Basic word statistics for information retrieval: thesaurus as a complex network

Adriano de Jesus Holanda, Ivan Torres Pisa, Osame Kinouchi,
Alexandre Souto Martinez , Evandro Eduardo Seron Ruiz

[1]Faculdade de Filosofia, Ciências e Letras de Ribeirão Preto
Universidade de São Paulo
Av. Bandeirantes, 3900
14040-901 Ribeirão Preto, SP, Brasil.

evandro@usp.br

Abstract. *Words are the building blocks to construct sentences and to transmit information. Here, two distinctive hard classification approaches are applied to words. First, we consider words as being the nodes and their relationships as being the links of a directed graph. This permits us define, in a natural manner, the thesaurus conformation. The statistics of the outcoming and incoming links are characterized by simple fitting functions. Later, from a large collection of articles from* The New York Times *online newspaper, classified by thematical sections, we have shown that current spoken words in natural language is distributed according to the same Zipf's law. A combination of both approaches seems to be a promising tool for automatic information retrieval.*

Resumo. *As palavras são entes para a construção de frases e transmissão de informação. Aqui dois tratamentos distintos são aplicados para a classificação de palavras. Consideramos as palavras como sendo os nodos e suas relações como ligações de um grafo direcionado. Isto permite definir, naturalmente, a conformação de um dicionário de sinônimos. As estatísticas das ligações entrantes e emergentes são obtidas através de funções simples de ajuste. De uma coletânea de artigos do* The New York Times *online, classificada em seções temáticas, mostramos que palavras utilizadas em linguagem natural são distribuidas seguindo a mesma lei de Zipf. A combinação dos métodos parece ser uma ferramenta promissora na recuperação automática de informação.*

1. Introduction

It is believed that words are the building blocks to construct sentences and to transmit information. During last decades much effort has been spent on the statistics of words. Concern has been centered in the similarities and differences among word distributions towards automatic information retrieval.

Zipf [Zipf, 1972] has shown that word frequency obeys a power law if words are ranked from the most to the less frequent ones. Information retrieval, at its lowest level, can be exemplified by the Zipf's exponent. This exponent is very sensitive to the writer's instruction degree but much less sensitive to language (culture) (see, for instance, Figure 4).

Beyond word level, word connectivity has been treated in several manners. These treatments include entropic measures [Montemurro and Zanette, 2001] and the construction of other interesting quantities, such as the distribution of documents over the frequency of words [Volchenkov et al., 2003]. The "Latent Semantic Analysis" [Landauer and Dumais, 1997] deals with word connections from a corpus. It shows that to adjust typical word combinations a high dimensional Euclidean space must be considered. As far as we are aware, this is the only mechanism which has lead to practical applications, such as automatic grading of high school texts [Kintsch, 2002].

Other studies have focused on a different approach. Words are tied to each other as links of a graph were the words are the nodes of it. Exhaustive studies over thesaurus [Sigman and Gecchi, 2001, Motter et al., 2002] indicate that words are related among themselves as a small world network [Watts and Strogatz, 1998]. This result has been validated by a sampling procedure [Kinouchi et al., 2002].

In this paper we adopt the network view point. We add an important ingredient to study these networks: the directionality of the links. This allows us to define, in a natural manner, the thesaurus conformation. The thesaurus is presented in Section 2 where we show that it can be properly described considering a directed graph approach. Several subtleties are pointed out. The statistics of the outgoing and incoming links is characterized by simple fitting functions. In Section 3 we shown that current spoken words in natural language is very well distributed according to Zipf's law. We have gathered a large collection of articles from *The New York Times* (NYT) online newspaper originally classified by thematical sections. Finally, we draw the final conclusions in Section 4.

2. Moby Thesaurus II

A *thesaurus* is a list of terms. A *term* can be a word, a composed word or even an expression. The list of related terms to a main entry term (head-word) provides alternatives for these entries. Following previous studies, we will consider terms as being words in a broad sense.

Our study is based on a related term thesaurus, the *Moby Thesaurus II* which is the largest[1] and most comprehensive free thesaurus data source in English available [Ward, 2002]. It has 30,260 (main) *entries*, also called *root words* or *head-words*[2] and 73,046 words which are referred from the entries but that does not account for entries. Words which are not entries are called *non-root words*. These add up to 103,306 different words. Each root word points, in average, to 83 words[3]. We stress that the working thesaurus is a *related term* thesaurus. Definition terms thesaurus are not considered here.

