Large Text Networks as an Object of Corpus Linguistic Studies

Alexander Mehler
Bielefeld University, D-33615 Bielefeld, Germany
Alexander.Mehler@uni-bielefeld.de

Contents

1 Introduction

In simple mathematical terms, a corpus of natural language texts can be defined as a set which abstracts from any order of its elements w.r.t each other so that each element is separately processed by some corpus linguistic operation (e.g., of collocation statistics). This view implies the absence of structure formation within the corpus or at least disregards it from the point of view of text representation and subsequent corpus building. Of course, a corpus of natural language texts is more than just a *set* of linguistic units. There is structure formation above the level of single texts which can be made accessible to corpus linguistic studies. According to Stubbs (1996), texts are oriented to routines and conventions; they are shaped by prior texts to which they make intertextual references possibly (or preferably) included in the same corpus. In this sense, Stubbs (2001) points out:

"Analysis cannot be restricted to isolated texts. It requires an analysis of intertextual relations, and therefore comparison of individual instances in a given text, typical occurrences in other texts from the same text-type, and norms of usage in the language in general." (Stubbs, 2001, 120).

With the advent of web-based communication, more and more corpora are accessible which manifest such intertextual relations and thus structure formation in large text networks. Moreover, the WWW does not only manifest a tremendous set of text types (genres and registers) which already existed before the appearance of WWW-based communication, but also a vast number of instances of newly emerging document types, e.g., *corporate sites*, *Wikis*, *weblogs* or *personal academic home pages* (Mehler and Gleim, 2006, 2005; Thelwall and Wouters, 2005). Theoretically, this makes the web the source of choice for extracting large corpora of certain genres, registers and other linguistic varieties. It also makes the web the reference point of studying the emergence of *hypertext types* as well as the growth, maturity stage and dying of their instances manifested by websites and their constitutive pages. Thus, the web has become increasingly important as a quasi inexhaustible resource of corpus formation (Baroni and Bernardini, 2004; Keller and Lapata, 2003; Kilgarriff and Grefenstette, 2003; Resnik and Smith, 2003; Santamaría et al., 2003).

Of course, the web is not the only resource of large networks of textual units. There exist special areas of textual networking which become accessible to corpus linguistic studies not only because of their web-based interfaces, but also due to digitised or e-text releases (Hockey, 2000). This includes the area of *scientific communication* (e.g. CiteSeer or CiteBase as examples of digital libraries), *press communication* (e.g. the New York Times or the German Süddeutsche Zeitung which link articles to thematically related ones), *technical communication* (e.g. the Apache Software Foundation's technical documentations of open source projects) and *electronic encyclopedias* (e.g. Wikipedia and its releases in a multitude of languages) which can be analysed in terms of corpora of networked units. These are examples of large corpora of interlinked texts which in the majority of cases utilise HTML in order to manifest intertextual relations as, for example, citation links (digital libraries), content-based add-ons (online press communication) and links to related lexicon articles (electronic encyclopedias). From a corpus linguistic point of view, several scientific questions come to the fore regarding the formation of such networks:

1. Preprocessing: *How to provide a uniform, generic interface to the analysis of intertextual relations as manifested in web-based communication?*

2. "Co-textualising" corpus linguistic analyses: *How to explore text linkage in large text networks as a source of corpus linguistic studies?*

   This question is related to the position of Fairclough (1992) who argues for an intertextual view of analysing, for example, pre-constructed phrases and fixed collocations. In this sense, digital manifestations of intertextual links provide a source of exploring linguistic structures which are confirmed by intertextually related texts. Cf. also Mehler and Gleim (2006) for the notion of collocation analyses which are sensitive to genre-specific structures.

3. Exploring structure formation in large text networks: *What are the regularities of the distributed formation of large text networks subject to the limitations of the medium in use?*

   As structure formation in large text networks cannot be reduced to the intentionality of single authors, the question for distributed processes of text production and processing — distributed over thousands of collaborating/competing authors — comes to the fore.

This article reviews the state of the art in these areas. As corpus linguistic studies of large text networks are at the very beginning, this will relate especially to the third question. Interestingly, arguments in support of the need of text network analyses come from computer science and, especially, from the field of text and web mining (Mehler and Wolff, 2005). This relates to the so-called *link-content conjecture* of Menczer (2004) who states that *the content of a web page is similar to the content of the pages that link to it*. As Menczer approaches content in terms of Information Retrieval (IR) and, thus, in lexical terms, this hypothesis can be reformulated as follows: *A page's lexical organisation is similar to the lexical organisation of the pages that link to it* (where lexical similarity is measured in the framework of the vector space model based on a tf-idf weighting scheme and the cosine measure — the tf-idf weighting scheme is a function of the term and document frequency of candidate terms; it is used to filter out non-descriptive terms in the sense of IR; e.g., words which are evenly distributed over all texts of the corpus and, thus, do not contribute to thematically separating them — for more details on this model cf. Baeza-Yates and Ribeiro-Neto 1999). Menczer (2004) presents data in support of this conjecture which also points at an exponential decay of the similarity in question from the point of view of a focal page when following hyperlinks from page to page. As will be motivated, this observation is in accordance with so-called small world models of social-semiotic networks (cf., for example, Albert et al., 1999). Supposing that Menczer's hypothesis is continuingly supported, it implies that additional data in support of the data being observed on a focal page comes from its neighborhood in the web — by analogy with the argument of Stubbs (2001) cited at the beginning of this section. Although this hypothesis has not yet been investigated in the case of more traditional text networks, it is nevertheless plausible to conjecture it in these cases too. This may look as follows: *A text's linguistic organisation is similar to the linguistic organisation of the texts that relate to it by means of intertextual relations*. In order to substantiate this conjecture, the notion of *linguistic similarity* needs to be operationalised as well as the aspect of linguistic organisation under consideration and the type of intertextual relation for which this conjecture actually holds. *Text network analysis* is a step into this direction as it investigates principles of intertextuality which should be taken into account in the course of corpus building in order to meet the requirement of Stubbs (2001) and related requirements.

This article puts emphasis on the state of the art of network analysis (Newman, 2000, 2003b) and

its utilisation in the area of linguistic systems (cf. Ferrer i Cancho et al., 2005). Amongst others, this includes approaches to the notion of the small world of social systems (Watts and Strogatz, 1998; Watts, 1999). From the point of view of quantitative linguistics, this comprises cluster and pathway analysis (Newman 2003). Moreover, non-linear regression analyses of degree distributions (based on the number of in- and outgoing links) which relate to Zipfian regularities (Rapoport, 1982) are reviewed too. The article demonstrates this analytical apparatus by example of some document networks. The aim is to exemplify the state of the art of quantitative network analysis in the area of linguistic networks. From the point of view of corpus linguistics, the significance of this kind of analysis is due to the fact that it gives hints at how to quantify validity constraints of corpora based on intertextual regularities of their constitutive texts.

Text network analysis is at its very beginning, in corpus linguistics as well as in computational and quantitative linguistics. Although the notion of intertextuality comes into age (Fix, 2000), it nevertheless has been addressed in terms of qualitative, descriptive, but not of exploratory corpus linguistics. Theoretical definitions which allow to demarcate the field of document network analysis are still missing. Accordingly, the subsequent section introduces some preliminary notions for this task.

1.1 A Short Note on the Corpus Linguistic Relevance of Complex Network Analysis

The analysis of complex text networks is about structure formation in corpora of textual units. For decades, text linguistics has argued that intertextuality is a source of structure formation *above* the level of single texts (de Beaugrande, 1980, 1997; Heinemann, 1997; Hoey, 1995; Holthuis, 1993; Jakobs, 1999; Raible, 1995). This kind of structure formation has two aspects: in terms of the development of text types and in terms of the networking of their textual instances. Following this line of argumentation, Fairclough (1992) points out that intertextual relations allow to explore related texts and, thus, to identify significant cotexts as additional, viable data resources of corpus linguistic studies: If two texts x and y are intertextually related due to their common or related functions or topics, they probably contain common or related linguistic manifestations of these functions or topics, respectively, and, thus, are more likely structured in a similar way (Biber, 1995; Brinker, 1991). This correlation may hold on the level of lexico-grammatical patterns (Halliday, 1966) as well as on the level of textual superstructures (van Dijk and Kintsch, 1983). In other words: Studying intertextually related texts provides additional data to lexical and grammatical patterns and their variation subject to the change of the underlying genres or registers, respectively (Biber, 1995; Ventola, 1987). However, in order to benefit from intertextuality as a data resource we need to explore its principles first. That is, we need to make it an object of computational linguistics, not only on the level of pairwise linked texts, but on the level of whole networks based thereon. *This is the task of complex text network analysis*.

On the other hand, knowing the principles of intertextual structure formation (i.e. of the development of complex text networks) provides knowledge about constraints of the "naturalness" or "non-artificiality" of text corpora *by analogy with Zipf's first law* in the case of lexical systems. This can be explained as follows: It is well confirmed in quantitative linguistics that the ranked frequency distribution of lexical text constituents is highly skewed in a way which departs from the normal distribution,

	text area	e-text area	hypertext area
atomic level	text component	*e*-text component	text module
intermediate level	text	*e*-text	hypertext document
network level	text network	*e*-text network	hypertext document network

Table 1: Levels of structuring of text, *e*-text and hypertext document networks (cf. Storrer, 2002; Mehler, 2005).

but is more reliably modeled by a power-law or some related distribution (Baayen, 2001; Rapoport, 1982; Wimmer and Altmann, 1999a; Zipf, 1972). Accordingly, a "text candidate" which heavily departs from this Zipfian distribution indicates to be a mixture of different texts (possibly written by different authors) or to be an artificial product which was produced under "unusual" conditions disturbing the process of text production (Orlov, 1982). Analogously, a corpus of texts whose intertextual networking departs from the principles of text networks may indicate artificiality in the sense to be a mixture of topically or functionally highly divergent and, thus, unrelated texts. Using such a corpus as a starting point of inductive reasoning in corpus linguistics (Stubbs, 2006) is, thus, problematic. Complex network analysis allows to explore such naturalness constraints of corpus formation.

In summary, the intertextual formation of linguistic patterns as well as quality constraints of text corpora are two reference points in support of the relevance of complex network analysis in corpus linguistics. This article surveys the state of the art in this field of research.

1.2 Text and Document Networks

Structure formation above the level of texts is based on intertextual relations which span *networks* in which nodes denote texts (or textual components thereof) and links manifest coherence relations of these nodes. With the advent of web-based communication, text networking is not only accessible by means of e-texts and their networks, but also by hypertexts which utilise hyperlinks in order to make intertextual relations explicit (Mehler, 2005). Starting from the notion of a document which integrates textual content with hypertextual add-ons (Kuhlen, 1991), all three kinds of networks are taken into account in this survey: *text*, *e-text* and (hypertext) *document networks* — see Table 1. For reasons of terminological simplicity we simply speak of *text* and *document networks* and use both terms interchangeably (while we make it explicit if only one, but not the other term is adequate). Generally speaking, such networks are characterised as follows:

- *Intertextuality:* Text and document networks are units to which intertextuality can be ascribed as a gradual, quantifiable property by analogy with textuality as a property of single texts.

  Intertextual cohesion or coherence relations interrelate different texts or documents in order to build (not necessarily mutual) constraints on their interpretations. For a formal model of such constraints (with a focus on intratextual ones) cf. Mehler (2007). A survey of this notion is out of reach of the present paper — cf. Mehler (2005) for such a survey. We only stress the fundamental distinction of *referential* and *typological intertextuality* (Heinemann, 1997): Whereas the former comprises immediate text-to-text relations, which authors manifest more or less explicitly

by surface structural markers, it is the shared usage of the same or alike patterns within different texts which mediates their typological, but not necessarily intended relatedness. Since intertextual relations are in many cases *implicit*, they first need to be explored in order to become an object of network analysis. Intertextual relations of web documents may, but do not need to be manifested by hyperlinks. As in the case of cohesion and coherence in general, there are many resources of intertextuality so that ascribing this property to a text or document network is bound by vagueness and under-specification due to a diversity of possibly competing criteria. Even in the case of citation relations, exploring intertextual relations can be a demanding task in terms of computational linguistics and machine learning (Giles et al., 1998). In any case, the starting point of analysing text and document networks is a network of textual units which is spanned by their intertextual relations. Thus, complex text or document network analysis is about structural analyses of networks whose links are spanned by cohesion or coherence relations which in the majority of cases are meaning- or content-based.

- *Chaining and clustering:* Intertextuality results from producing or processing intertextual relations. These relations generate chains or clusters of thematically related texts/documents which manifest the same, similar or otherwise associated themes, topics or fields. On the other hand, the chains or clusters may be induced by schematically ordered texts/documents which manifest the same or related text types, patterns or superstructures. Note that whereas chains are partially ordered, clusters are clumps of interrelated units. Finally, as the chains and clusters overlap or intersect, respectively, they constitute networks.

- *Variability:* As intertextual relations are genre-sensitive or -specific (e.g. citations in scientific communication vs. content-based links in online press communication), text and document networks as a whole are genre-sensitive, too. That is, for different genres (e.g. of scientific, technical or press communication) variations in topological and statistical characteristics of the networks of these genres are expected. That is, genres are expected to be distinguishable in terms of the characteristics of their document networks.

- *Distributed cognition:* The production and reception of text and document networks is necessarily *distributed* over possibly hundreds and thousands of agents. They result from cooperative or competitive sign processes in the sense of distributed cognition (Hollan et al., 2000) and, thus, manifest a kind of superindividual structure formation which cannot be reduced to intentional acts of individual interlocutors — *comparable to the language system, but on the level of its manifestation.* That is, as the lexicon of a language cannot be attributed to single interlocutors, text networks are (because of their size) structured in a way which is not controlled by any (group of) such interlocutors separately. But whereas the lexicon is part of the language system, texts and the networks they induce are manifestation units.

In order to grasp the principles of this kind of networking, a combined approach which integrates at least topological and statistical methods is needed. This can be motivated as follows: According to Bense (1998), the formal branch of text linguistics includes *algebraical*, *topological* and *statistical* aspects. Whereas algebraic approaches to discourse grammars rely on the notion of *constituency* and *dependency* (Polanyi, 1988), it is the notion of *distance* and *neighborhood* which underlies topological

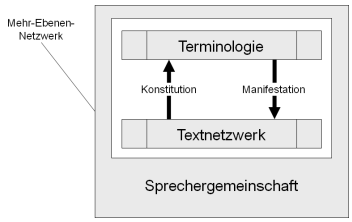

Figure 1: A three-level model of networking.

models (Brainerd, 1977). In contrast to this, the notion of *occurrence*, *co-occurrence* and *repetition* are the core of statistical approaches (Altmann, 1988). A central aim of quantitative linguistics is to investigate those types of repetition which bring about the statistical nature of linguistic structure formation in-between the extreme values of complete randomness and determinism.

Because of their non-linear, non-hierarchical structure formation, text and document networks are only adequately described by means of *graph theory* (Schenker et al., 2005) and *network analysis* (Newman, 2003b). Moreover, because of the size of these networks of hundreds and thousands of nodes there is no alternative to *automatic statistical* analyses. By analogy with stochastic discourse grammars which combine algebraic with statistical modeling, exploring these kinds of networks demands integrating topological and statistical approaches. Thus, the methods of *statistical network analysis* as elaborated in social science and subsequently sophisticated in physics builds the methodic core of this survey.

1.3 Delineation and Terminological Notes

This survey is about regularities of large networks of textual units as a special kind of complex networks. A network is called *complex* if it consists of hundreds and thousands or even millions of nodes in a way which affects its self-regulation and -organisation (Milgram, 1967; Newman, 2003b). The aim of analysing complex networks of texts or documents is to investigate indicators of structure formation which can be utilised for the task of corpus building and maintenance. In his survey of the structure and function of complex networks, Newman (2003b) reports on results of analysing social, informational, technical and biological networks. Amongst others, this comprises co-authorship and company director networks, WWW-based networks and citation networks, the Internet and peer-to-peer networks as well as protein interaction and neural networks. Focusing on social, technical and informational networks, Park (2003) interrelates these areas as follows: As a sort of informational network, hyperlink networks are based on the Internet as a kind of technological network which, in turn, manifests a communication network as a sort of social network in which nodes denote interconnected individuals.

Starting from this general view, we can delineate the object of the present survey as follows: Generally speaking, it does not regard social networks in which nodes denote individuals, agents, actors or communities thereof and where links represent social (communication) relations of these agents (Wasserman and Faust, 1999; Kautz et al., 1997; Otte and Rousseau, 2002). Other than these and related analyses, this review deals with networks whose nodes are linguistic units down from the level of words, up to the level of texts and hypertexts, respectively, where the main focus is on the latter.

Nevertheless, this review is not restricted to *hyperlink networks* (Park, 2003), but takes networking within *old* and *new* media into account thereby stressing the need to explore intertextual relations beyond hyperlinks as a source of networking within text corpora. As pointed out by the three-layer model in Figure 1, this does not deny the fact that text and document networks are manifestations of some linguistic system which, in turn, is enclosed by a corresponding social system (e.g. a speech community). Rather, it has to be understood as an indispensable reduction of the variety of network studies to be surveyed within this article. As will be shown in the subsequent section, this includes a wide area of text and document networks ranging from social software-based networks to networks in scientific and press communication.

A note on terminology: The terms *node* and *link* will be used when speaking about networks, while *vertex* and *edge* are used when speaking about graphs as formal models thereof. Further, as the apparatus of complex network analysis has predominantly been developed by example of social networks, we will use the term *social-semiotic* network in order to stress the encompassing field of linguistic, text and social networks as interrelated in Figure 1.

This article deals with complex networks of textual units. Such networks do not only form a special kind of complex networks but also large corpora. In other words, networks of textual units are a sort of *large linguistic corpora* whose specificity is due to their structuring based on the network inducing intertextual relations of their constitutive units. For the annotation and representation of large linguistic corpora in general see article 37 in this volume. A more specialised case of a complex network of textual units is given by networks whose nodes denote *web* documents. Analogously, we have a special kind of a *web corpus* structured by the hyperlinks of its constitutive elements when dealing with complex networks of web documents. For web corpora as an object of corpus linguistics in general see article 55 in this volume. See also article 21 in this volume on various corpora of computer-mediated communication including web corpora. This article deals more specifically with aspects of networking in corpora of textual units, its graph-theoretical representation and quantitative modeling in various areas of text and document networks.

