# Networks Generated from Natural Language Text

Chris Biemann and Uwe Quasthoff

Institute for Computer Science, NLP Department, University of Leipzig, Johannisgasse 26, 04103 Leipzig, Germany; biem@informatik.uni-lepzig.de, quasthoff@informatik.uni-leipzig.de

## 1 Introduction

The study of large-scale characteristics of graphs that arise in natural language processing is an essential step in finding structural regularities. Structure discovery processes have to be designed with an awareness of these properties. Examining and contrasting the effects of processes that generate graph structures similar to those observed in language data sheds light on the structure of language and its evolution.

In this chapter, we examine power-law distributions and small world graphs (SWGs) originating from natural language data. There are several reasons for the special interest in these structures.

1. Power laws appear in many rank-frequency statistics. Furthermore, we can construct graphs with words as nodes and use various rules to introduce edges between words. In many cases, this results in SWGs, which again often have a power-law distribution for their node degrees.
2. SWGs appear in many other real world data, like social networks of many kinds, in the link structure of the World Wide Web or in traffic networks. It is interesting to analyze all these networks in more detail to identify similarities and differences.
3. From an application-driven view, SWGs allow effective clustering strategies in nearly linear time. Because these clusters are often related to the growth process of the underlying graph, they are often meaningful. In the case of natural language these clusters usually reflect semantic and/or syntactic structures.

After discussing several data sources that exhibit power-law distributions with respect to rank frequency in Section 2, graphs with small world properties in language data are discussed in Section 3. We shall see that these characteristics are omnipresent in language data, and we should be aware of them when designing structure discovery processes. For example, the knowledge that a

N. Ganguly et al. (eds.), *Dynamics On and Of Complex Networks*,
Modeling and Simulation in Science, Engineering and Technology,
DOI: 10.1007/978-0-8176-4751-3_10,
© Birkhäuser Boston, a part of Springer Science+Business Media, LLC 2009

few hundred words make the bulk of words in a text allows one to use only these words as contextual features with only a minor loss in text coverage. Knowing that word co-occurrence networks possess the scale-free small world property has implications for clustering these networks.

An interesting aspect is whether these characteristics are only inherent to real natural language data or whether they can be produced with generators of linear sequences in a much simpler way than our intuition about language complexity would suggest. In other words, we shall see how distinctive these characteristics are with respect to tests deciding whether a given sequence is natural language or not.

## 2 Power Laws in Rank-Frequency Distribution

G. K. Zipf [31, 32] described the following phenomenon: if all words in a corpus of natural language are arranged in decreasing order of frequency, then the relation between a word's frequency and its rank in the list follows a power law. Since then, a significant amount of research has been devoted to the question of how this property emerges and what kinds of processes generate such Zipfian distributions. Hence, some datasets related to language will be presented that exhibit a power law on their rank-frequency distribution. For this discussion, basic units of language will be examined.

### 2.1 Word Frequency

The relation between the frequency of a word at rank $r$ and its rank is given by $f(r) \sim r^{-z}$, where $z$ is the exponent of the power law that corresponds to the slope of the curve in a log-log plot. The exponent $z$ was assumed to be exactly 1 by Zipf. In natural language data, slightly differing exponents in the range of about 0.7 to 1.2 are also observed [30]. B. Mandelbrot [21] provided a formula that more closely approximates the frequency distributions in language data after noticing that Zipf's law holds only for the medium range of ranks, whereas the curve is flatter for very frequent words and steeper for high ranks. Figure 1 displays the word rank-frequency distributions of corpora of different languages taken from the Leipzig Corpora Collection.[1]

There exist several exhaustive collections of research capitalising Zipf's law and related distributions[2] ranging over a wide area of datasets; here, only findings related to natural language will be reported. A related distribution is the *lexical spectrum* [16], which gives the probability of choosing a word from the vocabulary with a given frequency. For natural language, the lexical spectrum follows a power law with slope $\gamma = \frac{1}{z} + 1$, where $z$ is the exponent

[1] LCC, see `http://www.corpora.uni-leipzig.de` [July 7th, 2007].

[2] e.g. `http://www.nslij-genetics.org/wli/zipf/index.html` [April 1, 2007].

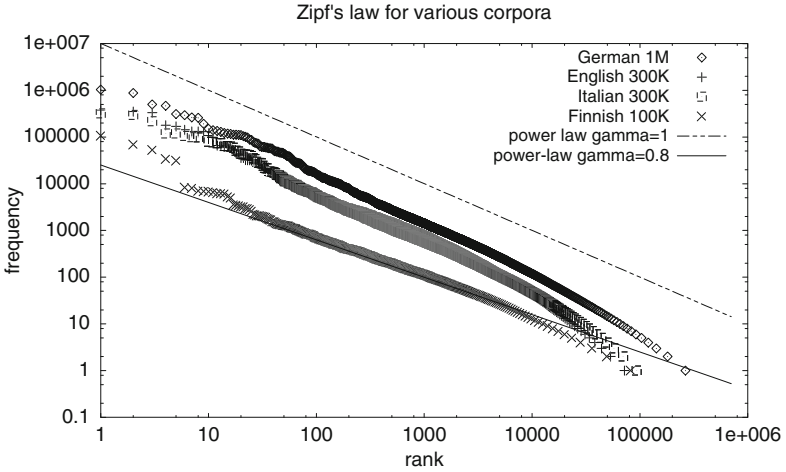


**Fig. 1.** Zipf's law for various corpora. The numbers next to the language give the corpus size in sentences. Enlarging the corpus does not affect the slope of the curve, but merely moves it upwards in the plot. Most lines are almost parallel to the ideal power-law curve with $z = 1$. Finnish exhibits a lower slope of $\gamma \approx 0.8$, akin to higher morphological productivity.

of the Zipfian rank-frequency distribution. For the relation between lexical spectrum, Zipf's law and Pareto's law, see [1].

But Zipf's law in its original form is just the tip of the iceberg of power-law distributions in a quantitative description of language. While a Zipfian distribution for word frequencies can be obtained by a simple model of generating letter sequences with space characters as word boundaries [21, 22], these models based on "intermittent silence" can neither reproduce the distributions on sentence length [26] nor explain the relations of words in sequence. Next, more power-law distributions in natural language are discussed and exemplified.

### 2.2 Letter $N$-Grams

To continue with a counter example, letter frequencies do not obey a power law in the rank-frequency distribution. This also holds for letter $N$-grams (including the space character), yet for higher $N$, the rank-frequency plots show a large power-law regime with exponential tails for high ranks. Figure 2 shows the rank-frequency plots for letter $N$-grams up to $N = 6$ for the first 10,000 sentences of the British National Corpus (BNC,[3] [10]).