The thesaurus derived network is defined considering each term as a node. Connections are established from an entry to its related list of terms. If only reciprocal terms (terms that refer to a given term and are referred by it) are considered, this forms a non-directed graph. The number of connections k of a node is called *degree of a node*. The structure of this thesaurus, as a non-directed graph, has been first studied by Motter et al.

[1] The file has 24,271 KB.

[2] Some curiosities are: 877 words which are not referred from other entries, 16 words are entry words but point only to non root words.

[3] From which 54 are root words and 29 are non-root words.

[Motter et al., 2002] where its small world nature has been pointed out. They have also obtained the node degree statistics showing an exponential behavior for small values of k and a power law behavior for large values of k.

2.1. Text processing

We have developed Perl programs to obtain the statistics of incoming and outgoing links from the thesaurus database. These programs turned out to be fast enough for our purposes and permit to establish a specific data structure that facilitates the walk through the word relations.

2.2. Directed graph

A more consistent and natural view of the thesaurus structure is to consider the access to any particular word from an entry word. A directed link from the head-word to the appropriate finding term would refer to what is called a *directed graph*.

We interpret the usual word classification of *root words* (the entry words) as the words with at least one emerging link ($k_{out} > 0$). On the other hand, words with no emerging links $k_{out} = 0$ are known to be *non-root words*.

From the directed graph terminology, root words are the *sources* and non-root words are the *sinks* of the thesaurus network, which may be connected by a giant strong component [Newman, 2001, Dorogovtsev and Mendes, 2001].

The non-directed graph considered in Reference [Motter et al., 2002] can be obtained from the directed structure considering only the co-linked terms (mutually referred terms).

This consideration of directed graph permits us to classify the terms according to the link properties as follows:

sink composed of the 73,046 terms with $k_{out} = 0$. For example: glucose, password, all-around, grape juice, send word, put to, lap dog, afterbirth;

source are the 30,260 terms with at least one outgoing link ($k_{out} > 0$), usually called main entries, entries, head-words or root words. The source can be divided into three categories;

- **absolute source** is related to the 877 terms without incoming links $k_{in} = 0$. For example: rackets, grammatical, double quick, half moon, blinded;
- **normal source** are 29,333 terms that receive links and send links to other source and sink terms. For example: ablation analogy, call out, factitious, laid low, make a deal;
- **bridge source** they are the 16 terms without outgoing links to source terms, listed: androgyny, Christian sectarians, Congress, detector, electric meter, enzyme, Esperanto, et cetera, Geiger counter, ghetto dwellers, harp, in fun, lobotomy, penicillin, perversely, Senate;

These definitions are illustrated by the subgraphs 1–4 in Figure 1.

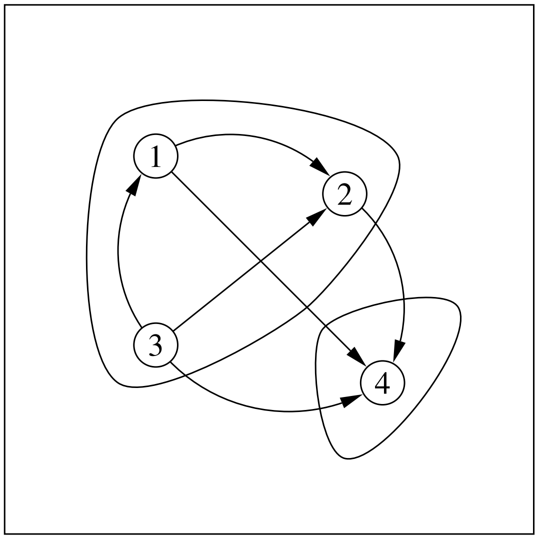

Figure 1: Coarse grained view of the thesaurus as a directed graph. The region composed by subgraphs 1 to 3 is the *source* and subgraph 4 is referred as the *sink*. The source contains: the *normal source*, named as subgraph 1; *bridge source*, named as subgraph 2 and *absolute source*, here called subgraph 3.