The article is organised as follows: Section 2 introduces graph theoretical notions and outlines some results of the theory of complex networks as needed subsequently. This relates especially to the so-called small-world property which allows to separate the area of random and social-semiotic networks. Section 3 surveys network-oriented studies in corpus, computational and cognitive linguistics as well as in computer science. This includes but is not limited to lexical, sentence and WWW-based networks. Finally, Section 4 gives a conclusion and prospects future directions within the present field of research.

2 Structure Formation in Large Networks

The concept of a social-semiotic network in general and that of a small world in particular is formally narrowed down in terms of graph theory. That is, networks to be analysed as candidates of small worlds are, first of all, modeled as *graphs*. The following sections survey this kind of modeling: Section 2.1 starts with an overview of some notions of graph theory as used subsequently. The classical model of small worlds as introduced by Watts and Strogatz (1998) which, since then, has been applied in various areas of network formation is described in Section 2.2. Section 2.3 overviews the alternative

Notation	Description
$C_{\mathrm{BR}}(G)$	The cluster value of graph G according to Watts and Strogatz (1998).
$C_{\mathrm{WS}}(G)$	The cluster value of graph G according to Bollobás and Riordan (2003).
$d(G)$	The average degree of vertices of graph G.
$d_G(v)$	The degree of vertex v_i of graph G.
$\Delta(G)$	The diameter of graph G.
$E(G)$	The set of edges of graph G.
$\epsilon(G)$	An alternative coefficient of the average degree of vertices of graph G.
γ	The exponent of a power law fitted to the degree distributions of a given graph.
γ_{in}	The exponent of a power law fitted to the in-degree distributions of a given graph.
γ_{out}	The exponent of a power law fitted to the out-degree distributions of a given graph.
$L(G)$	The average geodesic distance of vertices of graph G.
$r(G)$	The correlation coefficient of the degrees of interlinked vertices of G.
θ	The exponent of a power law fitted to the cluster coefficient $C(k)$ as a function of degree k.
$V(G)$	The set of vertices of graph G.

Table 2: Basic graph theoretical notions used throughout the article.

model of Barabási and Albert (1999) who — other than Watts and Strogatz — take the temporal aspect of network growth into account. In the meantime, several more indices have been introduced in order to quantitatively classify networks. This relates, especially, to what is called assortative mixing as a characteristic of social instead of technical networks. Section 2.4 gives a short summary of it. Next, Section 2.5 describes concepts of structure formation within complex networks above the level of local clusters as considered in the model of Watts and Strogatz. Finally, Section (2.6) reconsiders time-dependent constraints of network formation.

2.1 Graph Theoretical Preliminaries

This subsection briefly surveys fundamental notions of graph theory as they are needed for complex network analysis. Table 2 summarises these and other definitions introduced subsequently. For a more thorough introduction to graph theory confer Diestel (2005), Melnikov et al. (1998) and Bronstein et al. (1999).

Let $[X]^k$ be the set of all subsets of k elements of X. *A simple undirected graph* G is a pair $G = (V, E)$ where V is the set of *vertices* and E the set of *edges* such that $E \subseteq [V]^2$. If $G = (X, Y)$ is a graph, then $V(G) = X$ denotes its *vertex set* and $E(G) = Y$ its *edge set*. The *order* $|G|$ of a graph G is the number of its vertices. An edge $e = \{v, w\} \in E$ is *ending at* v and w which are both *incident* with e and thus *adjacent* or *neighbors*. We also say that two edges are adjacent if they end at least at one common vertex. $E(v)$ is the set of all edges to which v is incident. G is *complete* if all its vertices are pairwise adjacent. A complete graph of order n is denoted by K_n. A *triangle* is a complete graph K_3 of order 3. $N_G(v)$ is the *set of neighbors* of $v \in V(G)$. Usually, the subscript is omitted if the graph referred to is evident. The *degree* $d_G(v_i) = k_i$ of a vertex v_i is the number $|E(v)|$ of edges ending at v. Evidently, $|E(v)| = |N(v)|$ (note that E is a set and therefore does not contain multiple edges which are introduced below). A graph is called *regular* (or *k-regular*) if all its vertices have the same degree

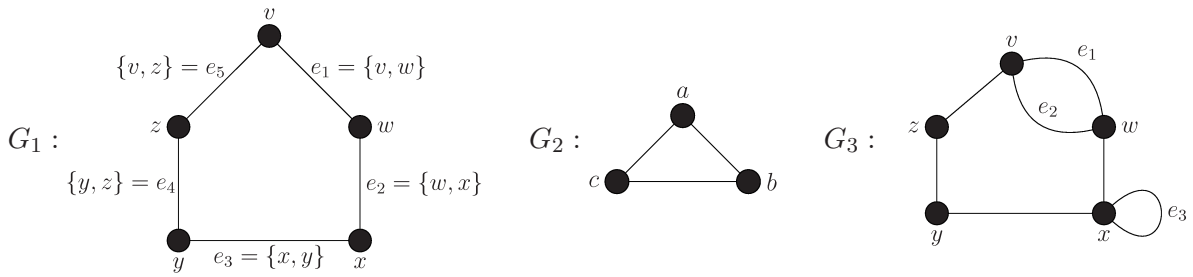

Figure 2: A graphical representation of an undirected graph $G_1 = (V, E)$ with $V = \{v, w, x, y, z\}$ and $E = \{e_1, \ldots, e_5\}$. G_1 is of order $|G_1| = 5$. Edge $e_3 = \{x, y\}$ is ending at vertex x and y. e_3 is adjacent with e_2 and e_4. Vertex x is adjacent to two edges, that is, $E(x) = \{e_2, e_3\}$. Further, $N_{G_1}(x) = \{w, y\}$ is the set of neighbors of x. The degree of x is $d_{G_1}(x) = |E(x)| = |N_{G_1}(x)| = 2$. G_1 is *not* complete. A triangle, that is, a complete graph of order 3, is exemplified by G_2. Note that G_1 and G_2 are both 2-regular graphs. Thus, $d(G_1) = d(G_2) = 2$ and $\epsilon(G_1) = \epsilon(G_2) = 1$. $(v, e_1, w, e_2, x, e_3, y)$ is a simple path with end vertices v and y. The distance $\delta(v, y)$ is 2 since (v, e_5, z, e_4, y) is the shortest path between v, y in G_1. The diameter $\Delta(G)$ of G is 2. Obviously, G_1 and G_2 are connected. G_3 demonstrates a multi- and pseudograph, respectively, with multiple edges e_1 and e_2 as well as a loop e_3.

(k). The *average degree* of a graph G is $d(G) = \frac{1}{|V|} \sum_{v_i \in V} d_G(v_i)$. In the following sections, we will alternatively refer to the ratio

$$\epsilon(G) = |E(G)|/|V(G)| = \frac{1}{2} d(G) \; . \tag{1}$$

A sequence $P = (v_{i_0}, e_{j_1}, v_{i_1}, e_{j_2}, \ldots, v_{i_{n-1}}, e_{j_n}, v_{i_n})$, $n > 0$, is called *walk* of *length* n between v_{i_0} and v_{i_n} in G, if for $k = 1, \ldots, n$: $e_{j_k} = \{v_{i_{k-1}}, v_{i_k}\} \in E(G)$. v_{i_0} and v_{i_n} are called *end vertices* of P. All other vertices are called *inner* w.r.t P. A walk is called *path* if all its edges are distinct. A path is called *simple* if all its inner vertices are distinct. A path is called *cyclic* if its end vertices are equal. The *distance* $\delta(v, w)$ of two vertices v, w, $v \neq w$, is the length of the shortest path ending at v and w. The *diameter* $\Delta(G) = \max_{v,w \in V(G), v \neq w} \delta(v, w)$ of a graph G is the maximal distance between any pair of vertices in $V(G)$. A non-empty graph G is *connected* if for any pair of vertices $v, w \in V(G)$ there exists a path ending at v and w. A maximal connected subgraph of G is called *component* of G. A graph G is called *bipartite* if its vertex set $V(G)$ is partitioned into non-empty disjunct subsets A, B such that every edge $\{v, w\} \in E(G)$ is ending at vertices $v \in A$ and $w \in B$. For reasons of clarity, we will call A and B the modes of the bipartite graph G and speak, more specifically, of the *bottom mode* and the *top mode* where the latter is seen to be placed "over" the former (see Figure 3).

So far we neither considered loops, nor multiple, parallel or directed edges which are grasped by the following definitions — these additional definitions are needed in order to map, for example, reflexive links from a web page to itself (i.e. loops) or different links between the same Wikipedia articles (i.e. parallel edges):

1. A *multigraph* is a pair (V, E) whose edge set E is defined as a *collection* of subsets of $[V]^2$ and, thus, may — in contrast to simple graphs — contain several copies of the same elements of $[V]^2$ where equal elements of E are called *multiple edges*.

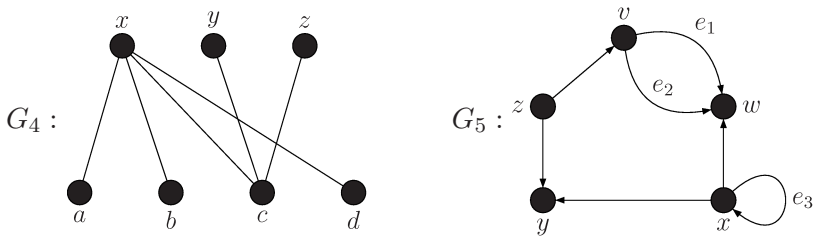

Figure 3: A bipartite graph G_4 whose vertex set $V(G_4)$ is separated by its edge set into the subsets $A = \{a, b, c, d\}$ and $B = \{x, y, z\}$. As a matter of convention we call A the top mode and B the bottom mode of G_4. G_5 exemplifies a directed graph where $\text{in}(e_1) = \text{in}(e_2) = v$ and $\text{out}(e_1) = \text{out}(e_2) = w$. Thus, e_1 and e_2 are not only multiple, but also parallel in G_5. G_5 is an orientation of G_3.

2. A *pseudograph* is a pair (V, E) where E is defined as a collection of unordered pairs of not necessarily different vertices of V. Thus, pseudographs may — in contrast to multigraphs — also contain loops.

3. A *directed graph* (or digraph) is a pair (V, E) of a vertex set V and an edge set E together with two functions in: $E \to V$ and out: $E \to V$ such that for every edge $e \in E$, $\text{in}(e)$ is the *initial vertex* and $\text{out}(e)$ the *end vertex* of e. Edges e_i, e_j, for which $\{\text{in}(e_i), \text{out}(e_i)\} = \{\text{in}(e_j), \text{out}(e_j)\}$, are called *multiple*. Edges e_i, e_j, for which $\text{in}(e_i) = \text{in}(e_j)$ and $\text{out}(e_i) = \text{out}(e_j)$, are called *parallel*. Finally, an *orientation* $D = (V, X)$ of an undirected graph $G = (V, E)$ is a directed graph such that for every edge $e \in X$: $\{\text{in}(e), \text{out}(e)\} \in E$. A graph is called *mixed* if it contains two sets E_1, E_2 of undirected and directed edges, respectively.

4. A graph is called *uniquely labeled* if its vertices have pairwise different labels. In order to simplify notation, we assume that indexed vertices $v_i, v_j \in V(G)$, $i, j \in \mathbb{N}$, are labeled by their indices. Thus, we sometimes will abbreviate v_i by i.

For additional notions of social network analysis which are used in order to characterise graphs in quantitative terms cf. Otte and Rousseau (2002). This relates, amongst others, to the notion of *components* and *cliques* on the one hand and of *density*, *centrality*, and *cohesion* (Egghe and Rousseau, 2003) on the other hand. See also Wasserman and Faust (1999) for a comprehensive overview on graph theoretical concepts in network analysis. For a survey of complex network analysis and its various fields of application see Newman (2000, 2003b). See also Watts (1999, 2003) and Strogatz (2001) for thorough introductions to this field. Further, see Thelwall et al. (2006) for a comprehensive overview of network analysis by example of the WWW. Finally, `www.cs.cornell.edu/courses/cs685/2002fa` is an excellent collection of links on complex network analysis.

The subsequent sections survey indicators of small world formation in complex networks as they were introduced in the literature. By default, these indicators are introduced for *simple undirected graphs* — in some cases, their derivation for directed and multi-graphs will be described.

2.2 Short Cuts and Clusters

What regards the overall structure of complex networks, these and related questions are, for the time being, investigated in terms of so-called *Small Worlds* (SW) (Newman, 2003b). Since its invention by

Milgram (1967), this notion awaited formalisation as a measurable property of large, complex networks which allows distinguishing them from random graphs. Milgram started from a social network of persons and their acquaintance (or friendship) links. He asked, so to speak, about the expected value of the shortest chain of such links which are connecting agents of a starting population from well specified target persons in a given population. More specifically, an agent of the starting population is presented with a description of a target person and asked to advance a letter to her by sending it to an acquaintance whom she considers more likely than herself to know the target. Each of these persons in turn advances the letter by the same procedure until the target person is reached. In social network analysis (Wasserman and Faust, 1999) this is a classical question about the cohesion of a network which affects the network's efficiency and vulnerability of information flow. In Milgram's model, the *short-cut* property, that is, the characteristic short average distance between any randomly chosen pair of nodes of a network, is seen as the central small world indicator. But this property alone does not delineate small worlds from random networks which also have the short-cut property, but obviously miss the kind of structure formation known from social networks. A first formalisation of small worlds was introduced by Watts and Strogatz (1998) who characterise them by *two* properties:

- Compared to *random graphs*, small worlds show a considerably higher level of *cluster formation*.
- Compared to *regular graphs*, any randomly chosen pair of vertices in a small world has, on average, a considerably shorter *distance*.

In order to operationalise these statements, Watts and Strogatz introduce two indicators of clustering and density, respectively. First, *clustering* in a simple undirected graph G is measured by the mean of the *cluster value* $C_{v_i}(G)$ of its vertices $v_i \in V(G)$. More precisely, clustering is measured as the mean of the ratio of the number $adj(v_i)$ of edges ending only at neighbors of v_i to the number of edges in a corresponding complete graph of order $|N_G(v_i)|$ (i.e. a graph in which all neighbors of v_i are adjacent) (note that in a simple undirected graph $adj(v)$ equals the number of triangles incident with v):

$$C_{v_i}(G) = \frac{adj(v_i)}{\binom{d_G(v)}{2}} = \frac{adj(v_i)}{d(v_i)(d(v_i)-1)/2} \in [0,1] \tag{2}$$

This allows to define the cluster value $C_{\mathrm{WS}}(G)$ of G as (note that $n = |V(G)|$)

$$C_{\mathrm{WS}}(G) = \frac{1}{n}\sum_{i=1}^{n} C_{v_i}(G) \in [0,1] \tag{3}$$

C_{WS} measures the average proportion of the neighbors of vertices that are themselves neighbors. It estimates the probability by which two vertices v, w are themselves adjacent when commonly linked with the same vertex u. In terms of friendship networks, for example, in which vertices denote individuals, a high cluster value C_{WS} means that the friend of a friend of a person is probably also the friend of that person.

The notion of clustering relates to the notion of *transitivity* in social networks: A triad of agents a, b, c is said to be transitive if whenever a links to b (i.e. $a \to b$) and $b \to c$, then $a \to c$ (Wasserman and Faust, 1999, 243). See Rapoport (1953) for an early study of transitivity patterns in social networks. As the term *clustering* has a completely different meaning, e.g., in explorative data analysis (Bock,

1994), the term *transitivity* is preferred in the network literature (Newman, 2003b). Following this manner of speaking, a *network* is said to be *transitive* to the amount of the probability that if any of its vertices a, b and b, c are linked, then a and c are linked, too. In this sense, C_{WS} is a candidate measure to estimate this probability.

Of course, the reason of edge formation varies with the network being modeled. In the area of social-semiotic networks, vertices may represent, for example, interlocutors which are seen to be linked whenever they communicate. Alternatively, vertices may represent the linguistic manifestations of this communication in the form of discourse units which are seen to be linked whenever they are related by one or more *intertextual* coherence relations. In this case, a high C_{WS} value means that if a focal discourse unit a is simultaneously related to some units b and c, then there is a high probability that there is an intertextual relation linking b and c, too.

A drawback of the definition of C_{WS} is that it does not appropriately operate on multigraphs. The reason is that it counts a triple only once even if spanned by multiple edges between the same vertices. Therefore, Bollobás and Riordan (2003) alternatively propose the cluster coefficient $C_{\mathrm{BR}}(G)$ as the fraction of the number of triangles within G and the number of pairs of adjacent edges:

$$C_{\mathrm{BR}}(G) = \frac{3 \times \text{number of triangles of } G}{\text{number of pairs of adjacent edges of } G} \in [0, 1] \tag{4}$$

High values of $C_{\mathrm{BR}}(G)$ and $C_{\mathrm{WS}}(G)$ indicate that linkage in G tends to be *transitive* in the sense that if any vertex $u \in V(G)$ is linked with vertices $v, w \in V(G)$, then v and w are probably linked, too. Obviously, this notion of clustering is not to be confused with cluster analysis in which clusters of any size are computed which optimise the proportion of cluster internal homogeneity and cluster external heterogeneity in the sense of the underlying similarity measure (Bock, 1994). In contrast to this, the reference point of clustering according to C_{BR} and C_{WS} is the local notion of a *triangle* based on *three* vertices. This "restriction" has been the starting point of developing more elaborate models of local structure formation — see, e.g., Milo et al. (2002) and Section 2.5.

A central observation of Watts and Strogatz (1998) is that in regular graphs of the sort they have examined, clustering is high, whereas in random graphs it is low. This value distribution is reversed by the average distances within regular and random graphs measured as follows:

$$L(G) = \frac{1}{\binom{|V(G)|}{2}} \sum_{\{v,w\} \in [V]^2} \delta(v, w) \tag{5}$$

Note that in the case of large networks (of hundreds and thousands of nodes), $L(G)$ is estimated by means of random samples of up to some thousand vertices of G and their geodesic distances.