Still, letter frequency distributions can be used to show that letters are not forming letter bigrams from the single letters independently, but there are restrictions on their combination. While this intuitively seems obvious for

[3] `http://www.natcorp.ox.ac.uk/` [April 1, 2007]

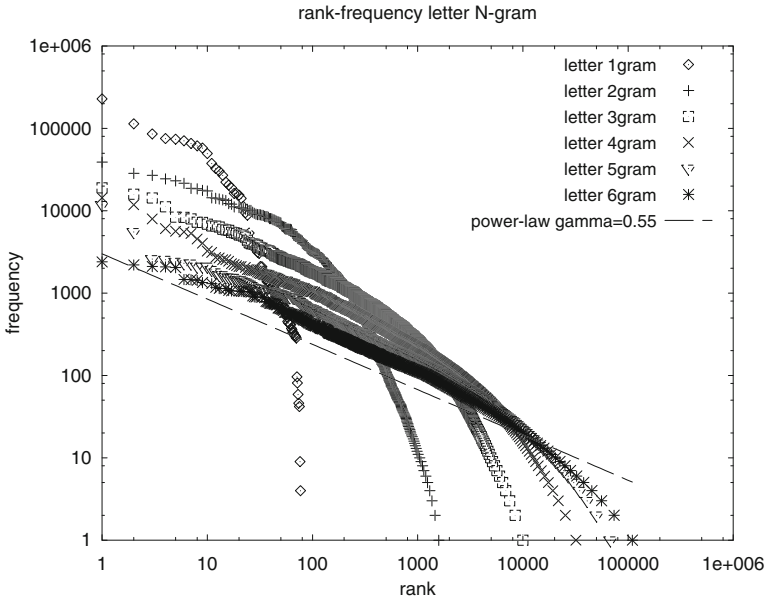


**Fig. 2.** Rank-frequency distributions for letter $N$-grams for the first 10,000 sentences in the BNC. Letter $N$-gram rank-frequency distributions do not exhibit power laws on the full scale, but increasing $N$ results in a larger power-law regime for low ranks.

letter combination, the following test is proposed for quantitatively examining the effects of these restrictions: from letter unigram probabilities, a text is generated that follows the letter unigram distribution by randomly and independently drawing letters according to their distribution and concatenating them. The letter bigram frequency distribution of this generated text can be compared to the letter bigram frequency distribution of the real text from where the unigram distribution was measured. Figure 3 shows the generated plot and the real rank-frequency plot, again from the small BNC sample.

The two curves clearly differ. The generated bigrams without restrictions predict a higher number of different bigrams and lower frequencies for bigrams of high ranks as compared to the real text bigram statistics. This shows that letter combination restrictions do exist, as not all bigrams predicted by the generation process were observed, resulting in higher counts for valid bigrams in the sample.

### 2.3 Word *N*-Grams

For word $N$-grams, the relation between rank and frequency follows a power law, just as in the case for words (unigrams). Figure 4 (left) shows the rank-frequency plots up to $N = 4$, based on the first 1 million sentences of the BNC. As more different word combinations are possible with increasing $N$,

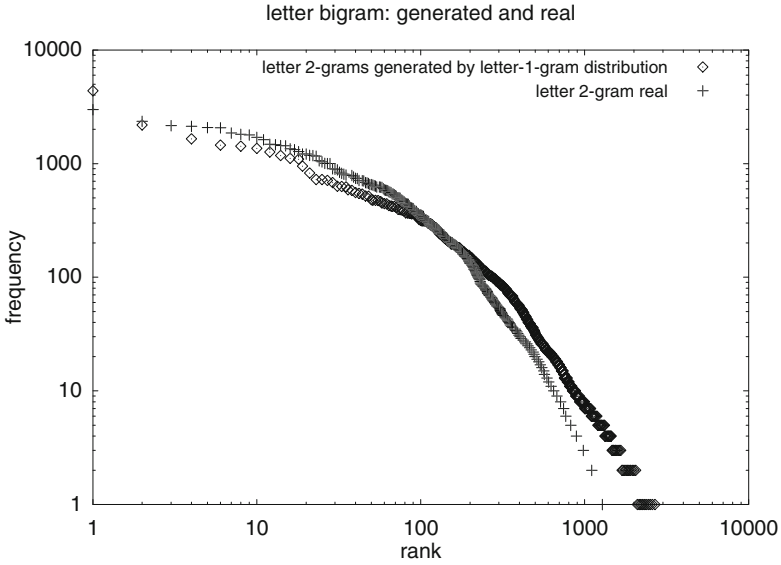


**Fig. 3.** Rank-frequency plots for letter bigrams, for a text generated from letter unigram probabilities and for the BNC sample.

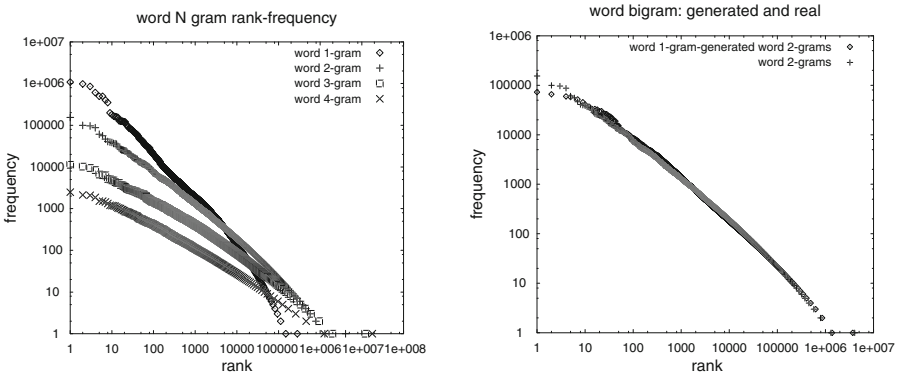


**Fig. 4.** Left: Rank-frequency distributions for word $N$-grams for the first one million sentences in the BNC. Word $N$-gram rank-frequency distributions exhibit power laws. Right: Rank-frequency plots for word bigrams, for a text generated from letter unigram probabilities and for the BNC sample.

the curves become flatter as the same total frequency is shared amongst more units, as previously observed (e.g. [27, 18]). Testing concatenation restrictions quantitatively as above for letters, it might at first seem surprising that the curve for a text generated with word unigram frequencies differs only very little from the word bigram curve, as Fig. 4 (right) shows. Small differences are only observable for low ranks: more top-rank generated bigrams reflect

that words are usually not repeated in the text. More low-ranked and less high-ranked real bigrams indicate that word concatenation takes place not entirely without restrictions, yet is subject to much more variety than letter concatenation. This coincides with the intuition that it is, for a given word pair, almost always possible to form a correct English sentence in which these words are neighbours. Regarding quantitative (as opposed to syntactic or semantic) aspects, the frequency distribution of word bigrams can be produced by a generation process based on word unigram probabilities.