2.2.1. Outgoing links

In Figure 2 we show that the distribution of outgoing links is well approximated by the distribution:

$$f(k_{out}) = \frac{N_0}{1 + \lambda k_{out}^q} \ , \tag{1}$$

where we have found the values: $N_0 = 468 \pm 3$, $\lambda = (2.0 \pm 0.4) \times 10^{-5}$ and $q = 2.55 \pm 0.03$. This is an interesting fitting because for small values of λk_{out}^q the distribution can be well approximated to a stretched exponential [Laherrère and Sornette, 1998]:

$$f(k_{out}) = N_0 \exp\left(-\frac{k_{out}}{\bar{k}_{out}}\right)^q \ , \tag{2}$$

with $\bar{k}_{out} = \lambda^{-1/q} = 70 \pm 6$. This value is close to the mean number of entries in the source which is 83.

Another fitting curve [Tsallis and de Albuquerque, 2000] has been tested:

$$f(k_{out}) = \frac{N_0}{[1 + (q-1)\lambda x]^{q/(q-1)}} \ , \tag{3}$$

and we have obtained for this fit: $N_0 = 654 \pm 7$, $\lambda = (1.66 \pm 0.01) \times 10^{-2}$ and $q = 0.95 \pm 0.01$.

Comparing both fitting functions one finds that Equation 3 is not as appropriate as Equation 1. This can be seen by the fitting quality parameters: $r^2 = 0.993$ and $\chi^2 = 77$ for the former case against $r^2 = 0.990$ and $\chi^2 = 99$ for the latter.

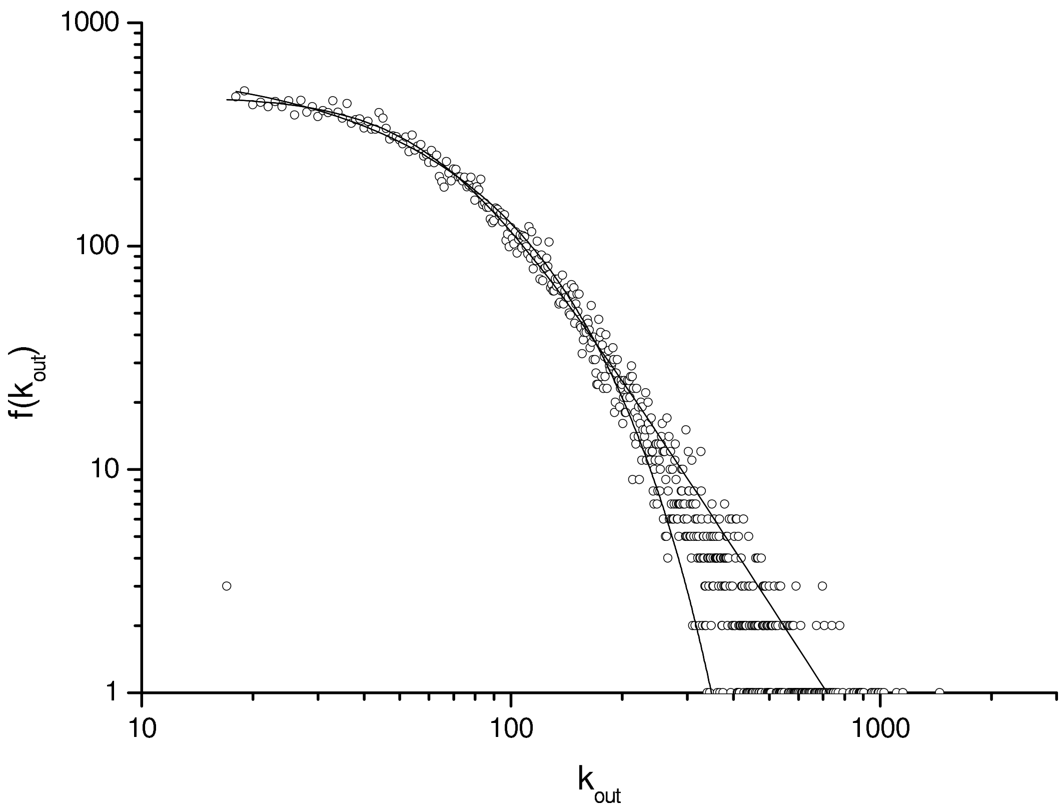

Figure 2: The frequency of outgoing links k_{out} (root words). is well described by Equation 1 which is the rightwards curve, contrasting with the curve of Equation 3. The point [$k_{out} = 17, f(k_{out}) = 3$] has been excluded in both fitting procedures. These words are: for good, for keeps, and grin.