Bollobás and Riordan (2003) point out that although by definition $L(G) \leq \Delta(G)$, $L(G)$ is mostly *not much smaller* than $\Delta(G)$. Therefore, $\Delta(G)$ is referred to as an alternative indicator of distance formation in small worlds (Albert et al., 1999).

In the case of hypertext networks, for example, small values of $L(G)$ indicate that the topic of pages changes fast, i.e., already after a couple of clicks when following the hyperlinks between their modules, supposing that the basic population is thematically diversified as, for example, in the case of the Wikipedia. Generally speaking, small average geodesic distances indicate rapid changes of a given variable V (e.g. topic, genre, register etc.) when following links between nodes of a network supposed that the values of V are diversified within this network.

Graph	Clustering	$C_{\mathrm{WS}}(G)$	$C_{\mathrm{BR}}(G)$	Distance	$L(G)$
regular	high	$k < \frac{2}{3}n \Rightarrow C = \frac{3(k-2)}{4(k-1)}$	cf. Bollobás and Riordan (2003)	long	$L = \frac{n}{2k} \gg 1$
SW	high	see Formula 3	see Formula 4	short	see Formula 5
random	low	$C = k/n$	cf. Bollobás and Riordan (2003)	short	$L \sim \frac{\ln(n)}{\ln(k)}$

Table 3: Cluster values and average distances in small worlds compared to regular graphs and random graphs. Estimators are given for corresponding random and regular graphs subject to the condition that $n \gg k \gg \ln n \gg 1$ (Watts and Strogatz, 1998) where $n = |V|$ and $k = d(G)$. $k \gg \log n$ ensures that the corresponding random graph is connected (Baldi et al., 2003).

Starting from $L(G)$ and $C_{\mathrm{WS}}(G)$, Watts and Strogatz (1998) narrow down their notion of a small world — henceforth called *WS model* (note that WS abbreviates the initials of the authors of this model). Their basic idea is to start with a regular graph whose edges are stepwise rewired with probability p such that for certain values of p small worlds emerge which simultaneously have high cluster values and short average distances. More specifically, they start from a $2r$-regular graph C_n^r, i.e. the rth power of an n-cycle, of fixed order $n > 2r$ in which vertices are adjacent whose distance within the n-cycle C_n is at most r (Bollobás and Riordan, 2003). Next, they derive a random graph $G(p)$ from C_n^r by rewiring a proportion $p \in (0, 1]$ of its edges. The "*'surprising' observation*" (Bollobás and Riordan, 2003) is that even for small values of p, that is, for the introduction of a small amount of randomness, small worlds emerge which share high cluster values with regular graphs and short distances with random graphs, that is:

$$C_{\mathrm{X}}(G_{\mathrm{regular}}) \quad \sim \quad C_{\mathrm{X}}(G_{\mathrm{sw}}) \quad \gg \quad C_{\mathrm{X}}(G_{\mathrm{random}})$$

and

$$L(G_{\mathrm{regular}}) \quad \gg \quad L(G_{\mathrm{sw}}) \quad \sim \quad L(G_{\mathrm{random}})$$

for $\mathrm{X} \in \{\mathrm{WS}, \mathrm{BR}\}$. This is summarised in Table 3 in which estimators of L and C_{WS} are given for the corresponding regular and random graphs of equal order (i.e. size) n and average degree $d(G)$ (as an index of sparsity or density) (cf. Bollobás and Riordan, 2003).

In summary, the WS model combines cluster formation with the formation of short distances by means of some short-cuts which provide efficient information flow within the network. In the area of social-semiotic networks, this property has been demonstrated by example of the WWW, collocation networks and thesauri. This is described in detail in Section 3.

The notion of a small world as emerging from introducing a small amount of randomness which generates short-cuts within initially regular graphs has been the starting point of a critical review of the WS model (Newman, 2000). One reason is that social networks are expected to be structured *far away from* the topology of regular graphs. Another reference point is the focus of the WS model on cluster and distance values and, thus, on node indices. $L(G)$ indicates a *global* network property in the sense that it aggregates values which interrelate all pairs of nodes of a network. In contrast to this, $C_{\mathrm{WS}}(G)$ indicates a *local* network property. The reason is that although cluster values are aggregated

for the whole network, their reference point are triangles and connected triples of vertices. Thus, both indices, $L(G)$ and $C_{\mathrm{WS}}(G)$, focus on single moments of the distributions of the corresponding input values and thus miss to describe these distributions in more detail. The following section describes a model which tackles this shortcoming.

2.3 Scale-Free Networks

Whereas the WS model describes small worlds under a static perspective, it is the dynamic perspective of network growth under which small worlds are described in the *preferential attachment model* of Barabási and Albert (1999) henceforth called *BA model*. It starts from the observation that the vertex connectivities of some complex networks are distributed according to a scale-free power law *in addition to their common property of short-cuts and local clustering*. More precisely, Barabási and Albert recur to the observation — confirmed by many social-semiotic networks, but not, for example, by instances of the random graph model of Erdős & Rényi (cf. Bollobás, 1985) — that the number of links per vertex can be reliably predicted by means of a power law. In other words: The probability $P(k)$ that a randomly chosen vertex interacts with k other vertices of the same graph representing a network is approximately:

$$P(k) \sim k^{-\gamma} \tag{6}$$

where γ is often between $[1.5, 3.5]$ (Newman, 2003b; Milo et al., 2002).

In the present case, in which power laws are fitted to the degree distributions of vertices of an undirected graph, this indicates that connectivity is scale-free and thus relates to the Zipfian nature of many social-semiotic phenomena (Rapoport, 1982) as, for example, in the case of the rank-frequency distribution of lexical units. Thus, networks with a power law-like degree distribution are called *scale-free networks* (Barabási and Albert, 1999). The exponent of the power law fitted to the degree distribution of a network is an indicator of the kind of its structuring which, in turn, is related to its *procedural* characteristics: Scale-free networks are known for their low vulnerability and fault tolerance (Albert et al., 1999). Generally speaking, a function $f(x)$ is called scale-free if it remains unchanged under rescaling of the variable x in the sense that $f(ax) = bf(x)$, $a, b \in \mathbb{R}$. Solutions to this equation have a power law form (Newman, 2005; van Raan, 2005). In the case of degree distributions, scale-freeness means that there are no typical nodes which represent all others because of their typical behavior (Barabási and Oltvai, 2004; Newman, 2005).

Power law-like degree distributions are contrasted by the Poisson distribution of node connectivity in random graphs (Bollobás, 1985). The Poisson distribution models the effect that the probability to find highly connected nodes decreases exponentially with k. This property also holds for the WS model — *contrary to empirical observations* which are better fitted by scale-free power laws (Barabási and Albert, 1999).

A power law can be fitted to the *rank-degree* distribution (where rank is determined by the decreasing order of node connectivity) or to the *size-degree* distribution (based on the number of vertices of degree one up to the number of vertices of highest degree). Fitting can be restricted to the distribution of vertex *in* or *out* degrees within directed graphs. Note that simple graphs do not distinguish multiple edges and may, therefore, displace the observed distribution. In the present context, successfully

fitting a power law indicates that the majority of nodes is poorly connected, while a selected minority of *hubs* is very highly connected (Watts, 2003). These hubs are mainly responsible for providing cohesion as they integrate the majority of nodes into the network (Ravasz et al., 2002). Thus, for a fixed number of links, the smaller the value of γ, the shallower the slope of the curve in a log-log plot, the higher the probability of higher connected hubs. In contrast to this, if the number of vertices of a certain degree decays exponentially with increasing degree, highly connected vertices (i.e. hubs) are very unlikely or do not exist. Three general remarks on power law-fittings:

- Firstly, although there is a definite relationship of rank (degree) distributions on the one hand and their cumulative correspondents or size-degree distributions on the other hand, the values of the exponents of power laws fitted separately to these distributions systematically depart from each other. This is explained in detail by Adamic (2000) and Newman (2005).
- Secondly, not only does the algebraic sign of a power law's exponent matter, but also its absolute value as it determines the existence and range of the expected value and variance of the corresponding theoretical distributions under the assumption of additivity — see Newman (2005) for these details. Thus, when comparing two studies we do not only need to know which empirical distribution (rank-degree or size-degree) was fitted, but also by means of which exponent.
- Thirdly, power laws are candidate distributions to be fitted to empirical distributions of, e.g., vertex degrees. Because of theoretical considerations as well as because of empirical observations or other restrictions, alternative distributions can be checked for their fitting as well — cf. Wimmer and Altmann (1999b) for the whole range of discrete probability distributions many of which became relevant in quantitative linguistics.

In order to derive a model which explains the *emergence* of power law-like node connectivities in networks *subject to their growth*, Barabási and Albert no longer view the number of vertices to be fixed and being rewired with a uniform probability as assumed by the WS model. Instead, they account for the dynamics of networks whose vertex set is continually growing by preferably linking new vertices with already highly connected ones. This preferential attachment produces a so-called *Matthew effect* (Simon, 1955) as it predicts that older nodes get rich in links at the expense of younger ones (Watts, 2003). In the case of text networks, the BA model says that newly added nodes tend to be added with texts providing a high amount of network coherence. As an example think of a citation network in which new documents tend to cite already frequently cited ones or of the Wikipedia in which new articles are predicted to preferably refer to already much discussed ones.

The basic idea of Barabási and Albert (1999) is that scale-invariant degree distributions result from the growth of networks subject to preferential attachment. More specifically, they assume that the probability $P(k_v)$ that a new vertex will be connected to vertex v is a function of the connectivity k_v of v (w runs over all vertices already inserted into the graph):

$$P(k_v) = \frac{k_v}{\sum_w k_w} \tag{7}$$

In several experiments, Barabási and Albert (1999) show that networks which grow according to this model evolve into "a scale-invariant state" in which node connectivity is distributed according to a power law with an exponent $\gamma = 2.9 \pm 0.1$. It is worth notifying that the BA model does not

produce networks which obey the WS model — such a combined model was proposed by Steyvers and Tenenbaum (2005) (see Section 3).

Although this model overcomes a central aspect of invariability of the WS model, it is open to many objections as it disregards other aspects of network dynamics. Amongst others, this relates to the fact that networks grow by the number of vertices *and* edges which may also decrease or stagnate if their birth and death rates accord. Another reference point for revising the BA model is its assumption that the choice of nodes to be linked with newly added ones solely depends on the connectivity patterns of the former. Actually, it is unrealistic to assume that a new vertex is linked with an old one simply because of the connectivity rate of the latter. Rather, linkage depends on the opportunities of old and new nodes to get in contact at all which, in turn, depends on the contexts in which members of the network can "meet" each other (Watts, 2003). In other words, high connectivity does not automatically mean to be met by newly added members of a network. Moreover, Sigman and Cecchi (2002) exemplified topologically quite different graphs which, nevertheless, share the same degree distribution. In this sense, the BA model is not selective enough. For a comprehensive mathematical review of the BA model and several alternatives to it see Bollobás and Riordan (2003).

These and related objections led to a stepwise search for further network characteristics which separate them more precisely from purely random graphs. This includes what is called *assortative mixing* and *community structure*.

2.4 Assortative Mixing

Newman (2002, 2003a) proposes a model in which the probability of a link between two nodes depends on the connectivity of both. This model serves to account for social networks in which vertices tend to be linked when they share certain properties, a tendency which is called *assortative mixing*. It reflects what is circumscribed by the expression *birds of a feather flock together*. According to Newman and Park (2003), this principle distinguishes social networks from non-social (e.g. artificial or biological) ones even if both are uniformly attributed as small worlds according to the WS model. Newman (2002) exemplifies this by assortative mixing of vertex degrees. He confirms that the degrees of interlinked nodes are highly positively correlated in the case of social, while being negatively correlated in the case of technical networks (e.g. the Internet) which show *disassortative mixing*. Newman derives a correlation coefficient in order to measure mixing in undirected graphs (for $r(G)$ of *directed* graphs G see Newman 2002, footnote 35):

$$r(G) = \frac{\frac{1}{m}\sum_i j_i k_i - \left[\frac{1}{m}\sum_i \frac{1}{2}(j_i + k_i)\right]^2}{\frac{1}{m}\sum_i \frac{1}{2}(j_i^2 + k_i^2) - \left[\frac{1}{m}\sum_i \frac{1}{2}(j_i + k_i)\right]^2} \in [-1, 1] \tag{8}$$

where i denotes the edge ending at vertices j and k of degree j_i and k_i, respectively, and $m = |E|, G = (V, E)$. Assortative mixing occurs if $r(G) \gg 0$, otherwise, if $r(G) \ll 0$, disassortative is diagnosed.

Although $r(G)$ separates social from other types of networks, it does not explain the emergence of mixing. Like all other coefficients presented so far, it stays on the level of graph indices and, thus, disregards higher order structure formation within complex networks. The starting-point of such an extended view is, as Newman and Park (2003) argue, *community structure*.

2.5 Community Building

The probability of the members of a social network to interact depends on the social groups (e.g. family, association etc.) and contexts (e.g. attending the same concert, waiting for the same metro etc.) in which they commonly participate (Watts, 2003). Sharing group or context membership raises the probability to interact. Thus, agents entering a network do not necessarily have uniform chance of interacting with any of its highly connected members — *in contrast to what is assumed by the BA model.* Analogously, textual manifestations of social interaction are recursively clustered according to the various genres and registers (Martin, 1992) they instantiate. Thus, the probability of an intertextual relation between two texts analogously raises with their common membership in the same or related genres or registers. The models presented so far do not account for such constraints on linkage within a network.

Newman et al. (2002) take this as a starting point for studying *affiliation networks* in order to overcome this deficit. Affiliation networks are exemplified by networks of collaborating scientists where membership in the same group or context is defined by co-authorship. Affiliation networks are modeled as *bipartite graphs* of group and actor vertices where every actor is linked to the group to which it belongs. This bipartite model is transformed into a unipartite graph in which nodes denote agents who are linked if they commonly belong to at least one group. Finally, the unipartite graph is input to calculating cluster values and average distances as before. A central conclusion of Newman et al. is that compared to random graphs (according to the model of Erdős & Rényi), clustering is always higher in such affiliation networks. The reason is that the higher the number of groups and their extent, the more actor triangles exist within the network. This accords with the expectation, that groups raise the probability of transitive closures, that is, an interaction of vertices v, w which are commonly adjacent to a vertex u of the same group. Another implication is that assortative mixing naturally emerges in networks with community structure although it may also be present in networks without it (Newman, 2003b). This further implies that networks with community structure supersede measuring assortative mixing.

The affiliation model leads back to a network model to which all standard indicators of small worlds are applied. Thus, it does not go far beyond the insights already inherent to the WS model. A more thorough approach to the formation of significantly *recurrent sub-networks* — which may represent, for example, structures of thematically or functionally homogeneous units — is presented by Milo et al. (2002) and Itzkovitz et al. (2003). They explore subgraphs G' of a graph G representing a network which occur more often in G than expected by chance, that is, than in collections of corresponding random graphs of equal order and the same number of edges (Bollobás, 1985) where degree is Poisson distributed and the vertices have the same single-vertex characteristics as in the input graph. A class of such subgraphs is called a *motif*. That is, a motif represents a class of sub-networks whose number is significantly higher within the input network than in its randomised counterpart. One of the observations of Milo et al. (2002) and Itzkovitz et al. (2003) is that exploring the motifs of different networks allow well distinguishing biological, technical and informational networks as they show different patterns of subgraphs recurrent within them although being uniformly attributed as small worlds.

The motif model looks for recurrent network patterns. It is a local model of networking as motifs

represent rather small subgraphs. Further, the motif model does not look for the recursive organisation of such motifs into highly connected modules organised into larger, less cohesive units up to the network as a whole. Such a model has been introduced by Ravasz et al. (2002) and Ravasz and Barabási (2003). They start from the dilemma that while scale-free networks miss any modularity as their hubs provide the predominant part of network cohesion, modular networks fail to have a scale-free degree distribution as they consist of inherently highly connected modules interlinked only by a couple of links so that vertices tend to have a uniform degree. Ravasz et al. present a graph model which has both a modular structure as well as a scale-free degree distribution. This graph model has an inherent hierarchical structure in the sense that it recursively builds around a kernel of a couple of highly clustered vertices more and more peripheral zones which are decreasingly clustered. A central observation of Ravasz et al. (2002) is that such *hierarchical networks* can be distinguished from non-hierarchical, though scale-free networks by the function $C(k)$ of the cluster coefficient C as a function of the degree k. Ravasz et al. (2002) observe that in hierarchical networks $C(k)$ decays — other than in purely scale-free and simply modular networks — as a power law with the degree k, that is,

$$C(k) \sim k^{-\theta} \tag{9}$$

Thus, this model reduces the measurement of structure formation within complex networks to a node-related coefficient. Such a hierarchical, modular network is exemplified by a text network in which modularity is defined by thematic criteria where the central module represents a general topic and where peripheral modules denote topics derived from the topic of their immediate neighbor, more central modules.

As an indicator of structure formation within linguistic networks, θ has been first computed by Ferrer i Cancho et al. (2004). These and related applications of models of complex networks to linguistic networks are reviewed within the subsequent sections.

2.6 Networks Evolving in Time

Apart from the BA model, all network characteristics considered so far focus on static graphs as snapshots of complex networks at certain points in time. In contrast to this, the BA model and its derivations (Bollobás and Riordan, 2003) start from a given set of vertices to which in every subsequent point in time a fixed number of nodes with a fixed number of links is added. Although the underlying model of preferential attachment already allows deriving degree distributions in correspondence to existing networks, this model nevertheless departs from empirical findings in support of what Leskovec et al. (2005) call the *densification* and *shrinking* of evolving networks:

- Firstly, Leskovec et al. (2005) observe that complex networks as, e.g., (scientific or patents) citation networks (cf. Section 3.4) tend to become more and more dense over time. This means that the average degree of their vertices is increasing with the aging network. Interestingly, Leskovec et al. (2005) successfully adapt a power law to this growth process with a positive exponent $1 < \alpha < 2$

  $$e(t) \sim n(t)^{\alpha} \tag{10}$$

  where $e(t)$ is the number of edges at time t while $n(t)$ denotes the number of vertices at that point in time. Leskovec et al. (2005) speak of a *densification* or *growth power law*.