### 2.4 Sentence Frequency

Larger corpora that are compiled from a variety of sources contain a considerable amount of duplicate sentences. In the full BNC, which serves as the data basis in this case, 7.3% of the sentences occur two or more times. The most frequent sentences are "Yeah.", "Mm.", "Yes." and "No.", which are mostly found in the section of spoken language. But also longer expressions like "Our next bulletin is at 10.30 p.m." have a count of over 250. The sentence frequencies also follow a power law with an exponent close to 1 (see Fig. 5), indicating that Zipf's law also holds for sentence frequencies.

### 2.5 Other Power Laws in Language Data

The preceding results strongly suggest that when counting document frequencies in large collections such as the World Wide Web, another power-law

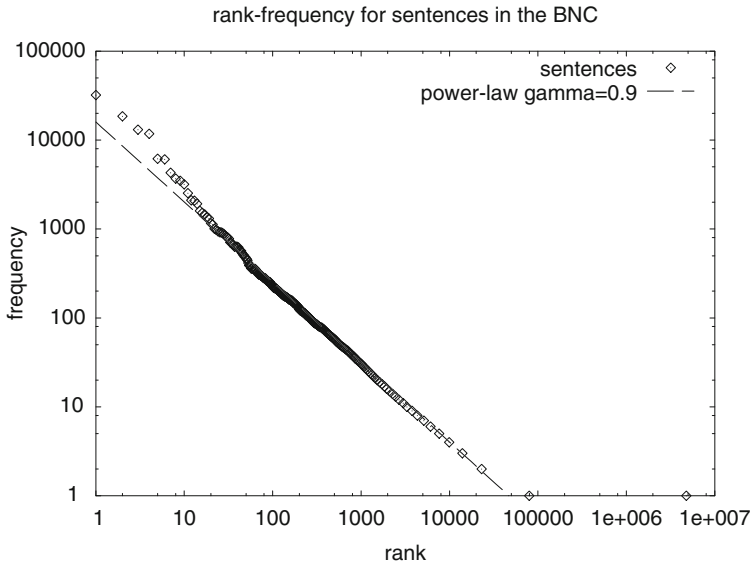


**Fig. 5.** Rank-frequency plot for sentence frequencies in the full BNC, following a power law with $\gamma \approx 0.9$, but with a high fraction of sentences occurring only once.

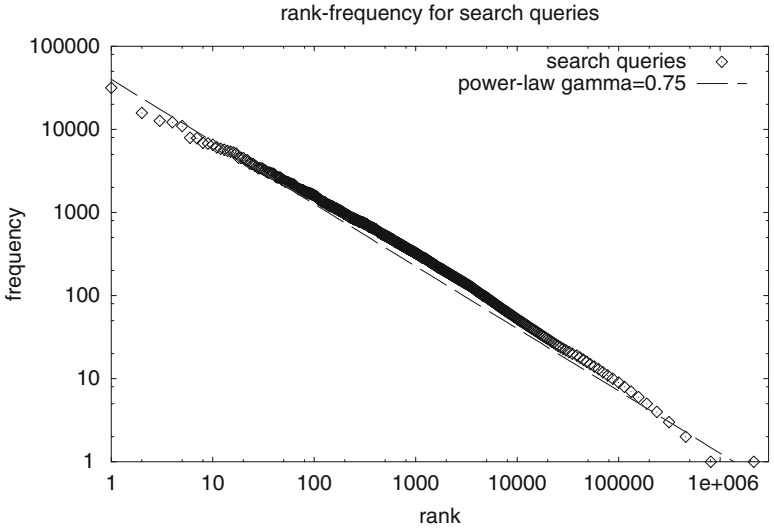


**Fig. 6.** Rank-frequency plot for AltaVista search queries, following a power law with $\gamma \approx 0.75$.

distribution would be found, but such an analysis has not been carried out and would require access to the index of a web search engine. Further, there are more power laws in language-related areas, some are mentioned here briefly to illustrate their omnipresence.

- Web page requests follow a power law, which was employed for a caching mechanism in [17].
- Related to this, frequencies of web search queries during a fixed time span also follow a power law, as exemplified in Fig. 6 for a 7-million queries log of AltaVista[4] as used by Lempel [19].
- The number of authors of Wikipedia[5] articles was found to follow a power law with $\gamma \approx 2.7$ for a large regime in [29]. The same paper further discusses various other power-law relationships.

## 3 Scale-Free Small Worlds in Language Data

The previous section discussed the shape of rank-frequency distributions for natural language units. Now the properties of graphs with units represented as vertices and relations between them as edges will be the focus of interest. Internal as well as contextual features can be employed for computing similarities between language units that are represented as (possibly weighted) edges

[4] `http://www.altavista.com`
[5] `http://www.wikipedia.org`

in the graph. Some of the graphs discussed here can be classified as scale-free SWGs; others have different characteristics and represent other, but related, graph classes.

### 3.1 Word Co-Occurrence Graph

The notion of *word co-occurrence* is used to model dependencies between words. If two words X and Y occur together in some contextual unit of information (as neighbours, in a word window of 5, in a clause, in a sentence, in a paragraph), they are said to co-occur. When regarding words as vertices and edge weights as the number of times two words co-occur, the *word co-occurrence graph* of a corpus is given by the entirety of all word co-occurrences. In the following, two specific types of co-occurrence graphs are considered: the graph as induced by neighbouring words, henceforth called the neighbour-based graph, and the graph as induced by sentence-based co-occurrence, henceforth called the sentence-based graph. The neighbour-based graph can be undirected or directed with edges going from the left to the right words as found in the corpus, the sentence-based graph is undirected.

