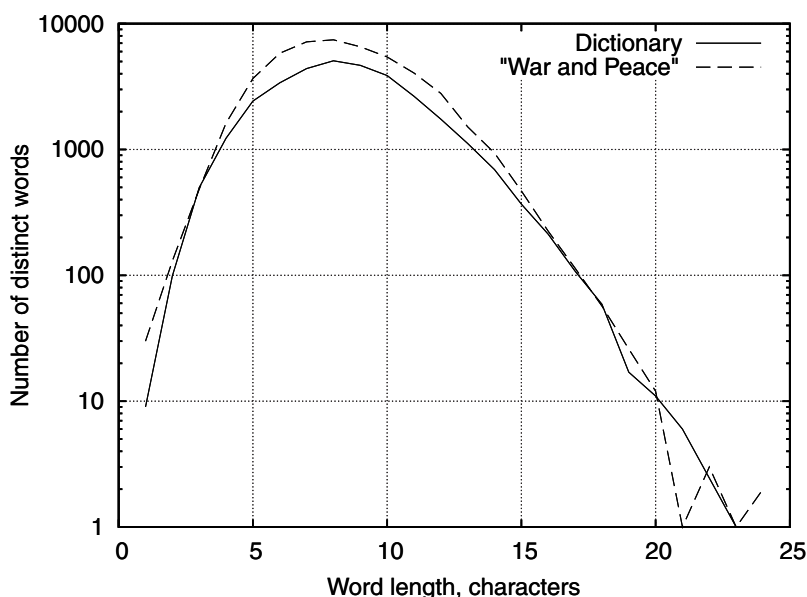
Fig. 2. Distribution of words by length.

1.2. Guiraud's semic matrices

A radically different approach was taken by the French linguist Pierre Guiraud. He suggested that Zipf's law "would be *produced* by the structure of the signified, but would be *reflected* by that of the signifier" (Guiraud, 1968). Specifically, suppose that all word meanings can be represented as superpositions of a small number of elementary meanings, or *semes*. In keeping with the structuralist paradigm, each seme is a binary opposition, such as *animate/inanimate* or *actor/process* (Guiraud's examples). Each seme can be positive, negative, or unmarked in any given word. Assuming that the semes are orthogonal, so that seme values can be combined with each other without constraints, with N semes, there can be $2N$ single-seme words (i.e., words where only one seme is marked), $4N(N-1)$ two-seme words, and so on. The number of words increases roughly exponentially with the number of marked semes. If all semes have the same probability of coming up in a speech situation, the probability of a word with m marked semes is also exponential in m. This leads to Zipf's distribution for words.

This model is very attractive conceptually and heuristically, but it is too rigid and schematized to be realistic. One problem with it lies in the assumption that any combination of semes should be admissible, but even Guiraud's own examples show that it would be very hard to satisfy this requirement. For example, if the *actor/process* seme is marked in a word with the value of *process*, then the word is a verb, and *animate/inanimate* has to be unmarked in it: there are no animate or inanimate verbs (i.e., there is no verb that would differ from, say, *laugh* only in that it's inanimate)—and that undermines the notion of unrestricted combinability of semes.

1.3. Models based on optimality principles

Different authors proposed models based on the observation that Zipf's law maximizes some quantity. If this quantity can be interpreted as a measure of "efficiency" in some sense, then such model can claim explanatory power.

Zipf himself surmised in 1949 that this distribution may be a result of "effort minimization" on the part of both speaker and listener. This argument goes approximately as follows: the broader[3] the meaning of a word, the more common it is, because it is usable in more situations. More common words are more accessible in memory, so their use minimizes the speaker's effort. On the other hand, they increase the listener's effort, because they require extra work on disambiguation of diffuse meanings. As a result of a compromise between speaker and listener, a distribution emerges.

Zipf did not construct any quantitative model based on these ideas. The first model of this sort was proposed by Mandelbrot (1953). It optimizes the cost of speech production per bit of information transferred. Zipf's law follows under the assumptions that (1) the cost of a word is proportional to its length, and (2) the number of words of a given length in language is also proportional to that length. In this form, the model turns out to be the reformulation of the random typing model. To quote Mandelbrot (1966), "These variants are fully equivalent mathematically, but they appeal to [. . .] different intuitions [. . .]."

A different optimality model was proposed by Arapov and Shrejder (1978). They demonstrated that Zipfian distribution maximizes a quantity they call *dissymmetry*, which is the sum of two entropies: $\Phi = H + H^*$, where H is the standard entropy that measures the number of different texts that can be constructed from a given set of word tokens (some of which are identical), while H^* measures the number of ways *the same* text can be constructed from these tokens by permutations of identical word tokens. The former quantity is maximized when all word tokens in a text are different, the latter one when they are all the same, and the Zipfian distribution with its steep initial decline and long tail provides the best compromise. This theoretical construct does possess a certain pleasing symmetry, but its physical meaning is rather obscure, though the authors claim that Φ should be maximized in "complex systems of natural origin."

Balasubrahmanyan and Naranan (2002) took a similar approach. They too, aimed to demonstrate that the language is a "complex adaptive system," and that Zipf's law is achieved in the state of maximum "complexity." Their derivation also involved defining and combining different entropies, some of which are related to the permutation of identical word tokens in the text. Both approaches of Arapov and Shrejder (1978) and Balasubrahmanyan and Naranan (2002), in our view, have the same two problems. First, the quantity being optimized is not compellingly shown to be meaningful. Second, no mechanism is proposed to explain why and how the language could evolve toward the maximum.

In a recent series of articles by Ferrer i Cancho with coauthors (see Ferrer i Cancho 2005a, 2005b, and references therein) the optimization idea is brought closer to reality. Ferrer i Cancho's models significantly differ from the other models in that they are based on the idea that the purpose of language is communication, and that it is optimized for the efficiency of communication. Ferrer i Cancho's models postulate a finite set of words and a finite set of objects or stimuli with a many-to-many mapping between the two. Multiple objects may be

linked to the same word because of polysemy, while multiple words may be linked to the same object because of synonymy. It is assumed that the frequency of a word is proportional to the number of objects it is linked to. Next, Ferrer i Cancho introduces optimality principles and, in some cases, constraints, with the meaning of coder's effort, decoder's effort, mutual entropy between words and objects, entropy of signals, and so on. By maximizing goal functions constructed from combinations of these quantities, Ferrer i Cancho demonstrated the emergence of Zipf's law in phase transition-like situations with finely tuned parameters.

The treatment in the present work, although quite different in spirit, shares two basic principles with Ferrer i Cancho's models and, in a way, with Guiraud's ideas. Namely, we take the view that word usage, and hence, frequency, is determined largely by its semantic properties. On the other hand, we do not assume any optimality principles, and neither do we use the notion of least effort. Instead, we show that Zipf's law can be obtained as a consequence of a purely linguistic notion of avoidance of excessive synonymy. It should be noted that our model may well be compatible with those of Ferrer i Cancho and Simon.

The present model is based on the notion of *meaning extent*, which we discuss in Section 2: a basic assumption is that words with broader, or more generic, meanings are more frequent than words with narrower, or more specific, meanings. We argue that words with different degrees of specificity are needed for efficient communication (Section 2); that Zipf's law can result from a particular way of arranging word meanings over the space of all meanings (Section 3); and that such arrangements can naturally arise from the historical dynamics under which (1) word meanings tend to expand, and (2) close synonyms are disfavored (Sections 4, 5).

If one is to claim that a word's frequency in texts is related to some properties of its meaning, a theory of meaning must be presented upfront. Fortunately, it doesn't have to be comprehensive, rather we'll outline a *minimal* theory that only deals with the meaning of words, rather than statements, and with just one aspect of meaning that we are concerned with here: its extent.