- Secondly, they find that what they call the *effective diameter* is decreasing over time. The effective diameter of a network is defined by means of the cumulative distribution of distances between connected nodes of a network: If $(d, \#(d))$ is the ordered pair of the number $\#(d)$ of vertices in the graph induced by the focal network which are at most d edges separated from each other, the effective diameter of this network is the number d_{eff} of vertices for which $d_{\text{eff}}/n = 0.9$ (i.e. 90% of the vertices in the graph — $n = |V|$).

As it is possible to generate networks which only have one of these two characteristics, it is worth considering them separately in empirical studies. Moreover, insofar as these characteristics depart from assumptions underlying traditional small world models, they give reason to reconsider and further develop the apparatus of complex network analysis in terms of time-dependent models — cf. Leskovec et al. (2005) for two models of such processes. With the accessibility of document networks as, for example, wiki-based systems (cf. Section 3.6) which make any change of their textual nodes and links transparent, this diachronic turn becomes a realistic endeavor of corpus linguistic analyses of complex document networks.

2.7 Summary

The progression of the models discussed in the last four sections mirrors a gradual revision of assumptions about constraints on vertex connectivity and structure formation in networks. Starting from the WS model which does not reflect constraints on degree distributions, extensions regarding aspects of network growth and community structure were discussed. For the time being, small-world formation is indicated by "sparsity, a single connected component containing the vast majority of nodes, very short average distances among nodes, high local clustering, and a power law degree distribution [...]." (Steyvers and Tenenbaum, 2005, 54). For alternative models of structure formation in large networks see Newman (2003b) and Bornholdt and Schuster (2003). These models were initially developed in order to analyse social, biological and technological networks, but also to analyse linguistic networks. The question is whether there exist principles of linguistic networks which can be explored by complex network analysis — by analogy to the Zipfian nature of many frequency distributions of linguistic units explored in quantitative linguistics. The following section reviews studies which deal, directly or indirectly, with this question.

3 Models of Networking of Linguistic Units

Graph	Source Network	Vertex	Edge	Orient.	$\|V(G)\|$	$\epsilon(G)$	$L(G)$	$C_{\mathrm{BR}}(G)$	$C_{\mathrm{WS}}(G)$	γ	$r(G)$	Reference
association graph	free-association data	word	association	undir.	5,018	22.0	3.04	n.s.	0.186	3.01	n.s.	Steyvers and Tenenbaum (2005)
association graph	free-association data	word	association	dir.	5,018	12.7	4.27	n.s.	0.186	1.79	n.s.	Steyvers and Tenenbaum (2005)
citation graph	ISI citation network	bibliographic record	citation (citing)	dir./bipar.	$1,099,017$	4.437	n.s.	n.s.	n.s.	$\gamma_{\mathrm{out}} \approx 3.5$	n.s.	van Raan (2005)
citation graph	ISI citation network	reference	citation (cited)	dir./bipar.	4,876,752	3.14	n.s.	n.s.	n.s.	$\gamma_{\mathrm{in}} \approx 3.1$	n.s.	van Raan (2005)
collocation graph	BNC corpus	word	collocation	undir.	460,902	70.13	2.67	n.s.	0.437	$1.5/2.7$	n.s.	Ferrer i Cancho and Solé (2001)
collocation graph	`wortschatz.uni-leipzig.de`	word	collocation	undir.			3.8	n.s.	0,05		n.s.	Heyer et al. (2006)
concept graph	WordNet	word	sense relation	undir.	122,005	S. 3.3	10.56	n.s.	0.0265	3.11	n.s.	Steyvers and Tenenbaum (2005)
co-occurrence graph	BNC corpus	word	co-occurrence	undir.	478,773	74.2	2.63	n.s.	0,687	$1.5/2.7$	n.s.	Ferrer i Cancho and Solé (2001)
newspaper graph	Süddeutsche Zeitung 1997	newspaper article	*quod vide* link	undir.	87,944	24.78	4.245	0.664	0.684	0.1146	0.699	Mehler (2006)
sentence graph	Czech tree bank	word	dependency relation	undir. i.a.	33,336	13.4	3.5	n.s.	0.1	≈ 2.29	0.06	Ferrer i Cancho et al. (2004)
sentence graph	German tree bank	word	dependency relation	undir. i.a.	6,789	4.6	3.8	n.s.	0.02	≈ 2.23	0.18	Ferrer i Cancho et al. (2004)
sentence graph	Romanian tree bank	word	dependency relation	undir. i.a.	5,563	5.1	3.4	n.s.	0.09	≈ 2.19	0.2	Ferrer i Cancho et al. (2004)
thesaurus graph	Moby's thesaurus	word	sense relation	undir.	30,244	59.9	3.16	n.s.	0.53	S. 3.3	n.s.	Motter et al. (2002)
thesaurus graph	Roget's thesaurus	word	sense relation	undir.	29,381	S. 3.3	5.60	n.s.	0.875	3.19	n.s.	Steyvers and Tenenbaum (2005)
web graph	search engine crawl	website	hyperlink	undir.	153,127	n.s.	3.1	n.s.	0.1078	n.s.	n.s.	Adamic (1999)
web graph	search engine crawl (SCC)	website	hyperlink	dir.	64,826	n.s.	4.228	n.s.	0.081	n.s.	n.s.	Adamic (1999)
web graph	search engine crawl (SCC)	`.edu` website	hyperlink	dir.	3,456	n.s.	4.062	n.s.	0.156	n.s.	n.s.	Adamic (1999)
wiki graph	German Wikipedia	wiki entry page	hyperlink	undir.	303,999	19.39	3.247	0.01	0.223	0.4222	-0.1	Mehler (2006)
wiki graph	German Wikipedia	wiki entry page	hyperlink	undir.	406,074	15.88	3.554	0.01	0.186	0.5273	-0.09	Mehler (2006)
wiki graph	German Wikipedia	wiki entry page	hyperlink	undir.	796,454	11.50	4.004	0.007	0.139	0.7405	-0.05	Mehler (2006)
wiki graph	`wiki.apache.org/jakarta`	wiki entry page	hyperlink	undir.	916	23.84	4.488	0.193	0.539	0.2949	-0.5	Mehler (2006)
wiki graph	`wiki.apache.org/struts`	wiki entry page	hyperlink	undir.	1,358	29.93	4.530	0.162	0.402	0.2023	-0.45	Mehler (2006)
wiki graph	`wiki.apache.org/ws`	wiki entry page	hyperlink	undir.	1,042	22.91	4.541	0.175	0.485	0.1989	-0.48	Mehler (2006)
wiki graph	11 Wikipedia releases	wiki entry page	hyperlink	dir.	n.s.	n.s.	4.53	cf. ref.	n.s.	$\gamma_{\mathrm{in}} \approx 2.15$	≈ -0.1	Zlatic et al. (2006)
wiki graph	11 Wikipedia releases	wiki entry page	hyperlink	undir.	n.s.	n.s.	3.32	cf. ref.	n.s.	$\gamma_{\mathrm{un.}} \approx 2.35$	≈ -0.1	Zlatic et al. (2006)

Table 4: Studies of complex networks of linguistic units.

In this section, small-world models are reviewed which focus on linguistic networks, that is, on graphs whose vertices represent, for example, words, sentences or texts. Table 4 summarises these approaches w.r.t the criteria of networking they apply and the network characteristics they compute. By the majority, these approaches analyse WWW-based graphs whose vertices represent *web pages* and whose edges stand for hyperlinks. The remaining set of approaches concentrates on networks spanned by lexical or sentential units and their lexical or syntactical relations. Generally speaking, all these approaches should (but often fail to) answer the following questions:

1. *What are the criteria of network formation?* In other words: *What do the vertices represent and subject to which criteria are they linked?*
2. *What is the reason of network analysis?* In other words: *Why are the networks analysed or what is the research interest in analysing these networks?*
3. *Which small-world or complex network indicators are investigated?*
4. *Which reasons are assumed to evoke the small-world property if observed?*
5. *Is there any account of network growth or of any other aspect of network dynamics?*

The review is ordered by increasing complexity of the signs denoted by the nodes of the networks: It starts with lexical networks in order to approach textual and document networks via so-called sentence networks.

3.1 Co-Occurrence Graphs and Collocation Graphs

Collocation analysis is a well established field of corpus linguistics (Sinclair, 1991; Stubbs, 1996, 2001). It follows the Firthian tradition according to which collocations manifest lexical semantic affinities beyond grammatical restrictions (Halliday, 1966). Collocation analysis aims at discovering semantically related words based on (e.g. similarity) functions of their co-occurrences. In computational linguistics, several measures exist for distinguishing collocations from insignificant, though recurrent co-occurrences (Manning and Schütze, 1999). Starting from pairwise computing lexical affinities by means of such measures, the network perspective is obvious: If two units a, b are related in terms of collocation statistics as the units b, c are, an indirect relation between a and c is implied even if not directly confirmed by a collocation of a and c – by analogy with semantic networks (cf. Section 3.3). Following this procedure, a network of units linked by collocation arises whose graph theoretical representation will, thus, be called *collocation network*. Following this line of argumentation, several approaches analyse the *topology of large collocation networks* (Dorogovtsev and Mendes, 2001; Ferrer i Cancho and Solé, 2001; Heyer et al., 2006). These networks are seen to be partitioned into a kernel lexis and more peripheral sociolects or topic specific terminologies. In such networks, lexical units are not immediately related to every other unit. Rather, there is mediation by means of common words of the kernel vocabulary in the role of hubs (Kleinberg, 1999) or long-range nodes which have connections to many local word clusters and, thus, interrelate the different fields of lexis (Tuldava, 1998). Moreover, the word clusters are themselves seen to be highly interwoven so that short paths emerge (Bordag et al., 2003). From this perspective, lexis is seen as a complex network which is based in part on collocational regularities (i.e. beyond sense relations) and, thus, can be asked for its SW properties.

A first experiment in this area is described by Ferrer i Cancho and Solé (2001) who analyse the British National Corpus (BNC) from which they extract two graphs:

- Firstly, a so-called *co-occurrence graph* G_1 in which words are linked if they co-occur in at least one sentence within a span of maximal three tokens — see also Widdows and Dorow (2002) who explore the BNC corpus in order to extract a co-occurrence graph whose extraction is constrained by means of PoS relationships.
- Secondly, a *collocation graph* G_2 is extracted in which only those links of G_1 are retained whose end vertices co-occur more frequent than expected by chance.

Generally speaking, a co-occurrence graph is a graph whose edges represent single co-occurrence events of word forms without abstracting over sets of alike events. In contrast to this, a collocation graph is a graph whose edges represent significant co-occurrences where significance is attributed according to evaluating some set of such events by means of some collocation measure.

Ferrer i Cancho and Solé observe the small-world property in the case of both networks — according to the WS model and the BA model (see Table 4). But other than according to the BA model, they separately fitted a power law to the degree distribution of the so-called kernel vocabulary (including the 5,000 topmost connected vertices) for which they yielded an exponent γ closer to the range of values predicted by the BA model. Dorogovtsev and Mendes (2001) took this empirical finding as a starting point and developed a theoretical model of network growth which reproduces both power laws with a greater exponent of the law fitted to the kernel vocabulary.

By analogy with the model of Ferrer i Cancho and Solé, Bordag et al. (2003) analyse a collocation graph extracted from a German corpus of newspaper articles (cf. `wortschatz.uni-leipzig.de`). Other than Ferrer i Cancho and Solé, they apply a log-likelihood-related measure for exploring collocations based on sentence co-occurrences. Their findings confirm the small-world property of the collocation graph being analysed. The numerical results of this study are published in Heyer et al. (2006).

3.2 Sentence Graphs

In the last section, co-occurrence graphs were described as a special case of lexical networks in which words, whose co-occurrences are observed in a given input corpus, are linked. These graphs were used as a starting point for deriving collocation graphs by retaining only those edges which manifest collocations (i.e. significant co-occurrences in the sense of some appropriate statistical measure). Another point of departure in dealing with co-occurrence graphs is to consider only those co-occurrences which manifest syntactic dependency relations, e.g. between the verb of a sentence and a noun manifesting its subject. This is the basic building principle of so-called *sentence graphs* (Ferrer i Cancho et al., 2004): A sentence graph is a directed graph in which vertices represent lexical units which are linked if they co-occur at least in one sentence in the role of a modifier (source vertex) and head (target vertex), respectively. In corpus linguistics, sentence graphs have already been analysed by Hoey (1991) who spans networks of sentences which are linked if they share at least two lexically cohesive words.

In the experiments reported by Ferrer i Cancho et al. (2004), directed edges are only considered

w.r.t power law fitting. As links represent syntactic dependency relations, each input sentence induces a subgraph of the sentence graph, thus justifying its name.

In order to determine the properties of such networks, Ferrer i Cancho et al. (2004) analyse tree banks of Czech, German and Romanian sentences annotated w.r.t their dependency structure. They show that sentence graphs have the small-world property according to the WS and the BA model. Additionally, Ferrer i Cancho et al. compute the clustering coefficient $C(k)$ as a function of the degree k. In Ravasz et al. (2002), the distribution of C over k was analysed as an indicator of latent hierarchical structures within networks — see Section 2.5. Ferrer i Cancho et al. do not observe this property in the case of sentence graphs. That is, $C(k)$ does not decay according to a scale-free power law, although being highly skewed. As a further indicator of structure formation, Ferrer i Cancho et al. observe disassortative mixing. Therefore, highly connected words tend to be linked with lowly connected ones — in accordance with what is expected according to the usage of function words, common nouns etc. An objection against the notion of a sentence graph is the yet unproven conjecture that their structure is a trivial consequence of patterns inherent to the input sentences so that their analysis is intrinsically obsolete as the sentences may be analysed in isolation. In general, network analysis has to prove that reductions of this kind are impossible, that is, that the small-world property and related characteristics only emerge in the network as a whole.

An extension of the model presented in Ferrer i Cancho and Solé (2001) and Ferrer i Cancho et al. (2004) is elaborated in Ferrer i Cancho et al. (2005) which starts from a simplified bipartite "form-meaning" graph in order to derive a network of linguistic units.

3.3 Concept Graphs, Thesaurus Graphs and Association Graphs

Whereas collocation graphs directly build on observable co-occurrences of lexical units in large text corpora, *lexical reference systems* or *terminological ontologies* (e.g. WordNet), *thesauri* (e.g. Roget's thesaurus) and related systems build — sometimes additionally — on expert knowledge of lexicographers in order to define *sense relations* (e.g. synonymy, antonymy, homonymy) between words or *conceptual relations* between concepts (e.g. hypernymy, co-hyponymy, metonymy). As in the case of collocation graphs, but other than in the case of co-occurrence and sentence graphs, sense relations are meaning-based. The difference of collocation graphs and the type of networks to be surveyed in this section relates to the distinction made by Halliday and Hasan (1976) between unsystematic lexical cohesion based on collocation and systematic lexical cohesion based on sense relations.

An alternative source of exploring meaning-based relations of lexical units which relate to, but are not identical with collocations, are regularities of *association* or, more specifically, *word priming*: In the case of word priming, lexical units are used as primes in order to let test persons associate sense or form-related words (Kintsch, 1988). In the line of this argumentation, association graphs of lexical units are built whose vertices represent primes and responses linked from the former to the latter.

Based on these preliminary considerations, the following graphs can be distinguished:

- *Thesaurus graphs* are graphs in which vertices denote words, whereas edges represent sense relations thereof (cf. Kinouchi et al., 2002).
- In contrast to this, *concept graphs* are graphs in which vertices represent concepts, whereas edges

denote conceptual relations thereof.

- Finally, *association graphs* are graphs in which vertices denote words — as in the case of thesaurus graphs —, whereas edges represent association or priming relations as observed in cognitive-linguistic experiments.

As these preliminary considerations motivate a network perspective on sense relations and association data, questions w.r.t structure formation within such networks and their overall topology likewise arise:

1. In the case of thesaurus graphs based on the expertise of lexicographers and corpus linguists, respectively, network properties can be interpreted as indicators of thesaurus quality or consistency. As networks of this kind represent lexical semantic knowledge of a given language, their analysis also provides an access to the semantic system of that language, that is, to the overall organisation of its lexical subsystem (Sigman and Cecchi, 2002).
2. In the case of association networks, a corresponding argumentation applies: According to the hypothesis that association is one of the principles of memory organisation, the question is raised which network topologies support an efficient organisation in terms of time and space complexity. This is the starting point of Motter et al. (2002) who interpret the small-world property of association networks as an indicator of efficient information storage and retrieval (cf. Sigman and Cecchi 2002; Steyvers and Tenenbaum 2005): Firstly, the existence of many local clusters is seen as a necessary condition of effective associations. Secondly, the existence of short path lengths is seen to guarantee fast information search (or spreading activation) since any "pieces of information" are on average separated only by a couple of associations — *irrespective how different they are*.

In addition to collocation graphs, these two research directions — the more language- and the more memory-oriented one — leave the narrow view on word-to-word relations in order to focus whole networks thereof and, thus, lexical subsystems based on corpus-linguistic collocations, cognitive associations or lexicographical sense relations. This section reviews these kinds of approaches. First of all, this includes the study of Motter et al. (2002) who analyse the so-called *Moby thesaurus* in the form of its e-text release as part of the Project Gutenberg (cf. `ftp://ibiblio.org/pub/docs/books/gutenberg/etext02/mthes10.zip`). Motter et al. extract an undirected graph from this thesaurus in which vertices represent *root words* which are linked if the one word occurs in the root word list of the other (cf. also Holanda et al., 2003). As shown in Table 4, network analysis indicates that this thesaurus has the small-world property. Regarding scale-freeness, Motter et al. do not directly fit a power law, but observe a crossover from a more exponential behavior of $P(k)$ to a more power law-like behavior for higher values of k (with $\gamma = 3.5$). Albert and Barabási (2002) report on a related experiment of Yook, Jeong & Barabási in which a thesaurus graph (which also shows the small-world property) is extracted from the *Merriam-Webster dictionary* by exploring synonymy relations.

Sigman and Cecchi (2002) extract a concept graph from WordNet (Miller et al., 1990). They extract a graph of *lexicalised concepts* or *word meanings* in which vertices stand for synsets and edges for their meaning relations — 'synset' is a short form of *synonym set* representing a single word meaning (cf. Miller et al., 1990). In such a graph, edges are typed according to the meaning relation they represent.