To find out whether the co-occurrence of two specific words A and B is merely due to chance or exhibits a statistical dependency, measures are used that compute, to what extent the co-occurrence of A and B is statistically significant. Many significance measures can be found in the literature; for extensive overviews consult e.g. [9] or [14]. In general, the measures compare the probability for A and B to co-occur under the assumption of their statistical independence with the actual probability of their joint co-occurrence in the corpus. In this work, the log likelihood ratio [13] is used to sort the wheat from the chaff. It is given in expanded form in [9]:

$$-2\log\lambda = 2\left[\begin{array}{l} n\log n - n_A\log n_A - n_B\log n_B + n_{AB}\log n_{AB} \\ +(n - n_A - n_B + n_{AB})\log(n - n_A - n_B + n_{AB}) \\ +(n_A - n_{AB})\log(n_A - n_{AB}) + (n_B - n_{AB})\log(n_B - n_{AB}) \\ -(n - n_A)\log(n - n_A) - (n - n_B)\log(n - n_B) \end{array}\right],$$

where $n$ is the total number of contexts, $n_A$ the frequency of A, $n_B$ the frequency of B and $n_{AB}$ the number of co-occurrences of A and B. As pointed out by Moore [23], this formula overestimates the co-occurrence significance for small $n_{AB}$. For this reason, often a frequency threshold $t$ on $n_{AB}$ (e.g. a minimum of $n_{AB} = 2$) is applied. Further, a significance threshold $s$ regulates the density of the graph; for the log likelihood ratio, the significance values correspond to the $\chi^2$ tail probabilities [23], which makes it possible to translate the significance value into an error rate for rejecting the independence assumption.[6] The operation of applying a significance test results in pruning edges

[6]For example, a log likelihood ratio of 3.84 corresponds to a 5% error in stating that two words do not occur by chance, a significance of 6.63 corresponds to a 1% error.

that exist due to random noise and keeping almost exclusively those edges that reflect a true association between their endpoints. Graphs that contain all significant co-occurrences of a corpus, with edge weights set to the significance value between their endpoints, are called *significant co-occurrence graphs* in the remainder. For convenience, no singletons in the graph are allowed, i.e. if a vertex is not contained in any edge because none of the co-occurrences for the corresponding word is significant, then the vertex is excluded from the graph.

As observed previously [15, 24], word co-occurrence graphs exhibit the scale-free small world property. This is in line with co-occurrence graphs reflecting human associations [25] and human associations in turn forming SWGs [28]. The claim is confirmed here on an exemplary basis with the graph for Leipziy Corpora Collection's (LCC's) 1 million sentence corpus for German. Figure 7 gives the degree distributions and graph characteristics for various co-occurrence graphs.

The shape of the distribution is dependent on the language, as Fig. 8 shows. Some languages—here English and Italian—have a hump-shaped distribution in the log-log plot where the first regime follows a power law with a lower exponent than the second regime, as observed in [15]. For the Finnish and German corpora examined here, this effect could not be found in the data. This property of two power-law regimes in the degree distribution of word co-occurrence graphs motivated the Dorogovtsev-Mendes (DM)-model, see [12]. There, the

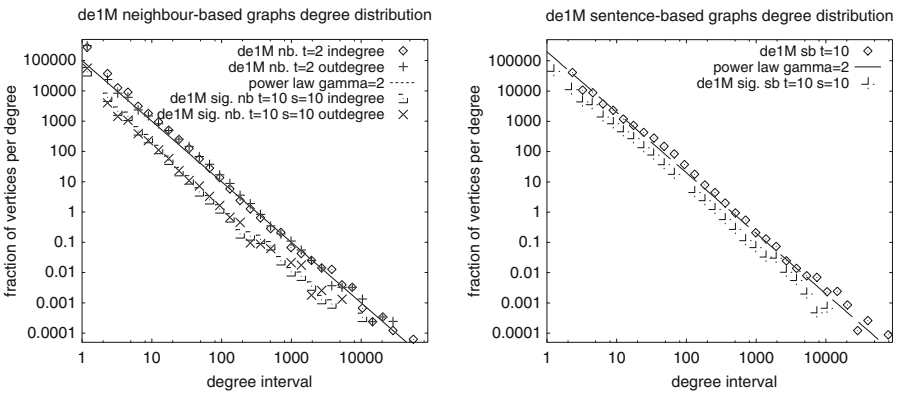


**Fig. 7.** Graph characteristics for various co-occurrence graphs of LCC's 1-million sentence German corpus. Abbreviations: nb = neighbour-based, sb = sentence-based, sig. = significant, $t$ = co-occurrence frequency threshold, $s$ = co-occurrence significance threshold. While the exact shapes of the distributions are language and corpus dependent, the overall characteristics are valid for all samples of natural language of sufficient size. The slope of the distribution is invariant to changes of thresholds. Characteristic path length and a high clustering coefficient at low average degrees are characteristic for SWGs.

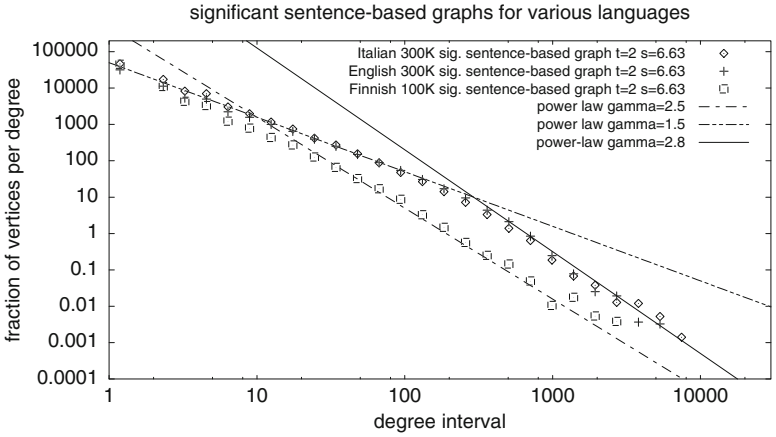


**Fig. 8.** Degree distribution of significant sentence-based co-occurrence graphs of similar thresholds for Italian, English and Finnish.