2. Synonymy, polysemy, semantic space

The nature of meaning has long been the subject of profound philosophical discourse. What meaning is and how meanings are connected to words and statements is not at all a settled question. But whatever meaning is, we can talk about "the set of all meanings," or "semantic space," because this doesn't introduce any significant assumptions about the nature of meaning (besides the assumption that it is reasonably stable in time and common across individuals). Of course, we should exercise extreme caution to avoid assuming any structure on this set that we don't absolutely need. For example, it would be unwise to think of semantic space as a Euclidean space with a certain dimensionality. Indeed, this would imply that meanings can be added together and multiplied by numbers to produce other meanings, which is probably not a tenable proposition. One could justify the assumption of a metric on semantic space, because we commonly talk about meanings being more or less close to each other, effectively assigning a distance to a pair of meanings. However, as we won't need it for the purposes of this work, no metric will be assumed.

In fact, the only additional structure that we do assume on semantic space S, is a measure. Mathematically, a measure on S assigns a non-negative "volume" to subsets of S, such that the volume of a union of two disjoint subsets is the sum of their volumes.[4] We need a measure so that we can speak of words being more "specific" or "generic" in their meanings. If a word w has a meaning $m(w) \subset S$, then the "degree of generality," or "extent," or "broadness" of its meaning is the measure $\mu(m(w))$, i.e., the "volume" that the word covers in semantic space.[5] Note that a measure does not imply a metric: thus, there is a natural measure on the unordered set of letters of the Latin alphabet (the "volume" of a subset is the number of letters in it), but to define a metric, i.e., to be able to say that the distance between a and b is, say, 1, we need to somehow order the letters.

We understand "meaning" in a very broad sense of the word. We are willing to say that any word has meaning. Even words like *the* and *and* have meanings: that of definiteness and that of combining, respectively. We also want to be able to say that such words as *together*, *join*, *couple*, and *fastener* have meanings that are subsets of the meaning of *and*. By that we mean that whenever one of these words comes up in a speech situation, *and* also comes up, even if it is not uttered, because each of them evokes a mental image with an essential component of separate entities being combined in some way.[6] We do not make a distinction between connotation and denotation, intension and extension, etc. This means that "semantic space" S may include elements of very different nature, such as the notion of a mammal, the emotion of love, the feeling of warmth, and your cat Fluffy. Such eclecticism shouldn't be a reason for concern, since words are in fact used to express and refer to all these kinds of entities and many more.

We only deal with isolated words here, without getting into how the meaning of *this dog* results from the meanings of *this* and of *dog*. Whether it is simply a set theoretic intersection of *thisness* and *dogness* or something more complicated, we don't venture to theorize. One of the problems here is that semantic space itself is not static. New meanings are created all the time as a result of human innovation in the world of objects, as well as in the world of ideas: poets and mathematicians are especially indefatigable producers of new meanings.[7] However, when dealing with individual words, as is the case with Zipf's law, one can ignore this instability, since words and their meanings are much more conservative, and only a small fraction of new meanings created by the alchemy of poetry and mathematics eventually claim words for themselves.

Note that up to now we didn't have to introduce any structure on S, not even a measure. Even the cardinality of S is not specified, it could be finite, countable, or continuous. But we do need a measure for the next step, when we assume that the frequency of the word w is proportional to the extent of its meaning, i.e., to the measure $\mu(m(w))$. The more generic the meaning, the more frequent the word, and vice versa, the more specific the meaning, the less frequent the word.

We don't have data to directly support this assumption, mostly because we don't know how to independently measure the extent of a word's meaning. However, one could obtain a rough estimate of meaning extent from such quantities as the length of the word's dictionary definition or the number of all hyponyms of the given word (for instance, using WordNet[8]). Indeed, Baayen and Fermin Moscoso Del Prado (2005) found a statistically significant correlation between the frequency of a verb and the number of synonymic sets it is a part of in English

and Dutch. We also provide a different kind of experimental evidence for proportionality between frequency and meaning extent in Appendix.

It is essential for this hypothesis that we do not reduce meaning to denotation, but include connotation, stylistical characteristics, etc. It is easy to see that the word frequency can't be proportional to the extent of its denotation alone: there are many more radiolaria than dogs in the world, but the word *dog* is much more frequent than the word *radiolaria*. In fact, the word *dog* is also more frequent than words *mammal* and *quadruped*, though its denotation (excluding figurative senses, though) is a strict subset and thus more narrow.[9] But the frequency of the word *mammal* is severely limited by its being a scientific term, i.e., its meaning extent is wider along the denotation axis, but narrower along the stylistic axis ("along the axis" should be understood metaphorically here, rather than technically). In the realm of scientific literature, where the stylistic difference is neutralized, *mammal* is quite probably more frequent than *dog*.

It's interesting to note in this connection that according to the frequency dictionary (Sharoff, n.d.), the word *собака* 'dog' is more frequent in Russian than even the words *животное* and *зверь* 'animal, beast,' although there is no significant stylistical differences between them. To explain this, note that of all animals, only the dog and the horse are so privileged (i.e., other animal nouns are less frequent than the word *животное* 'animal'). A possible reason is that the connotation of *animal* in the common language includes not so much the opposition 'animal as non-plant' as the opposition 'animal as non-human.' But the dog and the horse are characteristically viewed as "almost-human" companions, and thus in a sense do not belong to animals at all, which is why the corresponding words do not have to be less frequent.

If word frequency is a function of semantics, Zipf's law, with its wide range of frequencies, should reflect the structure of a natural language vocabulary. Indeed, word meanings range from very generic to extremely specific. There are at least two pretty obvious reasons for this. First, in some cases we need to refer to any object of a large class, as in *take a seat*, while in other cases we need a reference to a narrow subclass, as in *you're sitting on a Chippendale*. In the dialogue (3) two words, the generic one and the specific one, point to the same object.

— I want some Tweakles!
— Candy is bad for your teeth. (3)

Second, when context provides disambiguation, we tend to use generic words instead of specific ones. Thus, inhabitants of a large city environs say *I'm going to the city*, and avoid naming it by name. Musicians playing winds call their instrument a *horn*, whether it's a trumpet or a tuba. Pet owners say *feed the cat*, although the cat has a name, and some of them perform a second generalization to *feed the beast* (also heard in Russian as *накорми животное*). In fact, the word *candy* in the *Tweakles* example fulfills both roles at once: it generalizes to all candies, because all of them are bad for your teeth, but also it refers to this specific candy by contextual disambiguation. We even use the ultimate generic placeholders like *thingy* when we dropped it and need somebody to pick it up for us.[10] A colorful feature of Russian vernacular is the common use of desemantized expletives as generic placeholders, where whole sentences complete with adjectives and verbs can be formed without a significant word. What may not be generally appreciated is that this strategy may, at least in some cases, turn out to be highly efficient. According to the author V. Konetsky (1980), radio communications

of Russian WWII fighter pilots in a dogfight environment, where a split-second delay can be fatal, consisted almost entirely of such pseudo-obscene placeholder words, as evidenced by recordings. It hardly could have been so, were it not efficient.