Sigman and Cecchi take *antonymy*, *hypernymy*, *meronymy* and *polysemy* relations into account — note that antonymy and polysemy relations are symmetric, whereas hypernymy and meronymy relations have *hyponymy* and *holonomy* as their inverses. Nevertheless, the concept graph extracted by Sigman and Cecchi is an undirected graph whose vertices solely represent *noun* meanings.

Generally speaking, hyponymy relations induce a kernel hierarchical structure of the concept graph extracted from WordNet. This hierarchical skeleton is superimposed by polysemy relations defined between any word meanings whose synsets share at least one (polysemous) word form. Sigman and Cecchi explore these relations as an additional source of edge generation. Their findings indicate that the inclusion of polysemy relations convert the concept network into a small world whose degree distribution follows a power law and in which subgroups of fully connected meanings emerge — for numerical details see Table 4. This result fails to appear when antonymy or meronymy relations are added to the hierarchical skeleton instead of the polysemy relations. Sigman and Cecchi conclude that polysemy has the effect of generating the small-world property and view this to be an explanation of its emergence in natural language. In other words: Polysemy is seen to convert hyponymy-based concept networks into compact, clustered graphs which allow efficient storage and retrieval of lexical knowledge.

A comprehensive network analysis of lexical-semantic units is performed by Steyvers and Tenenbaum (2005). They analyse networks based on *Roget's thesaurus*, WordNet and *free-association data* of lexical units:

- *Experiment I:* Steyvers and Tenenbaum start with an *association graph* whose vertices denote cue and response words which are linked whenever at least two participants of the underlying free-association experiment associated the same response to the same input cue. This graph is by definition simple and undirected. Steyvers and Tenenbaum derive a directed graph from this graph by means of its orientation along the association from the cue to the response word. Thus, *two* variants of an association graph are analysed.

- *Experiment II:* As *Roget's thesaurus* defines a bipartite graph whose *top-mode* vertices represent semantic categories and whose *bottom-mode* vertices stand for words, Steyvers and Tenenbaum derive a unipartite *thesaurus graph* thereof in which words are linked whenever they are commonly classified by at least one category. Roget's thesaurus is also explored by Leicht et al. (2006) for the task of complex network analysis.

- *Experiment III:* Essentially, WordNet also has a bipartite structure based on the many-to-many relation of word forms and synsets. Thus, word form-to-word form edges (denoting, for example, antonymy relations) have to be distinguished from word form-to-synset and synset-to-synset edges (representing, for example, hypernymy or meronymy relations). Steyvers and Tenenbaum explore this bipartite structure in order to extract a unipartite graph of vertices denoting word forms (albeit they separately display $\langle d \rangle$ of the set of synsets). As Steyvers and Tenenbaum extract a graph of lexical, but not of conceptual nodes, it is not a concept graph — nevertheless we will retain this label as the predominant source of linkage are semantic relations as represented by means of links of synsets.

  Steyvers and Tenenbaum compute the average degree, average geodesic distances of all vertices

(experiment I) or a sample of 10,000 vertices (experiment II and III), diameters, cluster values C_{WS} and power law exponents γ for all four input graphs — see Table 4 for the results. All computations were made for the largest strongly connected component which covered at least 96% of the vertices. All networks demonstrate the small-world property, according to the WS model as well as according to the BA model. But Steyvers and Tenenbaum (2005) go beyond the present apparatus of complex network analysis as they develop a model of network growth which departs from the BA model of Barabási and Albert (1999) in that it additionally focuses on cluster formation as observed according to the WS model. In order to do that, they start from a linguistic assumption on the linkage of words newly added to a network at a given point in time. This model hints at promising extensions of the BA model from the point of view of (cognitive) linguistics and, thus, may serve as a starting point for critically extending complex network analysis in the light of linguistic research.

All studies surveyed so far focused on networking of linguistic units *below the text level*. The remaining three sections review studies which analyse text or document networks instead.

3.4 Citation Graphs and Sitation Graphs

The quantitative study of networks of scientific documents linked by bibliographic relations is one of the earliest approaches to document networking (Garfield, 1963; de Solla Price, 1965). This field of research is separated into *informetrics*, *bibliometrics*, *scientometrics* and *webometrics* depending on the provenance of the not necessarily textual units whose linkage is studied (cf. Björneborn, 2004):

- *Informetrics* is the most encompassing field of applying quantitative methods to studying processes of information transfer in networks of whatever information units — irrespective of the underlying transfer medium (Ravichandra Rao, 1996).
- In contrast to this, *bibliometrics* "is the quantitative study of literatures as they are reflected in bibliographies" (White and McCain, 1989, 119). Other than webometrics, it is mainly based on *printed*, but not on WWW data (Bar-Ilan, 2001).
- *Scientometrics* is a kind of bibliometrics with a focus on scientific communication. It aims at evaluating the impact factor of scientists, scientific discoveries or publication media. Further, it explores topological regularities of scientific document networks and maps the topological relatedness of authors, documents and publication media in order to derive recommendations for improving the retrieval of scientific publications (cf. Hummon and Doreian, 1989; Larson, 1996; Ravichandra Rao, 1996).
- With the advent of the WWW, the hyperlink-based classification of web documents became a further research topic not only in *web mining*, but also in *webometrics*. Its basic idea is to apply the methodical apparatus of bibliometrics to web documents and their hyperlinks by analogy with scientific documents and their citation relations — in spite of the many differences due to possibly bidirectional hyperlinks, the lack of peer reviewing and of knowledge about the motives of hyperlinking (for a critical review of this concept cf. Prime et al., 2002). According to Björneborn and Ingwersen (2001), webometrics aims at exploring the regularities of the content and structure of web documents and, thus, is connected to web *content* and *structure* mining (Kosala and Blockeel, 2000).

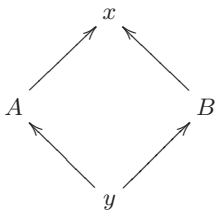

Figure 4: Two fundamental relations within citation networks according to Fang and Rousseau (2001): the *bibliographic coupling* of A and B via x and the *co-citation* of A and B via y.

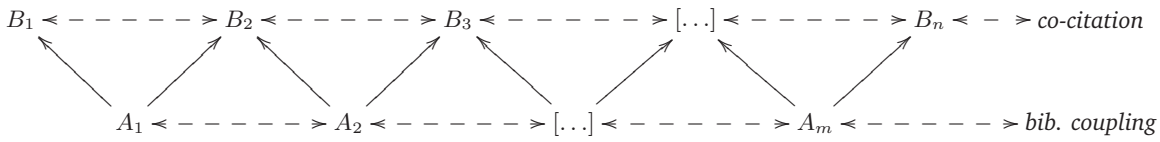

Figure 5: Co-citation and bibliographic coupling chains (cf. Björneborn, 2004).

The present section focuses on the scientometric study of networks of scientific documents (e.g. conference papers, journal articles, reviews, book chapters, scientific notes etc.) as part of scientific communication. These approaches are split into two groups of which only the second will be reviewed — for an overview of computational linguistic approaches to the first group cf. Leopold (2005):

- Firstly, there exists a group of approaches which explore the vocabularies of texts as a source of intertextual linkage (Leydesdorff, 2001).
- Secondly, there is the group of approaches which explore citation and other reference relations as a source of intertextual linkage provided that they are explicitly marked within the input texts.

The latter approaches deal with *referential intertextuality* in the sense of Heinemann (1997), while the former can be said to focus on *typological intertextuality*.

Input to the scientometric way of network analysis are so-called *citation networks* in which nodes denote scientific publications which are linked from the citing to the cited publication. By example of the `OpCit` project (`opcit.eprints.org`) based on the *Los Alamos Eprint Archive* (LANL, cf. `xxx.lanl.gov`), Harnad and Carr (2000) speak of "citation linked online digital" corpora which can easily be made an object of complex network analysis. Nevertheless, it is worth noting that the majority of these approaches explore bibliographic records (White and McCain, 1989) as collected, for example, by the *Institute for Scientific Information* (ISI; cf. Garfield (1994), `scientific.thomson.com/free/essays/citationanalysis/scientography` or `www.isinet.com`) and, thus, metadata descriptions without exploring the underlying documents directly — unless approaches to referential and typological intertextuality are amalgamated as, for example, by Glenisson et al. (2005). The ISI integrates the Science Citation, the Social Science Citation and the Arts & Humanities Citation Index. It covers articles, notes, letters, reviews, editorials, corrections, meeting-abstracts and related document types of scientific communication (Sigogneau, 2000). Note that, besides proceedings, scientometrics only occasionally analyses citation relations of books, but tends to concentrate on scientific articles and related types of publications (White and McCain, 1989). WWW-based resources of citation networks are digital libraries as, for example, CiteSeer, CiteBase or, as a kind of social software-based

digital library, CiteULike.

Citation networks allow distinguishing two fundamental scientometric relations as exemplified in Figure 4:

- First of all, two documents A and B are said to be *bibliographically coupled* if there exists at least a third document x commonly cited by A and B (Kessler, 1963). In terms of scientometrics: two *citing items* are bibliographically coupled if they have at least one reference (i.e. *cited item*) in common (Glänzel and Czerwon, 1996). Coupling relations induce *bibliographic coupling networks* or *BC networks* for short. These networks are induced by analogy with scientific collaboration or co-author networks (van Raan, 2005) in which author nodes are linked if they co-authored at least one publication. Within BC networks, link weighting naturally arises from counting and appropriately standardising the number of common references of two publications. BC graphs are inferred from BC networks by means of a bipartite graph model in which the top mode consists of vertices representing so-called *references* cited by *items* represented by vertices of the bottom-mode — see Figure 5. This allows interlinking any two *bibliographically coupled* publications whenever the corresponding vertices are linked to the same top-mode vertex. Next, publications sharing the same reference are collected into so-called *BC clusters* — by analogy with collaboration clusters in co-author networks whose affinity is affected by at least one publication whose authorship they share. Further, *BC chains* are paths $(v_{i_0}, e_{j_1}, v_{i_1}, e_{j_2}, \ldots, v_{i_{n-1}}, e_{j_n} v_{i_n})$ in which neighboring vertices $v_{i_j}, v_{i_{j+1}}$, $j \in \{0, \ldots, n-1\}$, denote bibliographically coupled items — see Figure 5.

  As references are nothing else but publications, BC networks are — in spite of their analogy to co-author networks — homogeneous in the sense that they consist of nodes of the same sort. Note further that any inference of a BC network necessarily includes two time windows: the window of those publications whose references are studied and the window spanned by the publication dates of these references. Redner (1998), for example, analyses a one year time window of references cited by publications within the years 1981-1997. In contrast to this, van Raan (2005) analyses a one year time window of publications and their references within the preceding publication years.

- A second type of relation is co-citation: two documents A and B are said to be *co-cited* if there exists at least a third document y commonly citing A and B (cf. Small, 1973) — see Figure 4. By analogy with BC networks, co-citation relations induce *co-citation networks* or *CC networks* for short. Further, CC graphs are inferred from the same bipartite model as BC graphs, but with the reverse perspective on the top-mode units. Likewise, *CC clusters* and *chains* are defined according to the same analogy — see Figure 5.

Obviously, citation relations are not symmetric and, thus, induce *directed* edges, while co-citation and bibliographic coupling are symmetric and, therefore, are represented by *undirected* edges. Further, edges representing co-citation or bibliographic coupling relations are straightforwardly weighted by means of functions of the number of their occurrences (Small, 1999). Note also that citation patterns vary among scientific communities so that findings in one of them do not necessarily characterise another one adequately. This observation is important when it comes to normalise numerical indicators in order to secure the comparability of different communities (Pinski and Narin, 1976).

In citation networks, the *in-degree* distribution is induced by the number of times a publication is cited, whereas the *out-degree* distribution is induced by the number of citations a publication is making. In this context, Redner (1998) distinguishes two types of distributions within citation networks related to the scale-freeness of complex networks:

- firstly, the so-called Zipfian *rank-frequency* distribution of the number of citations ranked in decreasing order where the first rank denotes the most cited publication down to those publications which are only cited once or not at all;
- secondly, the *size-frequency* distribution of the number $N(x)$ of publications with x citations which relates to the Lotka law of scientometrics which is a power law $y_x = c/x^\alpha$ where y_x is the number items with x occurrences of the focal type (Rousseau and Rousseau, 2000).

  In the case of co-citation networks, successful fitting a power law to a size-frequency distribution indicates that the majority of (pairs of) documents are not co-cited at all, many others are co-cited only once etc. till the region of those hubs are reached which are co-cited very often. Co-citation analyses often count only co-citations whose frequencies are beyond a certain threshold above which they are regarded as a significant source of (topical) relatedness (White and McCain, 1989).
- Additionally, a cumulative-frequency distribution can be considered which relates to Bradford's law in scientometrics.

In the case of the rank-frequency distribution, Redner (1998) fits a power law with exponent $\gamma \approx -0.5$ for the part of the distribution from rank 1 ($8,904$ citations) to rank 12,000 (of about 85 citations). For large x, this corresponds to fitting a power law $N(x) \propto x^{-\alpha}$ to the corresponding size-frequency distribution with $\alpha \approx 3$. These findings indicate that nearly one half of the papers is un-cited. In the area of webometrics which analyses hyperlinks by analogy with citations as sitations (see below), Prime et al. (2002) find three regimes of the out-degree distribution of external links within web pages which allow distinguishing, firstly, general portals (first regime) with a high number of external links from, secondly, more specialised portals and, thirdly, from web documents of varying size with few internal and external links. Prime et al. successfully fit a power law to the distribution of incoming links, that is, to the in-degree distribution of vertices representing web pages.

BC networks are analysed, for example, by van Raan (2005) — see Table 4. He explores a corpus of $1,099,017$ publications (manly articles, reviews, notes and letters) citing 4,876,752 references. Van Raan considers three degree distributions: Firstly, he adapts a power law to the in-degree distribution of references, that is, to the distribution of the number of citing publications per cited reference. Secondly, he considers the number of references per publication in order to adapt an out-degree distribution in the sense of the bipartite graph model presented in Figure 5. Thirdly, he studies the number of bibliographically coupled documents per publication. The central observation of van Raan is that within this corpus, power laws are only successfully fitted if all references are taken into account, while scale-freeness disappears when references are separately accounted by their age. He concludes that references to "younger" publications have other functions (e.g. including topic specific references) than references to older ones (e.g. referring to "classic" works in the corresponding field). This observation points to the significance of time dependent, nonstationary characteristics of document networking. In this sense, Seglen (1992) describes changes of citedness of articles dependent on their

age where “citedness declines steadily as a function of time since publication” indicating increasing obsolescence where only a few articles are accepted as classic work for a longer time period.

Citation, BC and CC networks are easily made input to multivariate statistics. Based on matrices of co-citation or coupling frequencies, correlation coefficients of documents can be computed as input to cluster or principle components analysis in order to derive clusters or factors, respectively, which represent sets of topically or otherwise related documents (Larson, 1996). Any such cluster can then be described in terms of its size, the mean year of publication of its elements or according to the distribution of these elements over the set of scientific genres. That is, citation, BC and CC networks are input to deriving large sub-networks thereof based on publication medium-, genre-, document type- or topic-related criteria. Hummon and Doreian (1989) analyse, for example, a citation network which solely consists of publications on DNA theory. Likewise, Schummer (2004) analyses citation networks incorporating articles on nanotechnology only. Further, citation networks may be confined to certain publication media (e.g. journals or proceedings) — leaving out other media of scientific communication (cf. Seglen, 1992) — or certain pragmatic parameters (e.g. common author, time or location of production etc.) (cf. Sigogneau, 2000).

Based on these considerations, the following approaches can be distinguished which focus on structure formation derived from citation relations: Glänzel and Czerwon (1996) present a method for identifying research topics as clusters of bibliographically coupled items. They define *core documents* as items with a higher-than-average number of links. Core documents are analysed w.r.t their distribution over journals, scientific subfields and corporate addresses (e.g. of universities). In summary, Glänzel and Czerwon analyse the formation of sub-networks within BC networks and their *macro*-level segmentation according to the criteria just mentioned. Identifying (mainstream) research topics and exploring their life cycle are further parts of this agenda. A further topic is to identify so-called sleeping beauties (van Raan, 2004), that is, publications which remain unnoticed for a longer period of time and then, suddenly, get cited to a high degree. A network perspective on the contributions to conference series is provided by Chen and Czerwinski (1998) who apply Latent Semantic Analysis (Landauer and Dumais, 1997) in order to automatically link topically related documents. Chen (1999) applies this apparatus to co-citation networks in order to explore predominant research fields as sub-networks. Small (1999) analyses co-citation chains in order to examine cross-disciplinary citations interrelating document networks of different scientific disciplines. He applies cluster analysis for recursively identifying sub-networks of high co-citation density which are traversed by means of *pathways* (Björneborn and Ingwersen, 2001) of cross-disciplinary citations.

A further research topic is the temporal dynamics of scientific communication. This relates, for example, to studying the life cycle of scientific fields based on their citation relations (Otte and Rousseau, 2002). As the set of scientific publications is rapidly growing, citation-based clusters are continually shaped by newly entering or dropping out documents. They may merge or split into new clusters or may even disappear completely (White and McCain, 1989).

Distributions of URL-based references to web documents within scientific articles are studied by Brown (2004). He gives a perspective on integrating document networks within more traditional and online media and, thus, leads over to the study of so-called *sitations*. In webometrics, hyperlinks between web documents (e.g. pages or websites) are analysed by analogy with citation relations as sitations (Rousseau, 1997; Faba-Pérez et al., 2003; Björneborn, 2004). Rousseau (1997) shows that

Level	Unit of Research	Approaches
macro level	network	complex network analysis
meso level	sub-network	exploring web communities, broad topics etc.
micro level	web document	segmentation & categorisation of websites and pages

Table 5: Reference points of document network analysis by example of the WWW.

the distribution of sitations obeys a power law with an exponent of 2.345. He analyses the distribution of the number $N(x)$ of websites with x sitations, that is, with x inlinks. Earlier, Larson (1996) applied co-citation analysis in order to cluster topically related pages. Inlink distributions are, further, analysed by Thelwall and Tang (2003), Tang and Thelwall (2004), Li et al. (2005a,b) who concentrate on academic websites in order to measure the impact factor of universities and their departments subject to situational parameters and membership in scientific disciplines. Outlink distributions are analysed by Ajiferuke and Wolfram (2004) who concentrate on web pages of top level domains. More recently, Björneborn (2004) has extend webometrics by defining co-linking and co-linked pages by analogy with bibliographic coupling and co-citation, respectively — see Figure 5. He explores *co-linkage chains* by analogy with *co-citation chains* (Small, 1999) in order to identify pathways of topically related, though *un*-co-linked pages (cf. Garfield, 1994). Björneborn gives a comprehensive scientometric perspective on the study of networking in WWW-based scientific communication and, thus, can be seen as leading over to the study of document networking in the WWW as surveyed in the next section.