2.2.2. Incoming links

We show in Figure 3 that the frequency of words with a given number of incoming links (k_{in}) is very well described by the stretched exponential curve:

$$f(k_{in}) = N_0 \exp\left(-\frac{k_{in}}{\bar{k}_{in}}\right)^q, \tag{4}$$

where we have found $N_0 = 12132 \pm 271$, $\bar{k}_{in} = 4.9 \pm 9.2$ and $q = 0.523 \pm 0.009$, with fitting parameters: $r^2 = 0.993$ and $\chi^2 = 4.576$. We shall stress that a fitting curve of the type of Equation 1 is not appropriate. A simple approximation may be used as: $f(k_{in}) \propto \exp(-\sqrt{k_{in}})$. Low values of incoming links ($k_{in} < 10$) are dominated by non-root words while great values ($k_{in} > 100$) are dominated by root words, as seen in Figure 3.

3. The New York Times

We have collected data from NYT online newspaper from December 30$^{\text{th}}$, 2002 to February 27$^{\text{th}}$, 2003. For this, we developed an agent software – here so called *Hunter* – using Borland Delphi (`www.borland.com`) that was able to simulate a human-access

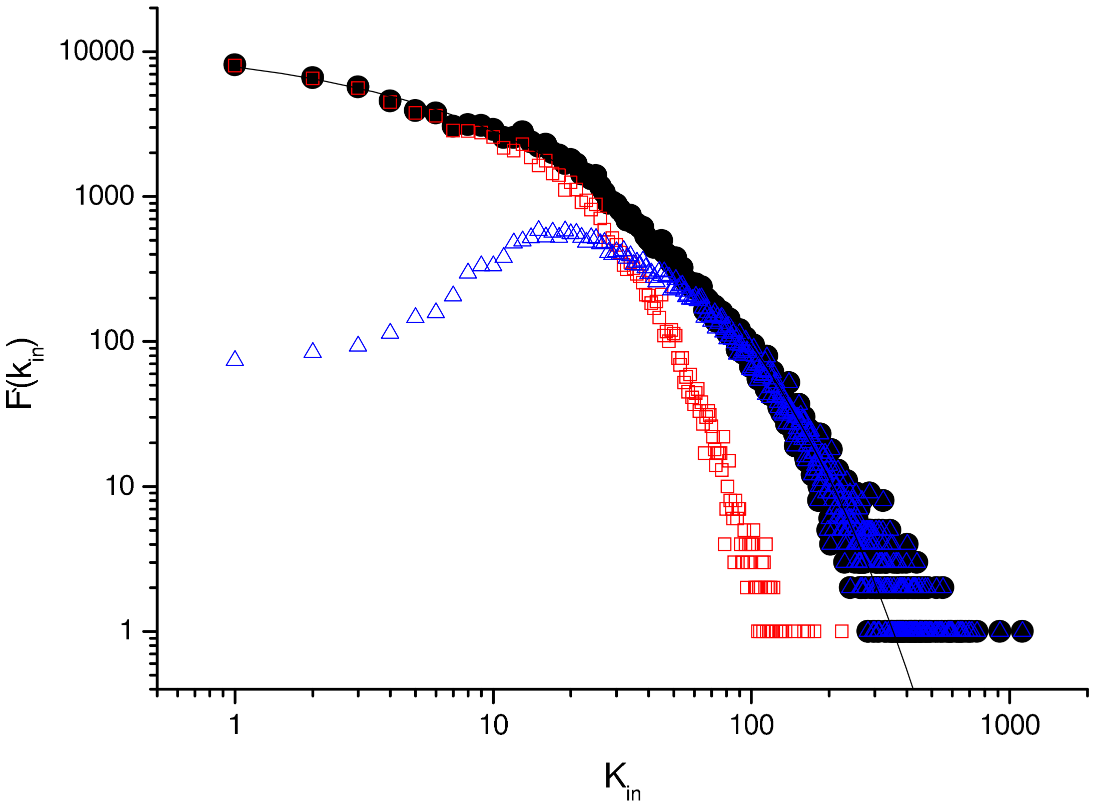

Figure 3: Frequency of incoming links k_{in} for all words (•), root words (△) and non-root words (□). The curve for *all words* is well described by a stretched exponential (line) expressed by Equation 4 ($N_0 = 12132 \pm 271$, $\bar{k}_{in} = 4.9 \pm 9.2$ and $q = 0.523 \pm 0.009$) which is dominated by non-root words for low k_{in} values $(k_{in} \leq 10)$ and by root words for great k_{in} values $(k_{in} \geq 100)$.