The reason for this tendency to generalize is very probably the Zipfian minimization of effort for the speaker. A so-called *word frequency effect* is known in psycholinguistics, whereby the more frequent the word, the more readily it is retrieved from memory (cf. Akmajian, Harnish, Demers, & Farmer, 1995; Carroll, 1994). However, contrary to Zipf, it doesn't seem plausible that such a generalization makes understanding more difficult for the listener. The whole idea of pitching the speaker against the listener in the effort minimization tug-of-war appears to fly in the face of communication as an essentially cooperative phenomenon, where a loss or gain for one party is a loss or gain for both. Again, we don't have hard data, but intuitively it seems that when there is only one city in the context of the conversation, it is even easier for the listener if it's referred to as *the city* rather than *Moscow* or *New York*. *I'm going to the city* means *I'm going you know where* while *I'm going to London* means *I'm going to this one of a thousand places where I could possibly go*. The first expression is easier not only for the speaker, but for the listener as well, because one doesn't have to pull out one's mental map of the world, as with the second expression. Or, put in information theoretic terms, *the city* carries much less information than *Shanghai* because the generic word implies a universal set consisting of one element, while the proper name implies a much larger universal set of dozens of toponyms—but most of this extra information is junk and has to be filtered out by the listener, if Shanghai is in fact The City; and this filtering is a wasted effort.

In the words of Stephen Levinson (2000), "inference is cheap, articulation expensive, and thus the design requirements are for a system that maximizes inference." He demonstrated that this principle applies to a wide range of linguistic constructions.

To summarize this discussion, the organization of words over semantic space in such a way that each element is covered by a hierarchy of words with different extents of meaning is useful for efficient communication. In this way, the speaker can select a word that refers to the desired element with the desired degree of precision. Or, rather, the most imprecise word that still allows disambiguation in the given context. The benefit here is that less precise words are more frequent, and thus more accessible for both the speaker and the listener, which can be said to minimize the effort for both. Another benefit is that such organization is conducive to building hierarchical classifications, which people are rather disposed to do (whether that's because the world itself is hierarchically organized is immaterial here).

3. Zipf's law and Zipfian coverings

It turns out that hierarchical organization of meanings over semantic space can also give rise to Zipf's law. As an illustration consider a simple example of such a mapping: let word number 1 cover the whole of S, words number 2 and 3 cover one-half of S each, words 4 through 7 cover one-quarter of S each, etc. (see Fig. 3). It is easy to see that this immediately leads to Zipf's distribution. Indeed, the extent of the k-th word's meaning is

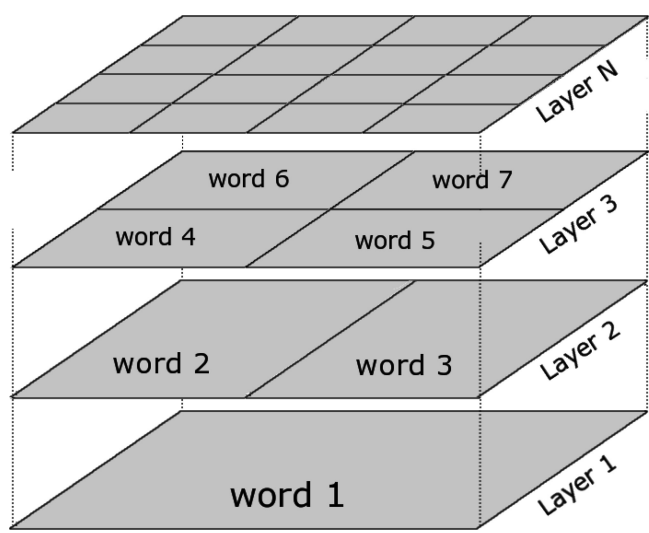$$\mu_k = 2^{-\lceil \log_2 k \rceil} \tag{4}$$

Fig. 3. Example of a hierarchical organization of semantic space.

Under the assumption that the frequency of a word f_k is proportional to the extent of its meaning μ_k, this is equivalent to (Equation 1), except for the piecewise-constant character of (4), see Fig. 4. What matters here is the overall trend, not the fine detail.

Of course, real word meanings do not follow this neat, orderly model literally. But it gives us an idea of what Zipf's distribution (Equation 1) can be good for. Consider a subset of all words whose frequency rank is in the range $[k, k\rho]$ with some k and $\rho > 1$. Zipf's distribution
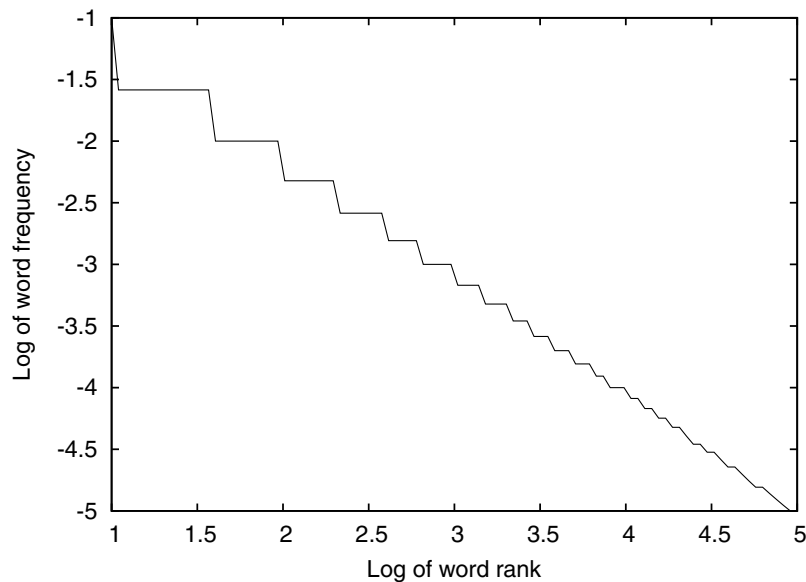

Fig. 4. Frequency distribution for hierarchical model Fig. 3.

has the following property: the sum of frequencies of words in any such subset depends only on the scaling exponent ρ (asymptotically with $k \to \infty$), since by Riemann's formula, it is bounded by inequalities

$$\ln \frac{k\rho}{k} = \int_k^{k\rho} \frac{dx}{x} < \sum_{j=k}^{k\rho} \frac{1}{j} < \int_{k-1}^{k\rho-1} \frac{dx}{x} = \ln \frac{k\rho - 1}{k - 1} \tag{5}$$

By our basic assumption, word frequency is proportional to the extent of its meaning. Thus, we can choose ρ so that the words in any subset $[k, k\rho]$ have a total measure equal to that of semantic space S, and so could cover S without gaps and overlaps. Of course, this does not guarantee that they *do* cover S in such a way, but only for Zipf's distribution does such a possibility exist.[11]

Let us introduce some notation at this point, to avoid bulky descriptions. Let S be a measurable set with a finite measure μ. Define a *covering* of S as an arbitrary sequence of subsets $C = \{m_i\}, m_i \subset S, \mu(m_i) \geq \mu(m_{i+1})$. Let the *gap* of C be the measure of the part of S not covered by C,

$$\text{gap}(C) = \mu(S) - \mu\left(\bigcup m_i\right), \tag{6}$$

and let *overlap* of C be the measure of the part covered by more than one m_i,

$$\text{overlap}(C) = \mu(\{x \mid x \in \text{more than one } m_i\}) = \mu\left(\left\{x \mid x \in \bigcup_{i \neq j} m_i \cap m_j\right\}\right) \tag{7}$$

Finally, define (ρ, k)-*layer* of C as subsequence $\{m_i\}, i \in [k, k\rho]$ for any starting rank $k > 0$ and some scaling exponent $\rho > 1$.

With these definitions, define *Zipfian covering* as an infinite covering such that for some ρ, both gap and overlap of (ρ, k)-layers vanish as $k \to \infty$. This means that all words with ranks in any range $[k, k\rho]$ cover the totality of S and do not overlap (asymptotically in $k \to \infty$). Or, to look at it from a different point of view, each point in S is covered by a sequence of words with more and more precise (narrow, specific) meanings, with precision growing in geometric progression with exponent ρ as illustrated in Fig. 3. Again, this organization of semantic space would be efficient, since it ensures the homogeneity of the "universal classification": precision of terms (inverse of meaning extent) increases by a constant factor each time you descend to the next level. This is why the exponent $B = 1$ in (Equation 2) is special: with other exponents one doesn't get the scale-free covering.