3.5 Web Graphs

A well explored area of network analysis is the World Wide Web (WWW). From the beginning of statistical analyses of small worlds on, it has been made an object of this kind of research (cf. the survey of Newman 2003b). Seen as a network of hypertext documents in the form of websites or pages, the WWW is by now the best studied document network. Because of the many surveys of WWW-oriented studies — cf., for example, Chakrabarti (2002) and Baldi et al. (2003) for two excellent books surveying this area — the present section concentrates on a general account of their reference points, that are, *macro*, *meso* and *micro* level units — see Table 5 — as input of what is called *web content* and *structure mining* (Kosala and Blockeel, 2000):

- On the *macro level*, the WWW as a whole is made an object of complex network analysis. The starting point of this kind of research is, more or less explicitly, the seminal paper of Botafogo et al. (1992) on so-called *hypertext graphs*. Botafogo et al. generally describe hypertexts in terms of vertices and edges denoting hypertext modules and their hyperlinks, respectively, in order to analyse their structural characteristics in terms of compactness and hierarchal structure formation. In web mining, this graph-theoretical format is utilised in order to represent the WWW or parts of it by means of so-called *web graphs* (Chakrabarti, 2002), that is, directed graphs whose vertices denote pages and whose edges denote hyperlinks in-between (Björneborn and Ingwersen, 2001) — cf. Park (2003) who speaks of *hyperlink network analysis* when it comes to exploring the WWW as a graph. Although the page level is the accentuated reference point of web-based link manifestation and, thus, of networking in the WWW, its networking may also be observed on the level

of websites and conglomerates thereof when viewed as vertices — cf. the layered graph model of Mukherjea (2000) and especially the document model of Björneborn (2004) and Björneborn and Ingwersen (2004) who describe a graph model in terms of webometrics which bridges analyses of the WWW in general and those of document networks in scientific communication. Generally speaking, macro level studies ask for the principles of the overall topology of the WWW (Barabási et al., 1999) in terms of clustering and geodesic distances (Adamic, 1999), its diameter (Albert et al., 1999), the degree distribution of pages (Barabási and Albert, 1999; Kleinberg et al., 1999; Adamic and Huberman, 2001; Barabási et al., 2000) and the web's characteristic motifs (Milo et al., 2002). A seminal paper in this area is Adamic (1999) who analyses the SW property of the web. He refers to websites as the operative units for vertex extraction where a site A is seen to be linked with a site B if it contains a page linked with a page in B. The resulting graph is analysed in three variants: As an undirected, as a directed and as a subgraph containing solely websites of a certain top level domain (i.e. `.edu`). A purely structural perspective on the overall topology of the WWW is induced by its so-called *bow tie-structure* (Broder et al. (2000)) which segments the WWW into four topological regions of roughly equal size (cf. Baldi et al., 2003): First, the *Strongly Connected Component* (SCC) contains all pages that are reachable from each other by directed paths. In contrast to this, the IN component includes all pages that can reach members of the SCC, but cannot be reached by them in terms of directed paths. Analogously, the OUT component contains all pages that are reachable from the SCC, but are not linked with any member of the SCC. Finally, there are, amongst others, components consisting of pages which are disconnected from the SCC as well as from the IN and OUT component. This model sheds light on the necessity of segmenting sub-networks of the WWW which obviously vary w.r.t their structural characteristics. This is done on a meso level of analysis:

- *On the meso level*, the WWW is studied as a heterogeneous network which does not simply consist of pages and their links, but is clustered into large, functionally as well as thematically heterogeneous sub-networks whose segmentation is the focus of interest on this level. Gibson et al. (1998) and Flake et al. (2000), for example, extract so-called *web communities*, that is, networks of websites which have more links to members of the same community than to sites outside of it. Web communities are induced by exploring the link structure of pages. In contrast to this, Chakrabarti et al. (2002) explore so-called *broad topics* manifested by large clusters of thematically homogeneous web pages whose size distribution they study. See also Mukherjea (2000) who likewise distinguishes sub-networks in terms of thematic criteria. A very interesting observation of Chakrabarti et al. (2002) as well as of Pennock et al. (2002) is that generically or topically demarcated sub-networks of the WWW show strikingly different regularities of their degree distributions in comparison to the WWW as a whole. This observation hints at the genre/register sensitivity of network analyses and, thus, on the necessity to further investigate and extract significant sub-networks of the WWW as the proper input of complex network analysis. Another reference point on the meso level (as a byproduct of inferring web communities and related units) is the classification of vertices of web graphs in terms of so-called *authorities* (i.e. "popular" pages linked by many other pages) and *hubs* (i.e. pages which list links to many authorities) Kleinberg (1999).

- The genre-sensitivity of large scale network characteristics hints at the fact that, by analogy with

texts, hypertexts manifest functionally/thematically demarcated *hypertext types* where instances of the same type tend to be similarly structured, while instances of different types are more likely dissimilarly structured. The general idea is that knowledge about hypertext types and their prototypical instances facilitate hypertext production and reception. This assumption is reflected by the notion of a *web genre* (Dillon and Gushrowski, 2000; Firth and Lawrence, 2003) which is defined in functional terms as a type of web documents serving a certain recurrent function of web-based communication. Manifestations of webgenres and related units are, more or less explicitly, analysed in terms of *compound documents* (Eiron and McCurley, 2003), *logical domains* (Li et al., 2000), *logical documents* (Tajima and Tanaka, 1999; Li et al., 2002) or *multipage segments* (Craven et al., 2000). In the majority of cases, instances of web genres are analysed on the level of single pages (Rehm, 2002). A smaller group of approaches analyses instances of webgenres in terms of websites as systems of pages whose links are likewise analysed in terms of genre-specific functions (Mehler and Gleim, 2006).

This series of approaches of increasing resolution of the units being segmented hints at the necessity to further study the functional/thematic structures of elementary web documents in order to better understand their networking. This is due to the fact that linkage of a page may be due to its membership in a web community, its role in manifesting a broad topic, its function as a hub or authority or as a component of a website manifesting a certain webgenre. Generally speaking, these considerations hint at the need to further integrate linguistic models of document types in order to ground document network analysis not only in terms of (social science and) statistics, but also of linguistics. Future elaborations of this apparatus will need to follow this direction in order to better grasp the genre/register sensitivity of the characteristics of document networks. This includes also networks whose generation is restricted by the kind of web-based software as surveyed in the subsequent section.

3.6 Social Software-Based Networks

With the advent of the so-called *Web 2.0* (O'Reilly, 2005; Bächle, 2006), a further, hardly foreseen media change takes place. In some areas of the web, a kind of a 'content provider' who generates her offer of information in cooperation with other members of the same social network takes the place of the classical WWW user in the role of a passive information recipient. That is, some parts of the web evolve as a medium of distributed cognition (Hollan et al., 2000) by utilising so-called *social software* which covers, amongst others, fora, networked blogs and countless wikis of knowledge or technical communication. The central aim of social software is to support the web-based buildup and self-organisation of social networks in the form of virtual communities of members which — without the need of face-to-face communication — cooperatively/competitively perform some task (e.g. writing a technical documentation, programming open source software, building an electronic encyclopedia etc.) *without involving any kind of central supervision*. In this sense, one may speak of *social software-mediated communication*. In this section, we review approaches to document networks which were cooperatively produced by means of some social software. This includes internet mailing lists (IML), fora, networked weblogs and wiki-based document networks. Other areas of the web 2.0 relevant for document networking, not taken into consideration, include *social bookmarking* and *social networking* systems.

3.6.1 Web Fora

A (web) (discussion) forum or (electronic bulletin) board is a website which supports asynchronous discussions on certain topics within groups of posters who possibly registered to the forum (Bächle, 2006; Fisher, 2003). A forum and its sub-fora are usually bound to a certain topic and its sub-topics, respectively (Bächle, 2006). It is built around the postings of its posters, where postings on the same subject as part of the same discussion are organised into a thread (see below). In this section, we will review approaches to forum-based document networking by example of Usenet newsgroups.

Usenet newsgroups span a worldwide system of electronic bulletin boards where each newsgroup organises discussions of a certain topic (Bar-Ilan, 1997). Usenet is hierarchically organised in a way which is reflected by the newsgroup names (Meinel and Sack, 2004; Smith, 2003). A newsgroup name is prefixed by the name of the topmost group to which the newsgroup belongs. It is followed by a sequence of period-separated subgroup names indicating — with increasing thematic resolution — the topic of the newsgroup (Bar-Ilan, 1997). For the number and diversity of newsgroups see (Kot et al., 2003; Smith, 2003). Each newsgroup may organise several discussions possibly in parallel to each other. Each discussion is organised as a thread which finally consists of single postings, i.e. messages or news items. As a user can answer to a message posted before, a post order of hierarchically threaded postings emerges. That is, a thread is a hierarchically ordered series of news items usually about a single topic instantiated by the thread's root or initial message where succeeding messages uniquely refer to a previous one. In its header, a news item identifies its submitter and subject while its body contains the message content. As there is no subscription procedure for newsgroups, one can easily participate in a discussion via email, although newsgroups may be moderated. The moderator may decide, for example, to crosspost a message, that is, to post it in different newsgroups.

Usenet contains science-related newsgroups and, thus, supports scientific communication. The BIOSCI/bionet newsgroup, for example, "is a series of freely accessible electronic communication forums (i.e., electronic bulletin boards or "newsgroups") for use by biological scientists worldwide" (cf. `www.bio.net/docs/biosci.FAQ.html` — cf. also Kot et al. 2003). Its aim is "to promote communication between professionals in the biological sciences" (ibid.). But as Usenet postings are not refereed, they cannot be compared with related forms of collaborative, peer-reviewed publication in scientific communication (Bar-Ilan, 1997). Moreover, other than in Wiki-based systems, once statements are posted in a newsgroup, they are, usually, no longer editable, let alone collectively.

Reflecting the thematic focus of Usenet newsgroups, related models of document networking concentrate on the time-related principles of postings on single topics. Bar-Ilan (1997), for example, analyses a corpus of about 16,000 Usenet messages on the *mad cow disease* in a period of hundred days starting around the beginning of this "food scandal". She studies the growth function of topic-specific messages within certain periods of time in order to analyse time-dependent phenomena as topic spread and burst.This approach is related to the study of Sengupta and Kumari (1991) of the growth rate of AIDS related publications. Other than Bar-Ilan, they observe an "epidemic", exponential growth of such publications during the period of 1976 to 1986. In a related context, but with a focus on the WWW, Bar-Ilan and Echermane (2005) analyse web pages linking to contributions on the anthrax scare.

The diversity of newsgroups is studied by Kot et al. (2003) by example of a group of 107 bioscience

related newsgroups as can be downloaded from `ftp://ftp.bio.net/BIOSCI/ARCHIVE`. They show that the number of postings of contributors ranked by decreasing number obeys Zipf's law. Kot et al. develop a model in terms of a stochastic process which accurately predicts the contribution of posters w.r.t their number, size and the total number of postings within the simulated newsgroup.

A power law-like characteristic of newsgroup postings is explored by Agrawal et al. (2003) who fit a power law (with exponent $\gamma \approx 0.8$) to the ranked-size distribution of (the number of) postings per author. But other than the studies just reviewed, Agrawal et al. do not consider document, but agent networks in which nodes denote posters which are linked if one has quoted from an earlier posting written by the other.

3.6.2 Internet Mailing Lists

An Internet Mailing List (IML) as uniquely identified by its name and address collects the list of email addresses of its members. A member can post a message which is then sent to all other members unless censored by the list moderator. The reason to submit may be to initiate a discussion by posting a question, hypothesis, or an issue for debate (Kuperman, 2005). Inter-discussion links occur subject to referential links (e.g. references *to* or quotations *from* preceding discussions) or thematic relatedness and, thus, give rise to networking beyond single discussions. As topic-based links are not explicit, they need to be explored in order to contribute to a networked corpus. The members of a list can actively participate in a discussion or passively follow it in the role of a so-called *lurker*. As submitters answer to messages posted before, a hierarchically threaded structure of *email* postings emerges by analogy with web fora. That is, an IML thread is a hierarchically ordered series of messages discussing a single topic with a unique initial message. Alternatively, threads may be simply linearly ordered as in the LINGUIST LIST in which postings are linked as sequels of the same discussion. According to Zelman and Leydesdorff (2000), threaded email messages are the fundamental communication units of IML-based computer-mediated communication (CMC). In scientific communication, they are characterised by their size and thematic homogeneity as their submitters are known for their expertise (Thelwall and Wouters, 2005) clearly affecting the self-organisation of these IMLs although the extent of this impact has still to be proven (Zelman and Leydesdorff, 2000).

IMLs are, usually, moderated to a higher degree than newsgroups, but to a lower degree than conventional scholarly publications. Moderation is based on publication policies. It may concern the format, content and size of postings as well as preventing repeated discussions. Comparable to newsgroups, but other than in conventional scholarly publications, IML-based postings are less restricted w.r.t their number, size, frequency and related restrictions induced by the publishing medium. On the other hand, scholarly IMLs are characterised to have more qualified contributions than unmoderated newsgroups (Hernandez-Borges et al., 1998). Thus, moderated IMLs of scientific communication can be settled in-between less moderated newsgroups and scholarly publications which are restricted in terms of their access, number, size, and frequency of publication (Kuperman, 2005). For the time being, a standard format for archiving and retrieving IMLs is missed (Zelman and Leydesdorff, 2000) as well as in the case of web fora.

Kuperman (2005) reports on a bibliometric analysis of the productivity of two IMLs. He analyses a corpus of $5,016$ emails of the LINGUIST LIST (cf. `linguistlist.org/issues/master.html`) and a

corpus of 3,023 emails of the History of the English Language List (cf. `listserv.linguistlist.org/archives/hel-l.html`). Kuperman shows that members of the power law family, e.g. Lotka's law, Zipf's law, Zipf-Mandelbrot's law or the Yule, Yule-Simon or the Waring distribution (Simon, 1955; Rapoport, 1982; Wimmer and Altmann, 1999b), poorly fit the ranked-size distribution of postings over authors in *unmoderated* lists. Goodness of fit is better in lists with a higher level of moderation. This result is in support of locating IMLs in-between the area of unmoderated newsgroups and conventional scholarly publications. It was the latter area for which power laws have been successfully fitted in scientometrics. In order to generalise this observation, we may hypothesize that the less restricted the publication process, the less distinctive the incentive to publish, the less "Zipfian" the order of publications (e.g. postings). This finding is supported by Zelman and Leydesdorff (2000) who analyse eleven IMLs of *scientific* communication which include, for example, IMLs on *Self-Organisation* and *Science & Technology Studies*. This corpus includes mailing lists of scientific projects as well as intermediate and field level lists. Zelman and Leydesdorff (2000) aim at describing the dynamics of IMLs by means of statistical indices. Amongst others, this includes counting the number of messages per thread and, subsequently, fitting a function in double-logarithmic scale to the distribution of the frequencies of thread size which allows deriving a corresponding power law with an exponent $-0.42 \leq \gamma \leq -0.47$.

3.6.3 Networked Blogs in Blogspace

A *weblog* or *blog* for short is a web site which, in the majority of cases, is authored by a single author, i.e. a *blogger*, with the help of a *weblog system* (Glance et al., 2004). As the word *blog* may denote the action of blogging, its end product (i.e. a blog) or the software that enables blogging (Gill, 2005), we will solely refer to the product perspective when using this term. According to Kumar et al. (2003) and Gill (2004), blogs consist of time-aligned, date-stamped, possibly archived entries that are reversely chronologically ordered and additionally contain links to related entries of the same or other blogs in conjunction with so-called blogrolls (as lists of links to recommended blogs).

According to Kumar et al., blogs are "quirky, highly personal, often consumed by regular repeat visitors and highly interwoven into a network of small but active micro-communities." This network of interrelated blogs is called *BlogSphere*, *blogosphere* or *blogspace*, respectively. Generally speaking, blogs can be characterised w.r.t their structure, content and the functions they provide. Nardi et al. (2004) point out the thematic heterogeneity of blogs and stress the wide range of motivations of bloggers to blog which make network analysis a hard task in this area. Likewise, Schmidt et al. (2005) state that blogs serve divergent functions including that of a personal diary, journalistic publishing as well as of knowledge or organisational communication. Accordingly, (personal) *online diaries* or *journals*, *blogs of pundits* (i.e. self-declared knowledge experts), *news filter blogs* (based on RSS aggregators), *writer* or *artist blogs*, *marketing blogs*, and *spam blogs* are distinguished as some examples of weblog genres (Gruhl et al., 2004; Glance et al., 2004; Bächle, 2006). Krishnamurthy (2002) presents a two-dimensional model of classifying weblogs according to their *personal* vs. *thematic* and *community* vs. *individual* orientation. Schmidt (2006) locates weblogs in-between a spectrum of media schemata spanned by standard websites on the one hand and media of asynchronous text-based CMC on the other hand. Other than IMLs as examples of the latter, weblogs are asymmetric (as they are usually authored by a single blogger whereby readers only have, if at all, the possibility to comment on en-

tries). Other than websites, weblogs are (intended to be) continuously updated, but are in a restricted sense multimedial.