behaviour to NYT website. We also have used MySQL Database Server and Connector/ODBC (`www.mysql.com`) for news archiving. Only real content have been accomplished disrespecting special SGML tags and formatting errors. Each news is composed of a title, an abstract, and a body. The news collected belongs to at least one of thirteen different categories according to NYT: arts, books, business, crosswords, education, fashion, health, New York region, obituaries, science, sport, technology, and world. The non-classified news were rejected. These categories can still be divided in 54 subcategories, which had not been used in this analysis. From about 9,000 news, we have effectively considered 5,485 distinct ones, resulting in 48MB of data. The others have been neglected mainly because of misspelling errors on the content of the fields. For this task we developed a *Cleaning* software that defines a simple signature to each news for unique identification. We also developed a *Word analyzer* software to count the words from the news collected. It is worth mentioning that the total number of words in the 5,485 considered news is 4,478,160. From these words:

- 42,982 words were in the title, with 8,998 of them been distinct of each other;
- 70,387 words were in the abstract, with 11,608 of them been distinct, and
- 45,146 words were distinct in the body[4].

It is interesting to point that the word frequency, according the their usage rank from each of the considered categories, follow an universal law (the Zipf's law). The only exception might be the newspaper crosswords category (see Figure 4).

4. Conclusion

A thesaurus is a tentative to synthesize terms and their relationships as natural as possible. Nevertheless this trial is artificial and subjective. The real term connections are indeed found in written current texts, and in future ones. Here we have taken samples of regarded popular literature which covers a large spectrum of knowledge. From a collection of articles about various themes, which we believe reflects current spoken language, and using established statistical tools, we have demonstrated that regular literature behaves evenly, no matter the subject concerned. Also, our original work of treating the thesaurus as a directed graph has provided a new insight into its macro structure. From this approach the counting of k_{in} and k_{out} could lead to a novel proposition of term arrangement and term connectivities in it. The standard thesaurus classification is made according to word acception. A similar mathematical equivalence is to describe the graph by the emerging links from the nodes, here called k_{out}. Writing a thesaurus from the k_{in} perspective could not lead to practical facilities, nevertheless is mathematically equivalent to the preceding k_{out} description. The k_{in} description has shown more adequate to distance measurement between terms than k_{out}. Up to now we are working on a plausible word distance definition using the above descriptors. We also stress that the above procedure would be more effective than the word frequency counting as proposed by Zipf. The latter may not lead to any significant text discriminator, since all texts obey the same law.

Now, from these preliminary conclusions, we firmly believe that these mathematical/statistical tools could be used to reflect a single upper level structure that ultimately represents natural language.

[4] When considering only 1/3 of all words from the field body, due to computer time.