The covering in Fig. 3 is an example of Zipfian covering, though a somewhat degenerate one. We will not discuss the existence of other Zipfian coverings in the strict mathematical sense, since the real language has only a finite number of words anyway, so the limit of an infinite word rank is unphysical. We need the notion of Zipfian covering as a *strict* idealized model which is presumably in an *approximate* correspondence with reality.

Note though that since $\sum_1^n 1/j$ grows indefinitely as $n \to \infty$, Zipf's law can be normalized so that all word frequencies sum up to 1, only if cut off at some rank N. The nature of this cut-off becomes very clear in the present model: the language does not need words with arbitrary

narrow meanings, because such meanings are more efficiently represented by combinations of words.

However, as noted above, demonstrating that Zipf's law satisfies some kind of optimality condition alone is not sufficient. One needs to demonstrate the existence of a plausible local dynamics that could be responsible for the evolution towards the optimal state. To this end, we now turn to the mechanisms and regularities of word meaning change. Specifically, we are interested in two basic processes: meaning expansion and avoidance of close synonymy.

4. Zipfian coverings and avoidance of excessive synonymy

Word meanings change as languages evolve. This is a rule, rather than an exception (see, e.g. Hock & Joseph, 1996, Maslov, 1987; most of the examples below come from these two sources). There are various reasons for semantic change, among them need, other changes in the language, social factors, "bleaching" of old words, etc. Some regularities can be observed in the direction of the change. Thus, in many languages, words that denote grasping of physical objects with hands develop the secondary meaning of understanding, "grasping of ideas with mind": Eng. *comprehend* and *grasp*, Fr. *comprendre*, Rus. *понимать* and *схватывать*, Germ. *fassen* illustrate various stages of this development. Likewise, Eng. *clear* and Rus. *ясный*, *прозрачный* illustrate the drift from optical properties to mental qualities. As a less spectacular, but ubiquitous example consider metonymic extension from action to its result, as in Eng. *wiring* and Rus. *проводка* (idem). There may also be deeper and more pervasive regularities (Traugott & Dasher, 2005). Paths from old to new meanings are usually classified in terms of metaphor, metonymy, specialization, ellipsis, etc. (Crystal, 2003).

Polysemy, multiplicity of meanings, is pervasive in language: "cases of monosemy are not very typical" (Maslov, 1987); "We know of no evidence that language evolution has made languages less ambiguous" (Wasow, Perfors, & Beaver, 2005); "word polysemy does not prevent people from understanding each other" (Maslov, 1987). There is no clear-cut distinction between polysemy and homonymy, but since Zipf's law deals with typographic words, we do not have to make this distinction. In the "meaning as mapping" paradigm employed in the present work, one can speak of different *senses*[12] of a polysemous word as subsets of its entire meaning. Senses may be disjoint (cf. *sweet*: 'tasting like sugar' and 'amiable'[13]), they may overlap (*ground*: 'region, territory, country' and 'land, estate, possession'), or one may be a strict subset of the other (*ball*: 'any round or roundish body' and 'a spherical body used to play with').

Note that causes, regularity, and paths of semantic change are not important for our purposes, since we are only concerned here with the extent, or scope, of meaning. And that can change by three more or less distinct processes: extension, formation, and disappearance of senses (although the distinction between extension and formation is as fuzzy as the distinction between polysemy and homonymy).

Extension can be illustrated by the history of Eng. *bread* which initially meant '(bread) crumb, morsel' (Hock & Joseph, 1996, p. 11), or Rus. *палец*, 'finger, toe,' initially 'thumb' (Maslov, 1987, pp. 197–198). With extension, the scope of meaning increases.

Formation of new senses may cause an increase in meaning scope or no change, if the new sense is a strict subset of the existing ones. This often happens through ellipsis, such as with

Eng. *car*, 'automobile' < *motor car* (Hock & Joseph, 1996, p. 299) or parallel Rus. машина < автомашина. In this case, the word initially denotes a large class of objects, while a noun phrase or a compound with this word denotes a subclass. If the subclass is important enough, the specifier of the phrase can be dropped (via the generalization discussed above), and this elliptic usage is reinterpreted as a new, specialized meaning.

Meanings can decrease in scope as a result of a sense dropping out of use. Consider Eng. *loaf* < OE *hlaf*, 'bread.' Schematically one can say that the broad sense 'bread in all its forms' disappears, while the more special sense 'a lump of bread as it comes from the oven' persists. Likewise, Fr. *chef*, initially 'head as part of body,' must have first acquired the new sense 'chief, senior' by metaphor, and only then lost the original meaning.

In the mapping paradigm, fading of archaic words can also be interpreted as narrowing of meaning. Consider Rus. перст, 'finger (arch., poet.).' The reference domain of this word is almost the same as that of палец 'finger (neut.)' (excluding the sense 'toe'), but its use is severely limited because of a strong flavor. Thus, meaning scope is reduced here along the connotation dimension. But since we consider both denotation and connotation as constituents of meaning, narrowing of either amounts to narrowing of meaning. Both types of narrowing are similar in that they tend to preserve stable compounds, like *meatloaf* or один, как перст 'lone as a finger.'

There is no symmetry between broadening and narrowing of meaning. Development of new senses naturally happens all the time without our really noticing it. But narrowing is typically a result of competition between words (except for the relatively rare cases where a word drops out of use because the object it denoted disappears). Whatever driving forces there were, but *hlaf* lost its generic sense only because it was supplanted by the expanding *bread*, *chef* was replaced by the expressive *tête* < *testa*, 'crock, pot,' and перст by палец (possibly, also as an expressive replacement). As another illustration, consider the "blocking" effect (Briscoe, 2006) where a regular word derivation, such as *stealer* from *steal*, is preempted (except in special cases) by an existing non-derivational synonym, such as *thief*.

This is summarized by Hock and Joseph (1996):

> [. . .] complete synonymy—where two phonetically distinct words would express exactly the same range of meanings—is highly disfavored. [. . .] where other types of linguistic change could give rise to complete synonymy, we see that languages— or more accurately, their speakers—time and again seek ways to remedy the situation by differentiating the two words semantically. (p. 236)

And by Maslov (1987):

> [. . .]since lexical units of the language are in *systemic relationships* with each other via semantic fields, synonymic sets, and antonymic pairs, it is natural that changes in one element of a microsystem entails changes in other related elements. (p. 201)

One important feature of this process of avoiding excessive synonymy is that words compete only if their meanings are similar in scope. That is, a word whose meaning overlaps with that of a significantly more general word will not feel the pressure of competition. As discussed earlier, the language needs (or rather its speakers need) words of different scope of meaning, so both the more general and the more specific words retain relevance. This is in a way similar

to the effect reported by Wasow et al. (2005) where it was found (both by genetic simulation and by studying polysemous word use in Brown Corpus) that polysemy persists if one of the senses is significantly more common than the other. Although this result is related to polysemy rather than to synonymy, it also can be interpreted as an example of meaning competition. As such, it shows that meanings do not interact (compete) if they are sufficiently different in scope, whether they belong to the same word (polysemy) or to different words (synonymy).