Evidently, network studies focusing on a single one of the dimensions just mentioned face the risk to overgeneralise to the disadvantage of the disregarded dimensions and their impact on networking within the blogspace (Schmidt et al., 2005). Thus, the sampling of blogs by example of which the network structure of the blogspace is investigated has to be carefully considered. From a structural point of view, the following types of links between blogs can be explored for this task (Glance et al., 2004): (friendship indicating) links as part of the *blogroll* of a blog, *trackbacks* (linking blogs whose bloggers have linked the focal entry), *permalinks* (as URIs which uniquely identify posts irrespective of whether they have been archived or not) as well as hyperlinks within a blog entry (to other blogs or web pages outside the blogspace). Blogs manifest intra links (interrelating entries of the same blog) as well as extra links which settle them, for example, in the neighborhood of related blogs (participating in the same discussion) (Gruhl et al., 2004). As all these kinds of links are not (necessarily) mutual, graphs derived thereof are necessarily directed. Starting from these structural notions, several reference points of complex network analysis come into play:

- Firstly, so-called *small communities* in the sense of Kumar et al. (2003), that is sub-networks of blogs which link to each others postings while discussing some topic within a certain period of time.
- Secondly, a large component of interlinked blogs or, alternatively, a *blog site* collecting hundreds and thousands of (links to) blogs (cf. Kumar et al., 2004) may be made an object of network analysis. See Adar et al. (2004) for an enumeration of such sites.
- Thirdly, the system of blog sites as interlinked by means of their component blogs may be made an object of network analysis.
- This leads, fourthly, to the whole blogspace as a candidate input of complex network analysis.

A system for building blog corpora is described by Glance et al. (2004). It includes an URL harvester, a blog crawler, a time aligner (for mapping blog entries to timestamps) and an indexer for making the collected blogs retrievable. The corpus builder also comprises text mining software for exploring thematic trends and, thus, time-dependent structure formation. As a sample corpus, Glance et al. crawl about 100,000 weblogs.

Herring et al. (2005) show that blog networks have SW-related characteristics as, for example, preferential attachment. They investigate the formation of so-called *blog dyads* constituted by their manyfold mutual links and "textual interaction" by means of reciprocal verbal exchange manifesting a sort of "conversation" between the corresponding bloggers. Herring et al. point out that — in contrast to what is propagated by the blogger community — blog linkage is an infrequent phenomenon making such dyads a rather seldom event: "the blogosphere appears to be selectively interconnected, with dense clusters in parts, and blogs minimally connected in local neighborhoods, or free-floating individually, constituting the majority." Likewise, Herring et al. (2004) present frequencies of various types of links which in spite of their wide range indicate that linking is a rare phenomenon in blogspace. This is more or less in accordance with successfully fitting power laws to the in-degree distribution of blogs — for related studies of power law fitting cf. Glance et al. (2004). See also Tricas

et al. (2004) who fit a power law with an exponent $\gamma \approx -0.58$, although they report on problems with fitting. The relevance of such analyses is confirmed by studies which show that users tend to focus on highly linked blogs. This hints at preferential attachment as a minority of authoritative blogs (Herring et al., 2005) is preferably linked from other blogs as well as from outside the blogosphere.

By analogy with web fora and internet mailing lists, networked blogs have also been made an object of investigating time-dependent structure formation. The aim of this research is to investigate the life cycle of thematic spreads and bursts within the blogosphere. This is made possible by the timestamps of blog entries. In this context, the study of Kumar et al. (2003, 2004) is of special interest. Kumar et al. (2003) introduce the notion of a *time graph* in order to describe link generation in the blogspace as a function of time. Time graphs are used to explore the build-up of blog communities and to separate recurrent periods of time within their life cycle. The formation of *small* communities (of about three to twenty members — cf. Kumar et al. 2004) is described as a characteristic of the blogspace. That is, blogs are seen to be networked — other than newsgroups — on the basis of small communities (of blogs whose authors mutually link each other within their blogrolls and respond to newly posted content within the corresponding community). Moreover, other than "classic" web communities (Gibson et al., 1998), blog communities show a strikingly temporal characteristic as they evolve subject to temporarily raising debates during which linkages of the blogs involved into community building rapidly grow before they decrease with the debate fading away. Kumar et al. (2004) distinguish three periods of time in the life cycle of blog communities as they, firstly, undergo a sudden burst of activity of rapid-fire discussion in a small period of time before they, secondly, lie, so to speak, dormant for weeks and are, thirdly, replaced by a subsequent burst. A characteristic trait of their study is that other than many other approaches they analyse a large corpus of about one million interwoven blogs. Extending the analysis of Kumar et al. (2003), they include the spatial and topical dimension into network analysis. Such a spatial restriction, which has already been taken into account in scientometrics, is also considered by Lin and Halavais (2004).

Adar et al. (2004) develop the notion of an information epidemics spreading over the blogosphere. They analyse a corpus of about 40,000 blogs with about 175,000 links in order to classify sitations (see Section 3.4) within blogs dependent on their time characteristics. Likewise, Gruhl et al. (2004) describe the long-term propagation of topics which are referred to in order to segment the blogspace on a macroscopic level. They distinguish spikes and chatters, that is, ongoing and short-term, but highly intensive discussions, respectively.

3.6.4 Wiki-based Document Networks

A fourth example of social software which became prominent by the online encyclopedia *Wikipedia* (`wikipedia.org`) is wiki software. By analogy with weblogs, one has to distinguish *wiki software* (e.g. MediaWiki, cf. `www.mediawiki.org/wiki/MediaWiki`, or TWiki, cf. `www.twiki.org`) from the *document networks* (as exemplified in Table 6) generated with this software. For a comparative overview of wiki software see `www.wikimatrix.org`. In the present review we refer to the product perspective when using the term *wiki* and, thus, refer by this term to document networks generated by means of some wiki software. Generally speaking, a wiki is a website which by means of the corresponding software allows collaborative writing, editing and revising the collection of pages and links this site con-

Wiki	URL	Language
a city wiki	ka.stadtwiki.net/Hauptseite	de
a wiki about the Wikimedia Foundation's projects	meta.wikimedia.org/wiki/Main_Page	en
a wiki-based dictionary of French	fr.wiktionary.org/wiki/	fr
Ward Cunningham's wiki (alias WikiWikiWeb)	c2.com/cgi/wiki	en
wiki of the Firefox project	www.firefox-browser.de/wiki/Hauptseite	de
wiki of the MediaWiki software	www.mediawiki.org/wiki/MediaWiki	en
wiki of the Mozilla project	wiki.mozilla.org/Main_Page	en
wiki of the OpenOffice.org	wiki.services.openoffice.org/wiki/Main_Page	en
wiki of the swarm project	www.swarm.org/wiki/Main_Page	en
wiki of the Wikibooks project of free textbooks	en.wikibooks.org/wiki/Wikibooks_portal	en
wiki of the wikis of the Apache.org projects	wiki.apache.org/general/	en
wiki of the W3C RIF Working Group (restricted access)	www.w3.org/2005/rules/wg/wiki/	en

Table 6: Some wikis of knowledge and technical communication.

sists of. By analogy with social software and its output, the generation of wikis is a self-organised process initiated and continued by a multitude of cooperating/competing authors who may, but in general do not have exclusive access to editing the wiki (Kuhlen, 2004). In other words, wikis manifest a sort of distributed, non-linear production and revision of hypertext documents and, thus, a sort of hypertextually manifested distributed cognition including social tagging (Mika, 2005) as exemplified by the category system of Wikipedia (tools.wikimedia.de/~daniel/WikiSense/CategoryTree.php). Other than "classic" websites, wikis are continually and cooperatively updated. Other than weblogs and maling lists, wiki software-mediated communication is, in principle, symmetric in the sense that every (registered or permitted) user can respond to, continue or edit the contribution of any other wiki author. Thus, the need of a thread-based organisation does not apply, although changes are archived by means of history pages accompanying each article page (see below). For a general discussion of wiki software-mediated communication, the underlying wiki software, some structural, statistical characteristics of wikis and their impact on knowledge communication see Ebersbach et al. (2005), Voss (2005), Holloway et al. (2005) and Kuhlen (2004), respectively.

This section concentrates on wikis built by means of the MediaWiki software. It is used by the Wikimedia Foundation (wikimediafoundation.org/wiki/Home) which hosts the Wikipedia project and its language specific releases which, together, are the largest wikis on the web. The dumps of these releases and those of many other wikis of the Wikimedia Foundation are accessible via download.wikimedia.org what — in spite of the size of these files — makes wiki network analysis a manageable task. An alternative way of downloading wikis is to explore wiki pages listing, if existing, all entry pages of the corresponding wiki as, for example, www.firefox-browser.de/wiki/Spezial:Allpages by example of the Firefox browser wiki. This gives access to all pages of the focal wiki website which need to be further analysed in order to explore their links.

Network extraction by example of Wiki-based networks faces the situation of the rich type system of node and link types as exemplified by the Wikipedia. That is, network extraction cannot be performed by simply extracting all wiki *article* pages as this may disregard other types of nodes and links (see Table 7). Thus, the question, which node and link types shall be taken into account, has to be carefully considered. A starting point for distinguishing most elementary node types in MediaWiki-based networks is what will be called a wiki document which consist of an article page (describing a

Type	Frequency
Pages total	796,454
Article	303,999
Redirect Node	190,193
Talk	115,314
Article Talk	78,224
User Talk	30,924
Image Talk	2,379
Wikipedia Talk	1,380
Category Talk	1,272
Template Talk	705
Portal Talk	339
Mediawiki Talk	64
Help Talk	27
Image	97,402
User	32,150
Disambiguation	22,768
Category	21,999
Template	6,794
Wikipedia	3,435
Mediawiki	1,575
Portal	791
Help	34

Type	Frequency
Links total	17,814,539
Interlink	12,818,378
Category Link	1,415,295
Categorises	704,092
Categorised by	704,092
Category Associates with	7,111
Topic of Talk	103,253
Talk of Topic	88,095
Hyponym of	26,704
Hyperonym of	26,704
Inter Portal Association	1,796
Broken	2,361,902
Outside	1,276,818
Inter Wiki	789,065
External	487,753
Intra	1,175,290
Kernel	1,153,928
Across	6,331
Up	6,121
Reflexive	5,433
Down	3,477
Redirect	182,151

Table 7: The system of node and link types and their frequencies by example of the German release of the Wikipedia (download in November 14, 2005).

certain entry of the wiki), a corresponding *discussion* (or *talk*), *history* and *edit this* or *view source* page which altogether form a flatly structured document.

Due to namespace conventions, some additional types of nodes can be distinguished by example of the Wikipedia. Table 7 lists all node types as found within its German release or additionally introduced in order to span a hierarchical type system. The central heuristic for extracting instances of node types relates to the URL of the corresponding document module and its namespace prefix, respectively. Category, portal and media wiki pages, for example, contain the namespace prefix `Kategorie` (*category*), `Portal` (*portal*) and `MediaWiki`, respectively. It is separated by a colon from the corresponding page name suffix. `de.wikipedia.org/wiki/Kategorie:Musik`, for example, references the page of the *music* category, whereas `de.wikipedia.org/wiki/Portal:Musik` identifies the German Wikipedia portal on music. Finally, `de.wikipedia.org/wiki/Musik` references the standard German Wikipedia article on music. That is, URLs of standard wiki articles do not include a special namespace prefix.

Table 7 lists the frequencies of the instances of the node types as found in the input release of the Wikipedia. The types are ordered into an inclusion hierarchy in which the child nodes dominated by the same type (e.g. `talk`) are ranked according to their frequencies in descending order. Further, the frequency of a dominating node is the sum of the frequencies of its child nodes. Analogously, table 7 lists all edge types as found within the input wiki or additionally introduced into the study in order to span a hierarchical type system.

From the point of view of network extraction, *redirect nodes* and *links* which manifest transitive and, thus, mediate links of content-based units are of special interest. An article node v may be linked, for example, with a redirect node r which in turn redirects to an article w. In this case, the document

network contains two edges $(v, r), (r, w)$ which have to be resolved to a single edge (v, w) if redirects are to be excluded in accordance with what the MediaWiki system does when processing them. That is, a user, when clicking on the corresponding lexical or phrasal anchor of the link (v, r) does not enter r, but is redirected to the node w. Such redirects may include more than one redirect node r.

A further peculiarity of the Wikipedia are portals which introduce a further level of structuring above the level of wiki documents and below the level of the wiki website as a whole. Entry pages of portals are identified by means of the corresponding namespace prefix. Other ways of typing nodes in wiki-based networks, which are not (necessarily) reflected by a namespace prefix, operate on user or entry page statistics. This is exemplified by stubs, that is, Wikipedia entries which are too short to be a useful encyclopedia article.

Complex network and other statistical analyses of wiki-based networks are performed, amongst others, by Voss (2005), Capocci et al. (2006), Zlatic et al. (2006) and — in comparison to other text and document networks — by Mehler (2006).

Voss (2005) suggests that Lotka's law also characterises authorship in Wikipedia w.r.t the distribution of the number of edits ranked by the number of authors which are responsible for the respective number of edits. Voss reports a low exponent $\gamma \approx 0.5$. Further, he considers the distribution of the number of authors ranked by the number of distinct articles they are authors of and also finds a highly skewed distribution (i.e. very many authors have contributed to only one article whereas only a small minority of persons authored very many articles while there is a smooth transition between these two extreme cases.

Capocci et al. (2006) analyse the topology of the English and of the Portuguese release of the Wikipedia in terms of the bow-tie model of Broder et al. (2000) (cf. Section 3.5). They observe that most of the entry pages of the Wikipedia belong to its SCC, that is, almost any of its pages can be reached from any other of these pages. Capocci et al. fit a power law to the in-degree and to the out-degree distribution of entry pages (which types of pages were actually considered is not mentioned in the paper). In both cases, fitting is successful with an exponent $2 \leq \gamma^{\text{in,out}} \leq 2.2$. Further, Capocci et al. observe a lack of correlation regarding the in-degree of vertices and the average in-degree of its neighboring vertices — this observation is in accordance with computing assortative mixing in the wiki medium (see below). Finally, Capocci et al. consider a model of network growth based on *directed* graphs. They simulate growth in terms of preferential attachment where the probability of acquiring a new edge is separately computed for incoming and outgoing edges subject to the present in-degree and out-degree of vertices, respectively. This model does not only distinguish the direction of newly added edges, but also wether they link already existing vertices or end at newly added ones. A central conclusion of Capocci et al. is that the Wikipedia resembles the WWW in terms of the characteristics they measured.

Zlatic et al. (2006) is the most comprehensive Wikipedia-based network study. They analyse the releases of ten languages (including English, German, Japanese, French and Spanish) by taking different node types into account — cf. Table 7. But although Zlatic et al. distinguish *article*, *talk*, *help*, *user*, *category*, *redirect* and *template* pages as well as *images* and *multimedia* resources, they report only on calculations with articles, redirects and templates thereby distinguishing broken and non-broken links. A central aim of their study is to distinguish network characteristics common to all releases from those which select singletons of them. Zlatic et al. fit power laws to the in and out-degree distributions of

directed graphs extracted from the Wikipedia releases as well as to the degree distributions of their undirected counterparts (see Table 4). Further, they fit a power law in order to predict the growth of links between entry pages as a function of the number of those pages and observe approximately a linear increase of the number of links and pages. Zlatic et al. compute the rate of assortative mixing within the different releases, compute their cluster values and estimate the average geodesic distance of their entry pages (see Table 4). Interestingly, they observe that clustering in Wikipedia generally decreases with the growth of the network. Generally speaking, the growth dependent variation of network characteristics will be a promising future direction of complex network research where the Wikipedia provides a tremendous set of information in support of this kind of research. Zlatic et al. (2006) also explore motifs of wiki networks by analogy with those found to be characteristic of the WWW. In spite of this and some other characteristics in support of assuming the resemblance of the WWW and the Wikipedia, Zlatic et al. find characteristic differences of these networks, for example, in terms of their reciprocity, that is the non-random existence of reciprocal edges between pairs of vertices (Garlaschelli and Loffredo, 2004). Further, they observe — other than in the WWW — a higher stability of the average geodesic distance and ask whether these and related differences are due to the specific growth dynamics of wiki-based networks or due to the structure of the underlying knowledge system approximated by them — these are just two more questions which are still open for future research.

Another perspective of exploring wiki-based networks is opened by Mehler (2006) who comparatively studies document networks in knowledge, technical, press and WWW-based communication. He analyses three variants of the German release of the Wikipedia dependent on the different types of nodes and links taken into account. This ranges from a variant based on article pages and their links only to the whole spectrum of entry pages and their different types of links as distinguished in Table 7 (except for broken and external links leading to pages outside the input wiki). Further, Mehler (2006) analyses three wiki networks of the `Apache.org` which belong to the topic of technical communication and, thus, considers wikis of different areas of WWW-based communication. He computes the cluster coefficients and average geodesic distances of input networks, fits power laws to the degree distributions of undirected graphs extracted from them and calculates the rate of assortative mixing within these networks (see Table 4). As in the case of the studies summarised above, Mehler (2006) observes a latent tendency towards disassortative mixing in wiki networks, while they can be definitely attributed as small worlds according to the WS model. Note that Mehler (2006) reports on very small values of the exponent γ of the BA model (i.e. $\gamma \approx 0.5$) in the case of the Wikipedia and even smaller values in the case of the wikis of technical communication. These differences are clarified by hinting at the choice of $k_{\min}$ and $k_{\max}$ as the degrees *from which* and *up to which* power laws are fitted. In Mehler (2006), the whole degree distributions are fitted while Zlatic et al. use a higher value of $k_{\min}$ and a lower value of $k_{\max}$ — Capocci et al. (2006) do not report on the choice of $k_{\min}$ and $k_{\max}$.

A central outcome of Mehler's study is that the different areas of document networking show strikingly differences w.r.t their network characteristics (except from their average geodesic distances). This result supports the view that the small-world property and related characteristics of complex document networks vary significantly with the underlying genre or area of communication and that these characteristics denote non-categorical, graded network properties.

4 Conclusion and Future Perspectives

This article surveyed approaches to text and document networks which, by the majority, refer to small-world models. The article reviewed studies in support of the view that these networks share topological characteristics so that one may speak of principles of the collaborative formation of intertextual structures in corpora of natural language texts. In spite of these commonalities, the article also hinted at the genre-sensitivity of these principles. In this sense, the small-world property, for example, cannot be attributed categorically, but varies with the underlying text and document genre which, thus, is not only mirrored on the level of *text internal characteristics* but also by *principles of intertextual networking*.