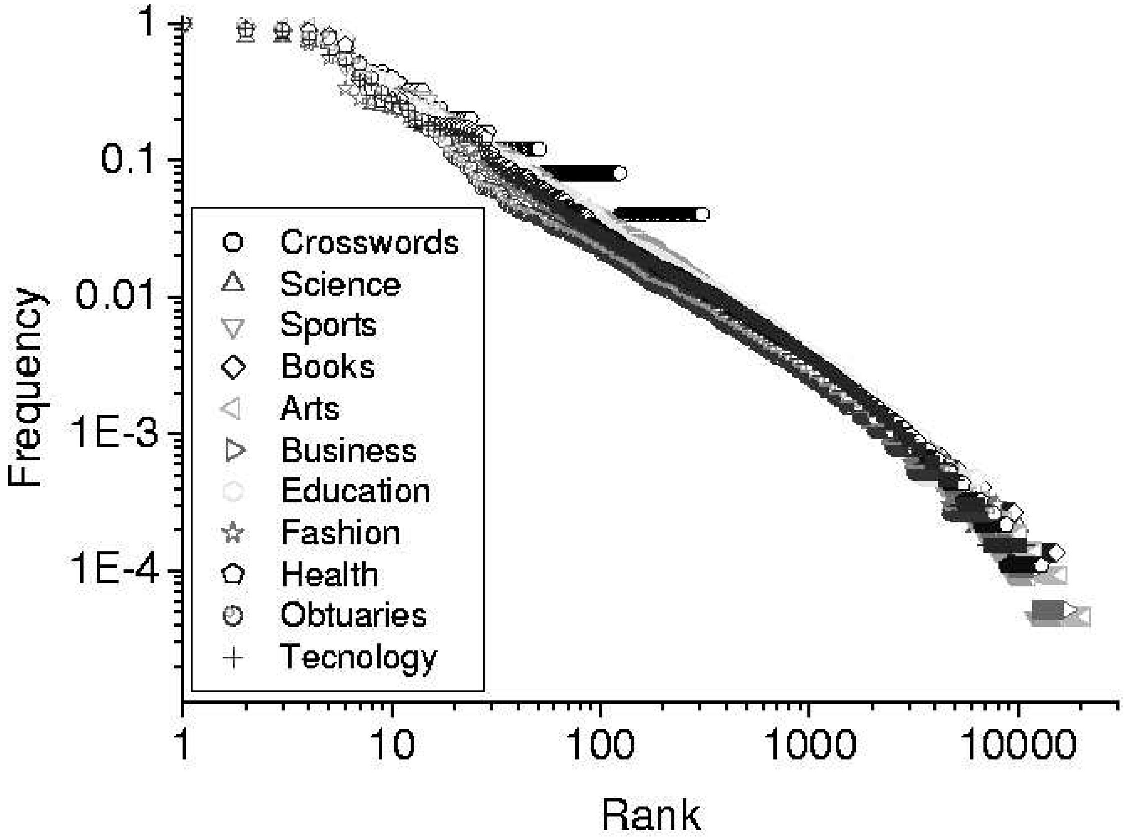

Figure 4: **New York Times frequency of words by rank. This shows that all the sections of this newspapers follow the same Zipf's law, maybe with the exception of crosswords section.**

5. Acknowledgements

The authors are deeply greatful to Vera Lúcia Coelho Villar from the Instituto Antônio Houaiss de Lexicografia, Brazil, for the fruitful discussions. This work has been partially funded by the Brazilian agencies: FAPESP, CAPES and CNPq.

References

Dorogovtsev, S. N. and Mendes, J. F. F. (2001). Languages as an evolving web. cond-mat/0105093.

Kinouchi, O., Martinez, A. S., Lima, G. F., Lourenço, G. M., and Risau-Gusman, S. (2002). Deterministic walks in random networks: an application to thesaurus graphs. *Physica A*, 315(3/4):665–676.

Kintsch, W. (2002). The potential of latent semantic analysis for machine grading of clinical case summaries. *Journal of Biomedical Informatics*, 104:3–7.

Laherrère, J. and Sornette, D. (1998). Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. *European Physical Journal B*, 2:525–539.

Landauer, T. K. and Dumais, S. (1997). A solution to plate's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2):211–240.

Montemurro, M. A. and Zanette, D. H. (2001). Entropic analysis of th role of words in literary texts. Entropic analysis of th role of words in literary texts, cond-mat/0109218.

Motter, A. E., de Moura, A. P. S., Lai, Y.-C., and Dasgupta, P. (2002). Topology of the conceptual network of language. *Physics Review E*, 65:065102(R).

Newman, M. E. J. (2001). Ego-centered networks and the ripple effect –or– why all your friends are weird. cond-mat/0111070v1.

Sigman, M. and Gecchi, G. A. (2001). Global organization of the lexicon. cond-mat/0106509.

Tsallis, C. and de Albuquerque, M. P. (2000). Are citations of scientific papers a case of nonextensivity? *European Physical Journal B*, 13:777–780.

Volchenkov, D., Blanchard, P., and Sharoff, S. (2003). Core lexicon and contagious words. Core lexicon and contagious words, cond-mat/0303454.

Ward, G. (2002). Moby Thesaurus II. ftp://ibiblio.org/pub/docs/books/gutenberg/-etext02/mthes10.zip. Project Gutenberg Literary Archive Foundation.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of "small world" networks. *Nature*, 393:440–442.

Zipf, G. K. (1972). *Human behavior and principle of least effort*. HafnerPublishing Company, New York, N.Y.