Summarizing the above, one can say that *meanings tend to increase in scope, unless they collide with other meanings of a similar scope, while meanings of significantly different scope do not interact*. But this looks just like a recipe for the development of approximately Zipfian coverings discussed in the previous section! Indeed, this kind of evolution could lead to semantic space being covered almost without gaps and overlaps by each subset of all words of approximately the same scope. In order to substantiate this idea two numerical models were developed.

5. Numerical models

The models simulate the two basic processes by which word meanings change in extent: generalization and specialization. They are very schematic and are not intended to be realistic. We model the semantic space by the interval $S = [0, 1]$ and word meanings by sub-intervals on it. The evolution of the sub-intervals is governed by the following algorithms.

Generalization model
1. Start with a number N of zero-length intervals $r_i \subset S$ randomly distributed on S.
2. At each step, grow each interval symmetrically by a small length δ, if it is not *frozen* (see below).
3. If two unfrozen intervals intersect, freeze one of them (the one to freeze is selected randomly).
4. Go to step 2 if there is more than one unfrozen interval left, otherwise stop.

Informally, words in the generalization model have a natural tendency to extend their meanings, unless this would cause excessive synonymy. If two expanding words collide, one of them stops growing. The other one can eventually encompass it completely, but that is not considered to be "excessive synonymy," since by that time, the growing word is significantly more generic, and words of different generality do not compete. The initial randomness simulates the effect of other factors not accounted for directly.

Specialization model
1. Start with a number N of intervals, whose centers are randomly distributed on S and lengths are uniformly distributed on $[0, 1]$.
2. For each pair of intervals r_i, r_j, if they intersect and their lengths l_i, l_j satisfy $1/\gamma < l_i/l_j < \gamma$, decrease the smaller interval by the length of their intersection.
3. Continue until there is nothing left to change.

The specialization model simulates avoidance of excessive synonymy where synonyms created by some other processes in language compete and one supplants the other in their
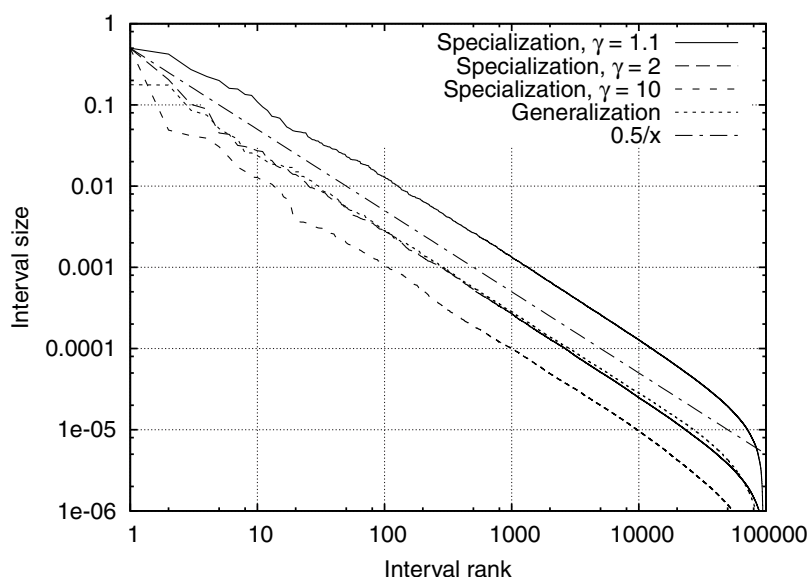
Fig. 5. Zipf's law generated by specialization and generalization models.

common area, such as with Fr. *chef* and *tête* mentioned above. Parameter γ determines by how much the two words can differ in extent and still compete.

When run, both these models reliably generate interval sets with sizes distributed by Zipf's law with exponent $B = 1$. The generalization model is parameter-free (except for the number of intervals, which is not essential as long as it is large enough). The specialization model is surprisingly robust with respect to its only parameter γ: we ran it with $\gamma \in [1.1, 10]$ with the same result—see Fig. 5. It is interesting to note that with $\gamma = 1.1$, specialization model even reproduces the low-rank behavior of the actual rank distributions, but it is not clear whether this is a mere coincidence or something deeper.

Both models also generate interval sizes that approximately satisfy the definition of Zipfian covering. That is, if we consider the subset of all intervals between ranks of k and ρk, they should cover the whole $[0, 1]$ interval with no gap and overlap—for some fixed ρ and asymptotically in $k \to \infty$. Fig. 6 shows the gap, i.e. the total measure of that part of S not covered by these intervals, as a function of the starting rank k. Scaling parameter ρ was chosen so that the sum of interval lengths between ranks k and $k\rho$ was approximately equal to 1. The fact that the gap indeed becomes very small demonstrates that the covering is approximately Zipfian. This effect does not follow from Zipf's law alone, because it depends not only on the size distribution, but also on where the intervals are located on S. On the other hand, Zipf's distribution does follow from the Zipfianness of the covering.

Of course, these models provide but an extremely crude simulation of the linguistic processes. However, the robustness of the result suggests that quite possibly the models represent a much larger class of processes that can lead to Zipfian coverings and hence to Zipf's distributions under the same very basic assumptions.
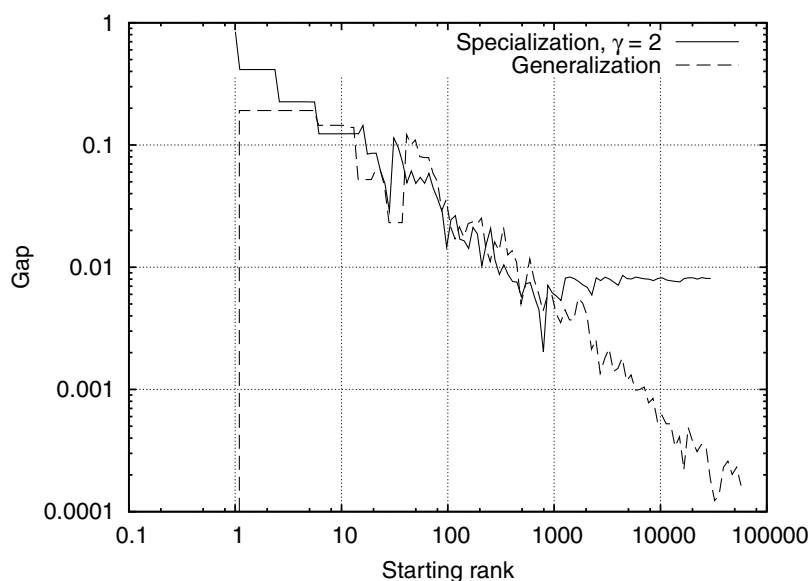
Fig. 6. The gap of (k, ρ)-layer decreases with increasing k.

6. Discussion

To summarize, we propose the following.

1. Word meanings have a tendency to broaden.
2. On the other hand, there is a tendency to avoid excessive synonymy, which counteracts the broadening.
3. Synonymy avoidance does not apply to any two words that differ significantly in the extent of their meanings.
4. As a result of this, word meanings evolve in such a way as to develop a multi-layer covering of the semantic space, where each layer consists of words of approximately the same broadness of meaning, with minimal gap and overlap.
5. We call arrangements of this sort *Zipfian coverings*. It is straighforward to show that they possess Zipf's distribution with exponent $B = 1$.
6. Since word frequency is likely to be in a direct relationship with the broadness of its meaning, Zipf's distribution for one of them entails the same distribution for the other.

This model is rooted in linguistic realities and demonstrates the evolutionary path for the language to develop Zipf's distribution of word frequencies. It not only predicts the power law, but also explains the specific exponent $B = 1$. Even though we argue that Zipfian coverings are in some sense "optimal," we do not need this optimality to be the driving force, and can in fact entirely do away with this notion, because the local dynamics of meaning expansion and synonymy avoidance are sufficient. The "meaning" of Zipf's distribution becomes very clear in this proposal.