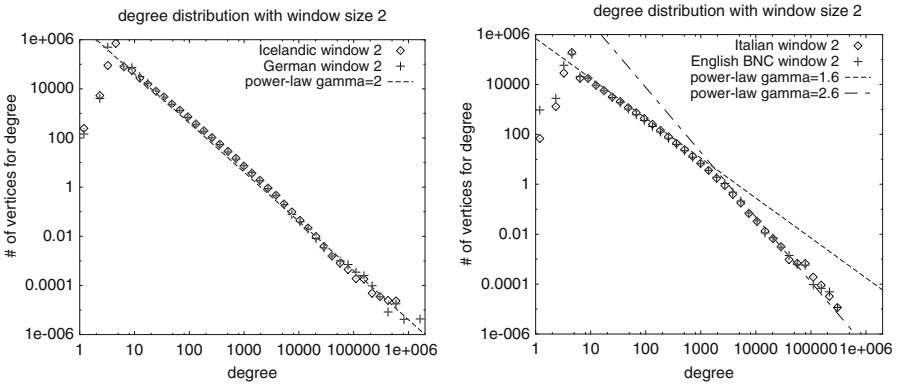


**Fig. 9.** Degree distributions in word co-occurrence graphs for window size 2. Left: The distribution for German and Icelandic is approximated by a power law with $\gamma = 2$. Right: For English (BNC) and Italian, the distribution is approximated by two power-law regimes.

crossover point of the two power-law regimes is motivated by a *kernel lexicon* of about 5000 words that can be combined with all words of a language.

The original experiments of [15] operated on a word co-occurrence graph with window size 2: an edge is drawn between words if they appear together at least once in a distance of one or two words in the corpus. Reproducing their experiment with the first 70 million words of the BNC and corpora of German, Icelandic and Italian of similar size reveals that the degree distribution of the English and the Italian graph is in fact approximated by two power-law regimes. In contrast to this, German and Icelandic show a single power-law distribution, just as in the experiments above; see Fig. 9. These results suggest

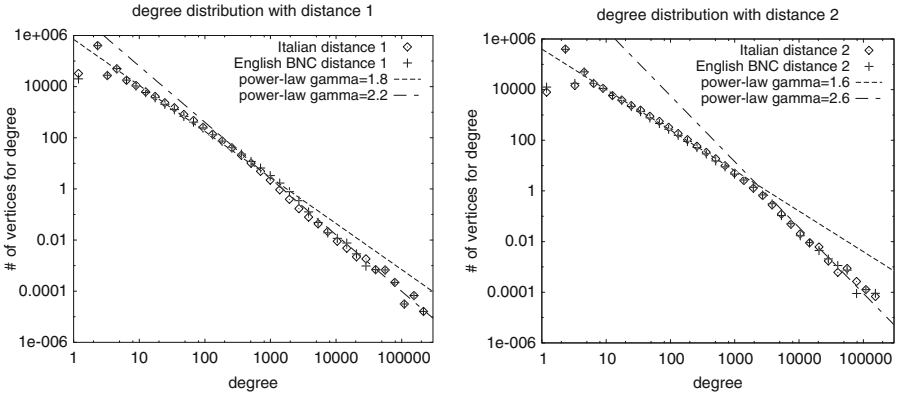


**Fig. 10.** Degree distributions in word co-occurrence graphs for distance 1 and distance 2 for English (BNC) and Italian. The hump-shaped distribution is much more distinctive for distance 2.

that two power-law regimes in word co-occurrence graphs with window size 2 are not a language universal, but only hold for some languages.

To examine the hump-shaped distributions further, Fig. 10 displays the degree distribution for the neighbour-based word co-occurrence graphs and the word co-occurrence graphs for connecting only words that appear in a distance of 2. As it becomes clear from the plots, the hump-shaped distribution is mainly caused by words co-occurring in distance 2, whereas the neighbour-based graph shows only a slight deviation from a single power law. Together with the observations from sentence-based co-occurrence graphs of different languages in Figure 8, it becomes clear that a hump-shaped distribution with two power-law regimes is caused by long-distance relationships between words, if present at all.

### 3.1.1 Applications of Word Co-Occurrences

Word co-occurrence statistics are an established standard and have been used in many language processing systems. The authors have used co-occurrences in practical applications like bilingual dictionary acquisition [4, 11], semantic lexicon extension [8] and visualisation of concept trails [7]. The aim of this chapter is to underpin their applications with a theoretical foundation.

## 3.2 Co-Occurrence Graphs of Higher Order

The significant word co-occurrence graph of a corpus represents words that are likely to appear near to each other. When one is interested in words co-occurring with similar other words, it is possible to transform the above-defined (first-order) co-occurrence graph into a second-order co-occurrence graph by drawing an edge between two words A and B if they share a common

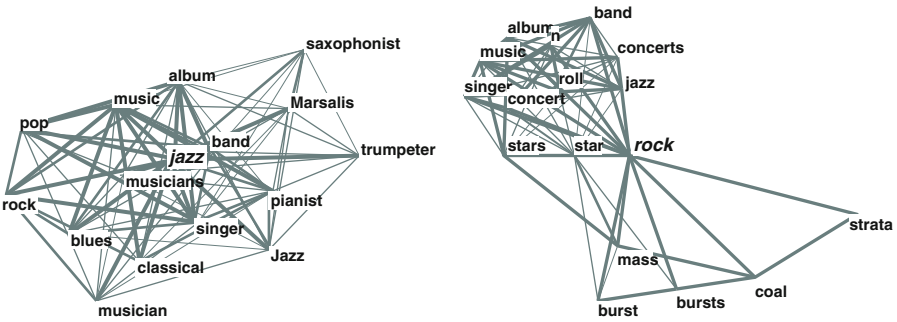


**Fig. 11.** Neighbourhoods of *jazz* and *rock* in the significant sentence-based word co-occurrence graph as displayed on LCC's English corpus website. Both neighbourhoods contain *album*, *music*, *singer* and *band*, which leads to an edge weight of 4 in the second-order graph.

neighbour in the first-order graph. Whereas the first-order word co-occurrence graph represents the global context per word, the corresponding second-order graph contains relations between words which have similar global contexts. The edge can be weighted according to the number of common neighbours, e.g. by $weight = |neigh(A) \cap neigh(B)|$. Figure 11 shows neighbourhoods of the significant sentence-based first-order word co-occurrence graph from LCC's English web corpus[7] for the words *jazz* and *rock*. Taking into account only the data depicted, *jazz* and *rock* are connected with an edge of weight 4 in the second-order graph, corresponding to their common neighbours *album*, *music*, *singer* and *band*. The fact that they share an edge in the first-order graph is ignored.