The greatest weakness of the model is that it is based upon a rather vague theory of meaning. The assumption of proportionality of word frequency to the extent of its meaning is natural (indeed, if one accepts the view that "meaning is usage," it becomes outright tautological), but it is unverifyable as long as we have no independent way to measure both quantities. However, some evidence in support of this assumption can be obtained, as shown in the Appendix. Further studies are necessary to clarify this issue. As one possibility, a direct estimate of word meaning extent might be obtained on the basis of the Moscow semantic school's Meaning–Text Theory (e.g., Mel'čuk, 1988; Mel'čuk, 1997), which provides a well-developed framework for describing meanings.

It should be noted that other, more specific, skewed distributions in the language have been observed, some approximating power or exponential laws more or less loosely (see, e.g., Briscoe (2006) and references therein), but the original Zipf's law still stands out in terms of both fit quality and stability of the exponent. It is not claimed here that other power-law distributions are also generated by the proposed mechanism. Because they can be produced by various processes, each case should be considered separately. Our purpose here was not so much to propose a new model for Zipf's law as to demonstrate that it can reveal some underlying properties of the language.

The treatment in this work was restricted to the linguistic domain. However, as is well known, Zipf's law is observed in many other domains. The mechanism of competitive growth proposed here could be applicable to some of them. Whenever one has entities that (a) exhibit the tendency to grow, and (b) compete only with like-sized entities, the same mechanism will lead to Zipfian covering of the territory and consequently to Zipf's distribution of sizes.

Notes

1. For example, Arapov and Shrejder (1978) and Wheeler (2002), respectively.
2. As a historical aside, it is interesting to mention that Simon and Mandelbrot exchanged rather spectacularly sharp criticisms of each other's models in a series of letters in the journal *Information and Control* in 1959–1961.
3. We will use *broad* or *generic* on the one hand and *narrow* or *specific* on the other to characterize the *extent* or *scope* of a word's meaning.
4. Many subtleties are omitted here, such as the fact that a measurable set may have non-measurable subsets.
5. We assume that meanings of words correspond to subsets of S. It may seem natural to model them instead with fuzzy subsets of S, or, which is the same, with probability distributions on S. Meanings may also be considered as prototypes, i.e., attractors in semantic space, but our model can be adapted to this view as well.
6. This statement could be tested experimentally using the standard semantic priming technique, by checking whether priming with the word *and* speeds up the recognition of the word *join*, while priming with words, say, *off* or *but* doesn't.
7. For a much deeper discussion see Manin (1981). In particular, it turns out that the rich paraphrasing capacity of language may paradoxically be evidence of high referential efficiency.

D. Y. Manin/Cognitive Science 32 (2008)

8. http://wordnet.princeton.edu/
9. I owe this example to Tom Wasow.
10. As Ray Bradbury wrote in his 1943 story *Doodad*: "Therefore, we have the birth of incorrect semantic labels that can be used to describe anything from a hen's nest to a motor-beetle crankcase. A doohingey can be the name of a scrub mop or a toupee. It's a term used freely by everybody in a certain culture. A doohingey isn't just one thing. It's a thousand things." WordNet lists several English words under the definition "something whose name is either forgotten or not known." Interestingly, some of these words (*gizmo, gadget, widget*) developed a second sense, "a device that is very useful for a particular job," and one (*gimmick*) similarly came to also mean "any clever (deceptive) maneuver."
11. Indeed, if it is possible for any subset $[k, k\rho]$ to cover S without gaps and overlaps, the total measure of any such subset needs to be the same independent of k (but dependent on ρ). The only frequency distribution $f(k)$ satisfying this condition is $f(k) = C/k$ with some constant C. To see this, consider $\rho = 1 + \epsilon, \epsilon \ll 1$. Then the sum of frequencies in the interval $[k, k\rho]$ is asymptotically given by $f(k)k(\rho - 1)$, and for it to be independent of k, $f(k)$ has to be inversely proportional to k.
12. This should not be confused with the dichotomy of *sense* and *meaning*. Here we use the word *sense* as in "the dictionary gave several senses of the word."
13. Definitions here and below are from 1913 edition of Webster's dictionary.
14. http://www.cogsci.rpi.edu/CSJarchive/Supplemental

Acknowledgment

References

Akmajian, A., Harnish, R. M., Demers, R. A., & Farmer, F. K. (1995). *Linguistics. An introduction to language and communication*. Cambridge, MA: MIT Press.

Arapov, M. V., & Shrejder, Y. A. (1978). Zakon cipfa i princip dissimmetrii sistem [Zipf's law and system dissymmetry principle]. In *Semiotics and informatics* (Vol. 10, pp. 74–95). Moscow: VINITI.

Baayen, R., & Fermin Moscoso Del Prado, M. (2005). Semantic density and past-tense formation in three Germanic languages. *Language*, *81*, 666–698.

Balasubrahmanyan, V. K., & Naranan, S. (2002). Algorithmic information, complexity and Zipf's law. *Glottometrics*, (4), 1–26.

Briscoe, E. J. (2006). Language learning, power laws and sexual selection. In K. Smith, A. D. M. Smith, & A. Cangelosi (Eds.), *Proceedings of the 6th international conference on the evolution of language* (pp. 19–26). World Scientific. (http://www.cl.cam.ac.uk/ ejb1/mind-soc-version.pdf)

Carroll, D. W. (1994). Psychology of language. In (pp. 242–248). Pacific Grove: Brooks/Cole Publishing Company.

Crystal, D. (2003). *The cambridge encyclopedia of language* (2nd ed.). Cambridge: Cambridge University Press.

Ferrer i Cancho, R. (2005a). Decoding least effort and scaling in signal frequency distributions. *Physica A*, *345*, 275–284.

Ferrer i Cancho, R. (2005b). Hidden communication aspects in the exponent of Zipf's law. *Glottometrics*, *11*, 98–119.

Ferrer i Cancho, R. (2005c). The variation of Zipf's law in human language. *The European Physical Journal B–Condensed Matter and Complex Systems*, *44*, 249–257.

Guiraud, P. (1968). The semic matrices of meaning. *Social Science Information*, 7, 131–139.

Hock, H. H., & Joseph, B. D. (1996). *Language history, language change, and language relationship*. Berlin: Mouton de Gruyter.

Konetsky, V. (1980). *Vcherashnie zaboty. Solenyj led.* [Yesterday's troubles. Salt-ice.] Moscow: Izvestiya.

Levinson, S. C. (2000). *Presumptive meanings: The theory of generalized conversational implicature*. Cambridge, MA: MIT Press.

Li, W. (1992). Random texts exhibit Zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, *38*(6), 1842–1845.

Mandelbrot, B. (1953). An informational theory of the statistical structure of languages. In W. Jackson (Ed.), *Communication theory* (pp. 486–502). Woburn, MA: Butterworth.

Mandelbrot, B. (1966). Information theory and psycholinguistics: A theory of word frequencies. In P. F. Lazarsfield & N. W. Henry (Eds.), *Readings in mathematical social sciences* (pp. 350–368). Cambridge, MA: MIT Press.

Mandelbrot, B. (1982). *The fractal geometry of nature*. New York: Freeman.

Manin, Y. I. (1981). Expanding constructive universe. In A. P. Ershov & D. Knuth (Eds.), *Algorithms in modern mathematics and computer science: proceedings, Urgench, Uzbek SSR, September 16–22, 1979* (Vol. 122). Berlin–New York: Springer-Verlag.

Maslov, Y. S. (1987). *Vvedenie v yazykoznanie* [Introduction to linguistics] (2nd ed.). Moscow: Vysshaya shkola.

Mel'čuk, I. (1988). Semantic description of lexical units in an Explanatory Combinatorial Dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, *1*, 165–188.

Mel'čuk, I. (1997). *Vers une linguistique sens-texte. Lecon inaugurale.* Paris: Collège de France. Retrieved from http://www.olst.umontreal.ca/FrEng/melcukColldeFr.pdf

Mitzenmacher, M. (2003). A brief history of generative models for power law and lognormal distributions. *Internet Math.*, *1*, 226–251.

Pado, S., & Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, *33*(2).

Sharoff, S. (n.d.). *The frequency dictionary for Russian.* Retrieved from http://www.artint.ru/projects/frqlist/frqlist-en.asp

Sharoff, S. (2002, May). Meaning as use: exploitation of aligned corpora for the contrastive study of lexical semantics. In *Proc. of Language Resources and Evaluation Conference (LREC02).* Las Palmas, Spain. Retrieved from http://www.artint.ru/projects/frqlist/lrec-02.pdf

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrica*, *42*, 425–440.

Simon, H. A. (1957). Models of man. In (chap. 6: *On a class of skew distribution functions*). New York: John Wiley and Sons.

Traugott, E. C., & Dasher, R. B. (2005). *Regularity in semantic change*. Cambridge: Cambridge University Press.

Wasow, T., Perfors, A., & Beaver, D. (2005). The puzzle of ambiguity. In O. Orgun & P. Sells (Eds.), *Morphology and the web of grammar: Essays in Memory of Steven G. Lapointe.* Stanford, CT: CSLI Publications. Retrieved from http://www.stanford.edu/Λwasow/Lapointe.pdf

Wheeler, E. S. (2002). Zipf's Law and why it works everywhere. *Glottometrics*, (4), 45–48.

Yule, G. (1925). A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. *F.R.S. Philosophical Transactions of the Royal Society of London (Series B)*, *213*, 21–87.

Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Reading, MA: Addison-Wesley.

Appendix

Meaning and frequency

In this Appendix we'll consider some evidence in favor of the hypothesis that word frequency is proportional to the extent of its meaning. Far from being a systematic study, this is rather a methodological sketch. This study was done in Russian, the author's native language. In the English text we'll attempt to provide translations and/or equivalents wherever possible. Due to space requirements, an abridged version is presented here, the complete text being available as a Web supplement.[14]

Strictly speaking, one could prove the hypothesis only if an explicit measure of meaning extent is proposed. However, the frequency hypothesis allows us to make some verifiable predictions. Suppose that some "head" word w_0 has a set of partial synonyms and/or hyponyms ("specific" words) $\{w_0^1, \ldots, w_0^n\}$, whose meanings together cover the meaning of w_0 without gaps and overlaps. We can make that judgement even without being able to measure meaning extent. Then, by definition, their total meaning extent is equal to that of w_0. In that case, the frequency hypothesis predicts that the sum total of hyponym frequencies should be close to the frequency of the head word.

One cannot expect to find many such examples in the real language. First, pure hyponyms are not very common; it is more common for words to have intersecting meanings, such as with *плохой*, 'bad, poor,' and *худой*, 'skinny; torn, leaky; bad, poor.' Second, only in rare cases can one state confidently that the hyponyms cover the whole meaning of the head word. For example, in the domain of fine arts, *натюрморт* 'still life,' *пейзаж* 'landscape,' and *портрет* 'portrait' are pure hyponyms of the word *картина* 'picture,' but there exist other genres of painting that can't be accounted for with frequency dictionary, since their names are phrases, rather than single words (*жанровая сцена* 'genre painting,' *батальное полотно* 'battle-piece').

Nevertheless, examples of this type do exist. For instance the net frequency of all specific tree names found in Sharoff's dictionary (n.d.) is 247/mln, while the net frequency of words *дерево, деревцо* 'tree, also dimin.' is 233/mln, which is remarkably close. In the Web supplement we present data for nouns, adjectives and verbs from different frequency ranges: *tree, flower, berry, meat, human, fish, fence; old, red, big/large, small, good, bad; say, think, rise and grow, shout and cry*. Some apparent counterexamples are also considered there. Here we focus on just one example, the most interesting from the methodological point of view.

Table A1 lists the hyponym frequencies for the word *плохой* 'bad,' collected from dictionaries. We omitted the four most frequent ones (*худой* 'skinny; leaky; bad,' *низкий* 'low, short; base, mean,' *дешевый* 'cheap, worthless,' *жалкий* 'pitiful; wretched'), because each of them has a primary meaning that does not directly imply badness. Something or somebody can be cheap and good, skinny and good, etc. But even without them, the net frequency of hyponyms is significantly greater than the head word frequency.

Notice though that the hyponyms can be roughly classified into two categories: those denoting more of an objective quality of an object, like *скверный* (cf. Eng. *poor* in its senses unrelated to pitying and lack of wealth), and those denoting more of a subjective feeling towards the object, like *мерзкий* 'loathsome, vile.' The head word itself falls more in the

Table A1
Translations of the word *bad*

Word	Freq./mln	Word	Freq./mln	Weight	Quality
плохой 'bad, poor'	102.22	дурной 'bad, mean'	40.40	0.911	+
		противный 'repugnant'	28.34	−0.0584	
		отвратительный 'disgusting'	21.85	−0.439	
		нехороший 'not good'	20.14	0.914	+
		мерзкий 'vile'	13.22	−1.946	
		скверный 'bad, poor'	13.16	0.896	+
		гнусный 'abominable'	12.73	−3.160	
		поганый 'foul'	11.51	−0.330	
		паршивый 'nasty'	10.16	−0.407	
		кошмарный 'nightmarish'	9.30	−0.180	
		негативный 'negative'	7.10	−0.183	
		неважный 'rather bad'	6.00	1.200	+
		омерзительный 'disgusting'	6.00	−0.432	
		гадкий 'repulsive; nasty'	5.33	−0.490	
		хреновый 'bad, poor (colloq.)'	5.14	2.358	+
		никчемный 'worthless'	5.08	0.144	+
		негодный 'worthless'	4.10	0.157	+
		дрянной 'rotten, trashy'	3.92	−0.110	
		никудышный 'worthless'	3.37	3.095	+
		захудалый 'run-down'	2.57	0.347	+
		неприглядный 'unsightly'	2.39	—	
		незавидный 'unenviable'	1.90	−0.161	
		дерьмовый 'shitty'	1.90	0.161	+
		фиговый 'bad, poor (colloq.)'	1.78	0.545	+
		неудовлетворительный 'unsatisfactory'	1.65	−0.077	
		паскудный 'foul, filthy'	1.59	−0.203	
		отвратный 'disgusting'	1.41	−0.165	
		грошовый 'dirt-cheap'	1.35	−0.172	
		бросовый 'worthless, trashy'	1.35	—	
		пакостный 'foul, mean'	1.35	−0.234	
		одиозный 'odious'	1.35	−0.122	
		сволочной 'mean, vile'	1.04	−0.318	
		аховый 'rotten'	0	−0.109	
		дефектный 'defective'	0	−0.179	
		завалящий 'worthless'	0	−0.078	
		мерзостный 'disgusting'	0	−0.270	
		мерзопакостный 'disgusting'	0	−0.302	
		низкопробный 'low-grade'	0	−0.406	
		отталкивающий 'revolting'	0	−0.198	
Sum	**102.22**		248.48		**103.64**

Note. Some translation are very approximate.

former category. To demonstrate this, consider the expression *плохой вор* 'a bad thief.' Its meaning is 'one who is not good at the art of stealing,' in contrast to *мерзкий вор* 'vile thief' = 'one whom I loathe because he steals,' Hence, only the frequencies of the hyponyms from the first category (denoting quality) should sum up to the frequency of the head word.