In general, a graph of order $N+1$ can be obtained from the graph of order $N$, using the same transformation. The higher-order transformation without thresholding is equivalent to a multiplication of the unweighted adjacency matrix $A$ with itself, then a zeroing of the main diagonal by subtracting the degree matrix of $A$. Since the average path length of scale-free SWGs is short and local clustering is high, this operation leads to an almost fully connected graph in the limit, which does not allow one to draw conclusions about the initial structure. Thus, the graph is pruned in every iteration $N$ in the following way. For each vertex, only the $max_N$ outgoing edges with the highest weights are taken into account. Notice that this vertex degree threshold $max_N$ does not limit the maximum degree, as thresholding is asymmetric. This operation is equivalent to only keeping the $max_N$ largest entries per row in the adjacency matrix $A = (a_{ij})$, then $A_t = (sign(a_{ij} + a_{ji}))$, resulting in an undirected graph. To examine quantitative effects of the higher-order transformation, the sentence-based word co-occurrence graph of LCC's 1-million German sentence corpus ($s = 6.63$, $t = 2$) underwent this operation. Figure 12 depicts the degree distributions for $N = 2$ and $N = 3$ for different $max_N$.

[7] `http://corpora.informatik.uni-leipzig.de/?dict=en` [April 1, 2007]

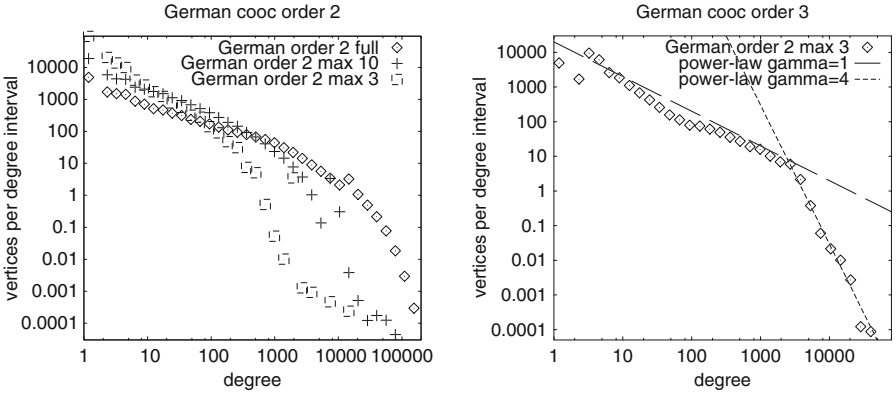


**Fig. 12.** Degree distributions of word-co-occurrence graphs of higher order. The first-order graph is the sentence-based word co-occurrence graph of LCC's 1-million German sentence corpus ($s = 6.63$, $t = 2$). Left: $N = 2$ for $max_2 = 3$, $max_2 = 10$ and $max_2 = \infty$. Right: $N = 3$ for $t_2 = 3$, $t_3 = \infty$, using the second-order graph with $max_2 = 3$.

Applying the $max_N$ threshold causes the degree distribution to change, especially for high degrees. In the third-order graph, two power-law regimes are observable.

Studying the degree distribution of higher-order word co-occurrence graphs revealed that the characteristic of being governed by power laws is invariant to the higher-order transformation, yet the power-law exponent changes. This indicates that the power-law characteristic is inherent at many levels in natural language data. To examine what this transformation yields on the graphs generated by other random graph models, Figure 13 shows the degree distribution of second-order and third-order graphs as generated by the graph generation models of [3] (Barabási-Albert (BA)-model), [28] (Steyvers-Tenenbaum (ST)-model) and [12] (DM-model). The underlying first-order graphs are the undirected graphs of order 10,000 and size 50,000 ($\langle k \rangle$=10) from these three models.

While the thorough interpretation of second-order graphs of random graphs might be subject to further studies, the following should be noted: the higher-order transformation reduces the power-law exponent of the BA-model graph from $\gamma = 3$ to $\gamma = 2$ in the second order and to $\gamma \approx 0.7$ in the third order. For the ST-model, the degree distribution of the full second-order graph shows a maximum around $2M$, then decays with a power law with exponent $\gamma \approx 2.7$. In the third-order ST-graph, the maximum moves to around $4M^2$ for sufficient $max_2$. The DM-model second-order graph shows, like the first-order DM-model graph, two power-law regimes in the full version, and a power-law with $\gamma \approx 2$ for the pruned versions. The third-order degree distribution exhibits many more vertices with high degrees than predicted by a power law.