But it is quite difficult to actually classify the words into these two categories. The "subjective" words tend to evolve toward emphatic terms, and further migrate to the "objective" group or close to it. So we need a method that would allow us to perform classification without relying on dubious judgements based on the linguistic intuition. To this end, notice that there exist three classes of nouns by their compatibility with the adjectives from Table A1. Neutral nouns, like *погода* 'weather' can be equally easy found in noun phrases with both *скверный* 'bad, poor' and *мерзкий* '≈disgusting.' However the nouns carrying distinct negative connotation, such as *предатель* 'traitor' are well compatible with *мерзкий* '≈disgusting,' but not with *скверный* 'bad, poor.' On the contrary, nouns with distinct positive connotation have the opposite preference: cf. *скверный поэт* 'bad poet' and ?**мерзкий поэт* 'disgusting poet.' It is possible to find out which of the adjectives in Table A1 tend to apply preferentially to positive or negative nouns, by using an Internet search engine.

We considered eight test nouns: negative *гадость* '≈filth,' *дрянь* '≈trash,' *предатель* 'traitor,' *предательство* 'treason' and positive *здоровье* 'health,' *врач* 'doctor,' *поэт* 'poet,' *актер* 'actor.' They were initially selected for maximum contrast in their compatibility with adjectives *скверный* and *мерзкий*. Then we used Russian-specific search engine Yandex (http://www.yandex.ru) to determine the frequencies of noun phrases constructed from each of the adjectives with each of the nouns.

It should be noted here that search engines can't be directly used as replacements for a frequency dictionary. First, they typically report the number of "pages" and "sites," but not the number of word instances. Meanwhile, web pages can be of very different size, and may contain multiple instances of a word or search phrase. Second, search engines trim the results to exclude "similar pages" and avoid duplicates, i.e., texts available in multiple copies or from multiple addresses. It's not clear whether this is correct behavior from the point of view of calculating frequencies. Finally, the corpus with which search engines work, the whole of the Web, is by no means well-balanced according to the criteria of frequency dictionary compilers. So the results from search engines can't be directly compared with the data from frequency dictionaries. But for our purposes we need only relative figures, and we are interested in their qualitative behavior only. The effect we are looking for, if it exists, should be robust enough to withstand the inevitable distortion.

The frequencies of noun phrases constructed from each of the adjectives a_i with each of the test nouns n_j form a matrix N_{ij} presented in Table A2. One can readily see that the rows "мерзкий" and "скверный" clearly separate the test nouns into two groups preferentially compatible with one or the other. Many other rows of the table (e.g., "гнусный" and "неважный") behave in the same way. But there are rows that do not, and that is precisely the reason to consider multiple test words. Thus the adjective *негодный* '≈worthless' is entirely compatible with all the positive test nouns, but also with the negative test noun *дрянь* '≈trash.' The adjectives *неприглядный* 'unsightly' and *бросовый* 'worthless,' as it turns out, are not compatible with any of them, so they are excluded from further analysis. Their low frequency can't appreciably change the result anyway.

Table A2
Compatibility of the hyponyms of *плохой* 'bad, poor' with test nouns on the Web ("the number of pages")

Word	дрянь	гадость	предатель	предательство	здоровье	врач	актер	поэт
дурной	0	1	0	0	81	66	187	172
противный	140	333	45	0	0	61	35	31
отвратительный	305	5,187	0	112	71	45	211	43
нехороший	15	38	11	19	24	282	4	11
мерзкий	627	1,354	849	316	7	42	30	16
скверный	4	3	15	3	232	54	250	137
гнусный	156	62	1,380	1,934	2	1	0	3
поганый	183	27	97	33	87	14	17	32
паршивый	493	32	156	12	27	8	314	38
кошмарный	26	39	0	4	0	7	8	1
негативный	2	0	0	0	0	0	0	0
неважный	0	0	0	0	1,589	22	62	141
омерзительный	149	183	10	80	0	5	9	0
гадкий	132	257	140	28	12	3	15	3
хреновый	1	0	0	0	226	166	431	381
никчемный	58	0	7	0	33	23	38	81
негодный	63	0	6	0	108	39	32	58
дрянной	114	2	0	0	6	1	18	62
никудышный	13	0	0	0	136	146	989	409
захудалый	0	1	0	0	0	8	22	150
неприглядный	0	0	0	0	0	0	0	0
незавидный	0	0	0	0	39	0	0	0
дерьмовый	8	1	51	1	13	14	89	70
фиговый	0	0	0	0	11	33	53	167
неудовлетворительный	0	0	0	0	212	0	0	0
паскудный	4	4	18	4	0	1	0	0
отвратный	28	92	0	0	2	10	12	1
грошовый	0	0	0	0	0	0	6	0
бросовый	0	0	0	0	0	0	0	0
пакостный	34	22	4	3	0	0	0	0
одиозный	0	0	7	0	0	0	28	8
сволочной	99	21	18	0	16	2	0	0
аховый	0	0	0	0	3	0	3	21
дефектный	0	0	0	0	3	0	0	0
завалящий	1	0	0	0	0	26	2	0
мерзостный	41	36	36	0	0	1	0	0
мерзопакостный	96	84	4	2	2	4	0	1
низкопробный	173	29	0	2	0	0	6	1
отталкивающий	0	3	19	0	0	0	1	0
Eigenvector	−0.300	−0.110	−0.418	−0.377	0.206	0.355	0.416	0.489

To recap, we want to classify the rows of Table A2 by whether each row is more similar to the row "скверный" (quality of the object) or to the row "мерзкий" (speaker's attitude towards the object). This can be done via a statistical procedure known as Principal component analysis (PCA) or Singular value decomposition (SVD), which has found many different uses in statistical NLP in the recent years (e.g., Pado and Lapata (2007)).

First, each row of Table A2 was normalized by subtracting the average and dividing by the standard deviation. As a result, the rows "мерзкий" and "скверный" become almost opposite to each other with opposite signs: positive on positive test nouns and negative on negative ones, or vice versa. Then, the correlation matrix of the table's columns was calculated (size 8×8) and its first eigenvector n_j^1. Finally, the eigenvector's scalar products with the i-th row of the table yields the weight of the corresponding adjective $a_i^1 = \sum_j n_j^1 N_{ij}$.

Mathematically, the result of this procedure is that the product $a_i^1 n_j^1$ provides the best (in terms of mean square) approximation of this kind to the matrix N_{ij}. In other words, each row of the normalized table A2 is approximately proportional to the pattern row n_j^1 multiplied by the weight a_i^1. The pattern row is given at the bottom of Table A2. As expected, it correctly classifies test nouns as positive and negative. This means that they actually behave in opposite ways relative to the adjectives of interest. Now we can classify all the adjectives with positive weights $a_i^1 > 0$ as proper hyponyms of the word *плохой* 'bad, poor.' The weights are shown in table A1 (in an arbitrary normalization). The table shows that the net frequency of these proper hyponyms is very close to the frequency of the head word.

So it can be seen that the frequency hypothesis is confirmed here as well, and this conclusion is not based on any intuitive judgement about word semantics.

To summarize, we demonstrated on several examples that the hypothesis of word frequency being proportional to the extent of its meaning is supported by available data. Of course, a much more thorough and systematic investigation is in order until the hypothesis can be considered proven. We only sketched some promising approaches to such an investigation. But it also should be noted that the examples considered span a wide range of word frequencies, include all three main parts of speech, and involve very common words, not specially hand-picked ones.