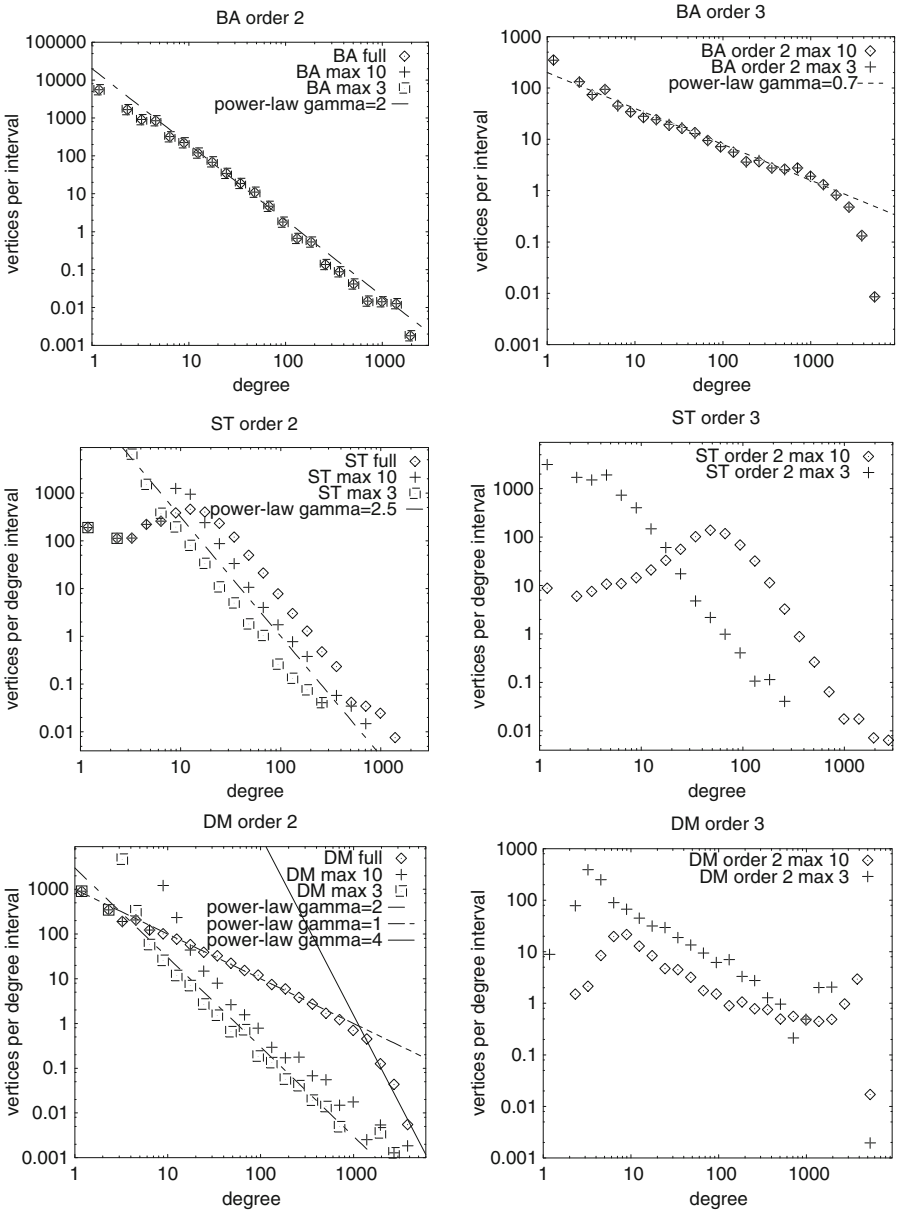


**Fig. 13.** Second- and third-order graph degree distributions for BA-model, ST-model and DM-model graphs.

In summary, all random graph models exhibit clear differences for word co-occurrence networks with respect to the higher-order transformation. The ST-model shows maxima depending on the average degree of the first-order graph. The BA-model's power law is decreased with higher orders, but is able to explain a degree distribution with power-law exponent 2. The full DM model exhibits the same two power-law regimes in the second order as observed for German sentence-based word co-occurrences in the third order.

### 3.2.1 Applications of Co-Occurrence Graphs of Higher Orders

In [6] and [20], the utility of word co-occurrence graphs of higher orders are examined for lexical semantic acquisition. The highest potential for extracting paradigmatic semantic relations can be attributed to second- and third-order word co-occurrences. In [9] second-order graphs are evaluated against lexical semantic resources.

## 3.3 Sentence Similarity

Using words as internal features, the similarity of two sentences can be measured by the number of common words they share. Since the few top frequency words are contained in most sentences as a consequence of Zipf's law, their influence should be downweighted or they should be excluded to arrive at a useful measure for sentence similarity. Here, the sentence similarity graph of sentences sharing at least two common words is examined, with the maximum frequency of these words bounded by 100. This maximum frequency threshold was arbitrarily chosen and could be replaced by a weighting scheme that attributes more weight to less frequent words. However, a hard threshold reduces the computational cost significantly. The corpus of examination is here LCC's 3-million sentences of German. Figure 14 shows the component size distribution for this sentence similarity graph, Figure 15 shows the degree distributions for the entire graph and for its largest component.

The degree distribution of the entire graph follows a power law with $\gamma$ close to 1 for low degrees and decays faster for high degrees; the largest component's degree distribution plot is flatter for low degrees. This can be attributed to limited sentence length: as sentences are not arbitrarily long, they cannot be similar to an arbitrary high number of other sentences with respect to the measure discussed here, as the number of sentences per feature word is bounded by the word frequency limit. However, the extremely high values for transitivity and clustering coefficient and the low $\gamma$ values for the degree distribution for low degree vertices and comparably long average shortest path lengths indicate that the sentence similarity graph belongs to a different graph class than all other graphs discussed above.

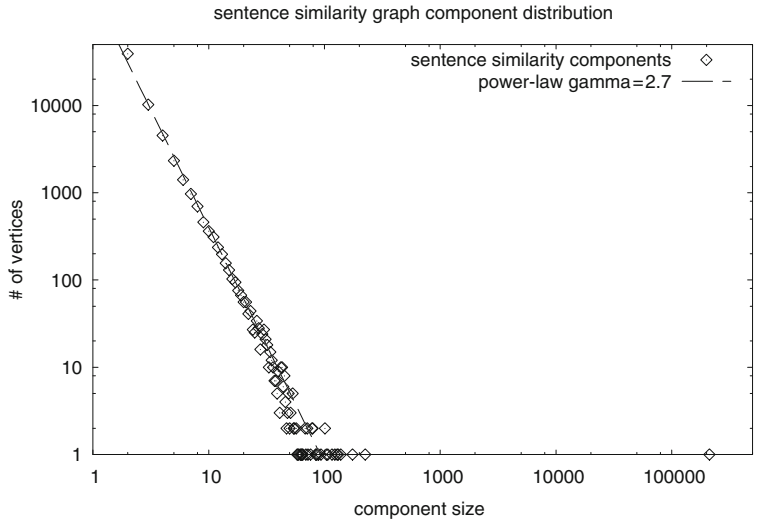


**Fig. 14.** Component size distribution for the sentence similarity graph of LCC's 3-million sentence German corpus. The component size distribution follows a power law with $\gamma \approx 2.7$ for small components, the largest component comprises 211,447 out of 416,922 total vertices. The component size distribution complies with the theoretical results of [2].

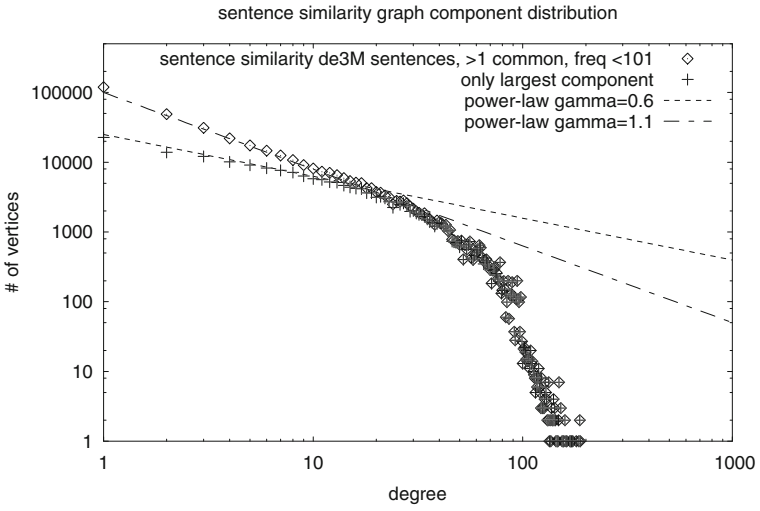


**Fig. 15.** Degree distribution for the sentence similarity graph of LCC's 3-million sentence German corpus and its largest component. An edge between two vertices representing sentences is drawn if the sentences share at least two words with corpus frequency <101; singletons are excluded.

### 3.3.1 Applications of the Sentence Similarity Graph

A similar measure is used in [5] for document similarity and obtains well-correlated results when evaluated against a given document classification. A precision-recall tradeoff arises when lowering the frequency threshold for feature words or increasing the minimum number of common feature words two documents must have in order to be connected in the graph: both improve precision but result in many singleton vertices, which lowers the total number of documents that are considered.

### 3.4 Summary of Scale-Free Small Worlds in Language Data

The preceding examples confirm the claim that graphs built on various aspects of natural language data often exhibit the scale-free small world property or similar characteristics. Experiments with generated text corpora suggest that this is mainly due to the power-law frequency distribution of language units. The slopes of the power law approximating the degree distributions can often not be produced using the random graph generation models. Specifically, all previously discussed generation models fail to explain the properties of word co-occurrence graphs, where $\gamma \approx 2$ was observed as the power-law exponent of the degree distribution. Of the generation models inspired by language data, the ST-model exhibits $\gamma = 3$, whereas the universality of the DM-model to capture word co-occurrence graph characteristics can be doubted after examining data from different languages.

## References

1. Adamic, L. A. (2000). Zipf, power-law, pareto – a ranking tutorial. Technical report, Information Dynamics Lab, HP Labs, HP Labs, Palo Alto, CA 94304.
2. Aiello, W., Chung, F., and Lu, L. (2000). A random graph model for massive graphs. In *STOC '00: Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, pages 171–180, New York, NY, USA. ACM Press.
3. Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**, 509.
4. Biemann, C. and Quasthoff, U. (2005). Dictionary acquisition using parallel text and co-occurrence statistics. In *Proceedings of NODALIDA'05*, Joensuu, Finland.
5. Biemann, C. and Quasthoff, U. (2007). Similarity of documents and document collections using attributes with low noise. In *Proceedings of the Third International Conference on Web Information Systems and Technologies (WEBIST-07)*, pages 130–135, Barcelona, Spain.
6. Biemann, C., Bordag, S., and Quasthoff, U. (2004a). Automatic acquisition of paradigmatic relations using iterated co-occurrences. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC-04)*, Lisbon, Portugal.

7. Biemann, C., Bhm, C., Heyer, G., and Melz, R. (2004b). Automatically building concept structures and displaying concept trails for the use in brainstorming sessions and content management systems. In *Proceedings of Innovative Internet Community Systems (IICS-2004)*, Springer LNCS, Guadalajara, Mexico.
8. Biemann, C., Shin, S.-I., and Choi, K.-S. (2004c). Semiautomatic extension of corenet using a bootstrapping mechanism on corpus-based co-occurrences. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING-04)*, Morristown, NJ, USA. Association for Computational Linguistics.
9. Bordag, S. (2007). *Elements of Knowledge-free and Unsupervised Lexical Acquisition*. Ph.D. thesis, University of Leipzig.
10. Burnard, L. (1995). *Users Reference Guide for the British National Corpus*. Oxford University Computing Service, Oxford, U.K.
11. Cysouw, M., Biemann, C., and Ongyerth, M. (2007). Using Strong's numbers in the Bible to test an automatic alignment of parallel texts. *Special issue of Sprachtypologie und Universalienforschung (STUF)*, pages 66–79.
12. Dorogovtsev, S. N. and Mendes, J. F. F. (2001). Language as an evolving word web. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, **268**(1485), 2603–2606.
13. Dunning, T. E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, **19**(1), 61–74.
14. Evert, S. (2004). *The Statistics of Word Co-occurrences: Word Pairs and Collocations*. Ph.D. thesis, University of Stuttgart.
15. Ferrer-i-Cancho, R. and Sol, R. V. (2001). The small world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, **268**(1482), 2261–2265.
16. Ferrer-i-Cancho, R. and Sol, R. V. (2002). Zipf's law and random texts. *Advances in Complex Systems*, **5**(1), 1–6.
17. Glassman, S. (1994). A caching relay for the world wide web. *Computer Networks and ISDN Systems*, **27**(2), 165–173.
18. Ha, L. Q., Sicilia-Garcia, E. I., Ming, J., and Smith, F. J. (2002). Extension of Zipf's law to words and phrases. In *Proceedings of the 19th International Conference on Computational Linguistics (COLING-02)*, pages 1–6, Morristown, NJ, USA. Association for Computational Linguistics.
19. Lempel, R. and Moran, S. (2003). Predictive caching and prefetching of query results in search engines. In *Proceedings of the 12th International Conference on World Wide Web (WWW-03)*, pages 19–28, New York, NY, USA. ACM Press.
20. Mahn, M. and Biemann, C. (2005). Tuning co-occurrences of higher orders for generating ontology extension candidates. In *Proceedings of the ICML-05 Workshop on Ontology Learning and Extension using Machine Learning Methods*, Bonn, Germany.
21. Mandelbrot, B. B. (1953). An information theory of the statistical structure of language. In *Proceedings of the Symposium on Applications of Communications Theory*. Butterworths.
22. Miller, G. A. (1957). Some effects of intermittent silence. *American Journal of Psychology*, **70**, 311–313.
23. Moore, R. C. (2004). On log-likelihood-ratios and the significance of rare events. In D. Lin and D. Wu, editors, *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, pages 333–340, Barcelona, Spain. Association for Computational Linguistics.

24. Quasthoff, U., Richter, M., and Biemann, C. (2006). Corpus portal for search in monolingual corpora. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC-06)*, pages 1799–1802, Genoa, Italy.
25. Rapp, R. (1996). *Die Berechnung von Assoziationen: ein korpuslinguistischer Ansatz*. Olms, Hildesheim.
26. Sigurd, B., Eeg-Olofsson, M., and van de Weijer, J. (2004). Word length, sentence length and frequency – Zipf revisited. *Studia Linguistica*, **58**(1), 37–52.
27. Smith, F. J. and Devine, K. (1985). Storing and retrieving word phrases. *Inf. Process. Manage.*, **21**(3), 215–224.
28. Steyvers, M. and Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, **29**(1), 41–78.
29. Voss, J. (2005). Measuring Wikipedia. In P. Ingwersen and B. Larsen, editors, *ISSI2005*, volume 1, pages 221–231, Stockholm. International Society for Scientometrics and Informetrics.
30. Zanette, D. H. and Montemurro, M. A. (2005). Dynamics of text generation with realistic Zipf's distribution. *Journal of Quantitative Linguistics*, **12**(1), 29–40.
31. Zipf, G. K. (1935). *The Psycho-Biology of Language*. Houghton Mifflin, Boston.
32. Zipf, G. K. (1949). *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, Cambridge, MA.