These and related findings raise the question for the corpus linguistic importance of complex network analysis. From a corpus linguistic point of view, the small-world property of text networks can be seen as an argument in favor of representative samples as input to computing, for example, cognitively plausible models of lexical association. Although it is known from quantitative linguistics that such samples are hardly possible — cf., for example, Orlov (1982) — the small-world property can at least be utilised as a necessary condition which has to be fulfilled by a corpus in order to be judged as a reliable data base for computing lexical memory models showing the small-world property on their own. Moreover, knowing that a given corpus has the small-world property one can infer a certain rate of change of a given variable (e.g. topic) when following intertextual relations. That is, the notion of *intertextual neighborhood* relevant to study a given text, as claimed by Stubbs (2001), can be approached in this framework.

From this perspective, at least three challenging research questions can be identified as an object of future research in this field:

- What do more realistic, linguistically grounded network models look like which do not only account for the genre-sensitivity of intertextuality in more detail, but can be used to simulate the generation of such networks in order to investigate network states which, for the time being, are empirically unobservable (e.g. because of their complexity or the impossibility of parameter variation in the case of real networks)?
- What are the interrelationships of quantitative principles on the level of texts (e.g. vocabulary growth) and those restricting their networking? Is there a unifying, so to speak Zipfian theory which grasps principles of intra- *and* intertextual structure formation?
- What do text-technological representation formats and their operations look like which are expressive enough to manage document networks of the complexity mentioned above and to compute related network characteristics?

These and related questions open up many research opportunities in utilising and enhancing the apparatus of complex network analysis in corpus linguistics.

Acknowledgement

The author thanks Vinko Zlatic (Rudjer Boskovic Institute, Theoretical Physics Division), Ramon Ferrer i Cancho (Universitat de Barcelona), the editors of this handbook as well as the anonymous reviewers of this article for their careful proof-readings and fruitful hints.

References

Adamic, L. A. (1999). The small world of web. In Abiteboul, S. and Vercoustre, A.-M., editors, *Research and Advanced Technology for Digital Libraries*, pages 443–452. Springer, Berlin.

Adamic, L. A. (2000). Zipf, power-law, Pareto — a ranking tutorial. `http://www.hpl.hp.com/research/idl/papers/ranking/`.

Adamic, L. A. and Huberman, B. A. (2001). The Web's hidden order. *Communications of the ACM*, 44(9):55–59.

Adar, E., Zhang, L., Adamic, L. A., and Lukose, R. M. (2004). Implicit structure and the dynamics of blogspace. In *Proc. of the Workshop on the Weblogging Ecosystem at the 13th Int. Conf. on World Wide Web (WWW'04)*, New York.

Agrawal, R., Rajagopalan, S., Srikant, R., and Xu, Y. (2003). Mining newsgroups using networks arising from social behavior. In *Proc. of the 12th Int. Conf. on World Wide Web (WWW'03)*, pages 529–535, New York. ACM Press.

Ajiferuke, I. and Wolfram, D. (2004). Modelling the characteristics of web page outlinks. *Scientometrics*, 59(1):43–62.

Albert, R. and Barabási, A.-L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74:47.

Albert, R., Jeong, H., and Barabási, A.-L. (1999). Diameter of the World Wide Web. *Nature*, 401:130–131.

Altmann, G. (1988). *Wiederholungen in Texten*. Brockmeyer, Bochum.

Baayen, H. (2001). *Word Frequency Distributions*. Kluwer, Dordrecht.

Bächle, M. (2006). Social software. *Informatik Spektrum*, 29(2):121–124.

Baeza-Yates, R. and Ribeiro-Neto, B., editors (1999). *Modern Information Retrieval*. Addison-Wesley, Reading, Massachusetts.

Baldi, P., Frasconi, P., and Smyth, P. (2003). *Modeling the Internet and the Web*. Wiley, Chichester.

Bar-Ilan, J. (1997). The "mad cow disease", Usenet newsgroups and bibliometric laws. *Scientometrics*, 39(1):29–55.

Bar-Ilan, J. (2001). Data collection methods on the web for infometric purposes – a review and analysis. *Scientometrics*, 50(1):7–32.

Bar-Ilan, J. and Echermane, A. (2005). The anthrax scare and the web: A content analysis of web pages linking to resources on anthrax. *Scientometrics*, 63(3):443–462.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Barabási, A.-L., Albert, R., and Jeong, H. (1999). Scale-free characteristics of random networks: The topology of the World Wide Web. *Physica A*, 281:69–77.

Barabási, A.-L., Albert, R., Jeong, H., and Bianconi, G. (2000). Power-law distribution of the World Wide Web. response to Adamic & Huberman (2000). *Science*, 287(12):2115a.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews. Genetics*, 5(2):101–113.

Baroni, M. and Bernardini, S. (2004). BootCaT: Bootstrapping corpora and terms from the web. In *Proceedings of LREC 2004*, pages 1313–1316, Lisbon. ELDA.

Bense, M. (1998). *Ausgewählte Schriften. Band 3. Ästhetik und Texttheorie*. Metzler, Stuttgart.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge University Press, Cambridge.

Björneborn, L. (2004). *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*. PhD thesis, Royal School of Library and Information Science, Department of Information Studies, Denmark.

Björneborn, L. and Ingwersen, P. (2001). Perspectives of webometrics. *Scientometrics*, 50(1):65–82.

Björneborn, L. and Ingwersen, P. (2004). Towards a basic framework for webometrics. *Journal of the American Society for Information Science and Technology*, 55(14):1216–1227.

Bock, H. H. (1994). Classification and clustering: Problems for the future. In Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., and Burtschy, B., editors, *New Approaches in Classification and Data Analysis*, pages 3–24. Springer, Berlin.

Bollobás, B. (1985). *Random Graphs*. Academic Press, London.

Bollobás, B. and Riordan, O. M. (2003). Mathematical results on scale-free random graphs. In Bornholdt, S. and Schuster, H. G., editors, *Handbook of Graphs and Networks. From the Genome to the Internet*, pages 1–34. Wiley-VCH, Weinheim.

Bordag, S., Heyer, G., and Quasthoff, U. (2003). Small worlds of concepts and other principles of semantic search. In Unger, H. and Böhme, T., editors, *Innovative Internet Computing Systems Second International Workshop (IICS '03)*, Berlin. Springer.

Bornholdt, S. and Schuster, H. G. (2003). *Handbook of Graphs and Networks. From the Genome to the Internet*. Wiley-VCH, Weinheim.

Botafogo, R. A., Rivlin, E., and Shneiderman, B. (1992). Structural analysis of hypertexts: Identifying hierarchies and useful metrics. *ACM Transactions on Information Systems*, 10(2):142–180.

Brainerd, B. (1977). Graphs, topology and text. *Poetics*, 1(14):1–14.

Brinker, K. (1991). Aspekte der Textlinguistik. Zur Einführung. In Brinker, K., editor, *Aspekte der Textlinguistik*, pages 7–17. Georg Olms, Hildesheim.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., and Wiener, J. (2000). Graph structure in the web. *Computer Networks*, 33:309–320.

Bronstein, I. N., Semendjajew, K. A., Musiol, G., and Mühlig, H. (1999). *Taschenbuch der Mathematik*. Harri Deutsch, Frankfurt a. M.

Brown, C. (2004). The Matthew effect of the *Annual Reviews* series and the flow of scientific communication through the World Wide Web. *Scientometrics*, 60(1):25–30.

Capocci, A., Servedio, V. D. P., Colaiori, F., Buriol, L. S., Donato, D., Leonardi, S., and Caldarelli, G. (2006). Preferential attachment in the growth of social networks: the case of Wikipedia. *Physical Review E*, 74:036116.

Chakrabarti, S. (2002). *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, San Francisco. Das Buch behandelt das Teilgebiet bzw. das Anwendungsgebiet des Web Mining.

Chakrabarti, S., Joshi, M., Punera, K., and Pennock, D. M. (2002). The structure of broad topics on the web. In *Proc. of the 11th Internat. World Wide Web Conference*, pages 251–262. ACM Press.

Chen, C. (1999). Visualising semantic spaces and author co-citation networks in digital libraries. *Information Processing and Management*, 35:401–420.

Chen, C. and Czerwinski, M. (1998). From latent semantics to spatial hypertext: An integrated approach. In Grønbæk, K., Mylonas, E., and Shipman, F. M., editors, *Proceedings of 9th ACM Conference on Hypertext and Hypermedia*, pages 77–86, New York. ACM.

Craven, M., DiPasquo, D., Freitag, D., McCallum, A. K., Mitchell, T. M., Nigam, K., and Slattery, S. (2000). Learning to construct knowledge bases from the World Wide Web. *Artificial Intelligence*, 118(1-2):69–113.

de Beaugrande, R. A. (1980). *Text, Discourse, and Process. Toward a Multidisciplinary Science of Texts*, volume 4 of *Advances in Discourse Processes*. Ablex, Norwood.

de Beaugrande, R. A. (1997). *New Foundations for a Science of Text and Discourse: Cognition, Communication, and the Freedom of Access to Knowledge and Society*. Ablex, Norwood.

de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, 149(3683):510–515.

Diestel, R. (2005). *Graph Theory*. Springer, Heidelberg.

Dillon, A. and Gushrowski, B. A. (2000). Genres and the WEB: Is the personal home page the first uniquely digital genre? *Journal of the American Society of Information Science*, 51(2):202–205.

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.

Ebersbach, A., Glaser, M., and Heigl, R. (2005). *WikiTools*. Springer, Berlin.

Egghe, L. and Rousseau, R. (2003). A measure for the cohesion of weighted networks. *Journal of the American Society for Information Science and Technology*, 54(3):193–202.

Eiron, N. and McCurley, K. S. (2003). Untangling compound documents on the web. In *Proceedings of the 14th ACM conference on Hypertext and Hypermedia, Nottingham, UK*, pages 85–94.

Faba-Pérez, C., Guerrero-Bote, V. P., and Moya-Anegón, F. D. (2003). "Sitation" distributions and Bradford's law in a closed web space. *Journal of Documentation*, 59(5):558–580.

Fairclough, N. (1992). *Discourse and Social Change*. Polity Press, Cambridge.

Fang, Y. and Rousseau, R. (2001). Lattices in citation networks: An investigation into the structure of citation graphs. *Scientometrics*, 50(2):273–287.

Ferrer i Cancho, R., Riordan, O., and Bollobás, B. (2005). The consequences of Zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society*, 272:561–565.

Ferrer i Cancho, R. and Solé, R. V. (2001). The small-world of human language. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 268(1482):2261–2265.

Ferrer i Cancho, R., Solé, R. V., and Köhler, R. (2004). Patterns in syntactic dependency-networks. *Physical Review*, E(69):051915.

Firth, D. and Lawrence, C. (2003). Genre analysis in information systems research. *Journal of Information Technology Theory and Application*, 5(3):63–87.

Fisher, D. (2003). Studying social information spaces. In Lueg and Fisher (2003), pages 3–19.

Fix, U. (2000). Aspekte der intertextualität. In Brinker, K., Antos, G., Heinemann, W., and Sager, S. F., editors, *Text- und Gesprächslinguistik / Linguistics of Text and Conversation – Ein internationales Handbuch zeitgenössischer Forschung*, volume 1, pages 449–457. De Gruyter, Berlin/New York.

Flake, G., Lawrence, S., and Giles, C. L. (2000). Efficient identification of web communities. In *Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–160, Boston, MA.

Garfield, E. (1963). Citation indexes in sociological and historical research. *American Documentation*, 14(4):289–291.

Garfield, E. (1994). Scientography: Mapping the tracks of science. *Current Contents: Social & Behavioural Sciences*, 7(45):5–10.

Garlaschelli, D. and Loffredo, M. I. (2004). Patterns of link reciprocity in directed networks. *Physical Review Letters*, 93:268701.

Gibson, D., Kleinberg, J., and Raghavan, P. (1998). Inferring web communities from link topology. In *Proceedings of the 9th ACM conference on Hypertext and Hypermedia: links, objects, time and space – structure in Hypermedia systems, Pittsburgh, Pennsylvania*, pages 225–234. ACM Press.

Giles, C. L., Bollacker, K., and Lawrence, S. (1998). CiteSeer: An automatic citation indexing system. In Witten, I., Akscyn, R., and Shipman III, F. M., editors, *Digital Libraries 98 – The Third ACM Conference on Digital Libraries*, pages 89–98, Pittsburgh, PA. ACM Press.

Gill, K. E. (2004). How can we measure the influence of the blogosphere? In *Proceedings of the Workshop on the Weblogging Ecosystem at the 13th international conference on World Wide Web (WWW'04)*, New York.

Gill, K. E. (2005). Blogging, RSS and the information landscape: A look at online news. In *Proceedings of the 2nd Annual Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics 14th international conference on World Wide Web (WWW'05)*, Chiba, Japan.

Glance, N., Hurst, M., and Tomokiyo, T. (2004). BlogPulse: Automated trend discovery for weblogs. In *WWW 2004 Workshop on the Weblogging Ecosystem: Aggregation, Analysis and Dynamics*.

Glänzel, W. and Czerwon, H.-J. (1996). A new methodological approach to bibliographic coupling and its application to the national, regional and institutional level. *Scientometrics*, 37(2):195–221.

Glenisson, P., Glänzel, W., and Persson, O. (2005). Combining full-text analysis and bibliometric indicators. A pilot study. *Scientometrics*, 63(1):163–180.

Gruhl, D., Guha, R., Liben-Nowell, D., and Tomkins, A. (2004). Information diffusion through blogspace. In *Proceedings of the 13th international conference on World Wide Web (WWW'04)*, pages 491–501, New York. ACM Press.

Halliday, M. A. K. (1966). Lexis as a linguistic level. In Bazell, C. E., Catford, J., Halliday, M. A. K., and Robins, R., editors, *In Memory of J. R. Firth*, pages 148–162. Longman, London.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*. Longman, London.

Harnad, S. and Carr, L. (2000). Integrating, navigating and analyzing eprint archives through open citation linking (the opcit project). *Current Science*, 79(5):629–638.

Heinemann, W. (1997). Zur Eingrenzung des Intertextualitätsbegriffs aus textlinguistischer Sicht. In Klein, J. and Fix, U., editors, *Textbeziehungen: linguistische und literaturwissenschaftliche Beiträge zur Intertextualität*, pages 21–37. Stauffenburg, Tübingen.

Hernandez-Borges, A. A., Macias, P., and Torres, A. (1998). Are medical mailing lists reliable sources of professional advice? *Medical informatics*, 23(3):231–236.

Herring, S. C., Kouper, I., Paolillo, J. C., Scheidt, L. A., Tyworth, M., Welsch, P., Wright, E., and Yu, N. (2005). Conversations in the blogosphere: An analysis "from the bottom up". In *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05)*.

Herring, S. C., Scheidt, L. A., Bonus, S., and Wright, E. (2004). Bridging the gap: a genre analysis of weblogs. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04)*.

Heyer, G., Quasthoff, U., and Wittig, T. (2006). *Text Mining: Wissensrohstoff Text*. W3L, Herdecke.

Hockey, S. (2000). *Electronic Texts in the Humanities*. Oxford University Press, Oxford.

Hoey, M. (1991). *Patterns of Lexis in Text*. Oxford University Press, Oxford.

Hoey, M. (1995). The lexical nature of intertextuality: A preliminary study. In Wårvik, B., Tanskanen, S.-K., and Hiltunen, R., editors, *Organization in Discourse. Proceedings from the Turku Conference*, pages 73–94.

Holanda, A. d. J., Torres Pisa, I., Kinouchi, O., Souto Martinez, A., and Seron Ruiz, E. E. (2003). Basic word statistics for information retrieval: thesaurus as a complex network. In *Anais XVI Brazilian Symposium on Computer Graphics and Image Processing*.

Hollan, J., Hutchins, E., and Kirsh, D. (2000). Distributed cognition: toward a new foundation for human-computer interaction research. *ACM Transaction on Computer-Human Interaction*, 7(2):174–196.

Holloway, T., Božičević, M., and Börner, K. (2005). Analyzing and visualizing the semantic coverage of Wikipedia and its authors. `http://tw.arxiv.org/abs/cs.IR/0512085`.

Holthuis, S. (1993). *Intertextualität. Aspekte einer rezeptionsorientierten Konzeption*. Stauffenburg, Tübingen.

Hummon, N. P. and Doreian, P. (1989). Connectivity in a citation network: The development of DNA theory. *Social Networks*, 11:39–63.

Itzkovitz, S., Milo, R., Kashtan, N., Ziv, G., and Alon, U. (2003). Subgraphs in random networks. *Physical Review E*, 68:026127.

Jakobs, E.-M. (1999). *Textvernetzung in den Wissenschaften*. Niemeyer, Tübingen.

Kautz, H., Selman, B., and Shah, M. (1997). Referral web: combining social networks and collaborative filtering. *Communications of the ACM*, 40(3):63–65.

Keller, F. and Lapata, M. (2003). Using the web to obtain frequencies for unseen bigrams. *Computational Linguistics*, 29(3):459–484.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, 14:10–25.

Kilgarriff, A. and Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–347.

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*, 315:665–676.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2):163–182.

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46(5):604–632.

Kleinberg, J. M., Kumar, R., Raghavan, P., Rajagopalan, S., and Tomkins, A. S. (1999). The web as a graph: Measurements, models, and methods. In Asano, T., Imai, H., Lee, D. T., Nakano, S., and Tokuyama, T., editors, *Computing and Combinatorics: 5th Annual International Conference (COCOON'99), Tokyo, Japan, July 1999*, Berlin/New York. Springer.

Kosala, R. and Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining*, 2(1):1–15.

Kot, M., Silverman, E., and Berg, C. A. (2003). Zipf's law and the diversity of biology newsgroups. *Scientometrics*, 56(2):247–257.

Krishnamurthy, S. (2002). The multidimensionality of blog conversations: The virtual enactment of september 11. In *Internet Research 3.0, Maastricht*.

Kuhlen, R. (1991). *Hypertext: ein nichtlineares Medium zwischen Buch und Wissensbank*. Springer, Berlin.

Kuhlen, R. (2004). Kollaboratives Schreiben. In Bieber, C. and Leggewie, C., editors, *Interaktivität – ein transdisziplinärer Schlüsselbegriff*, pages 216–239. Campus-Verlag, Frankfurt.

Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. (2003). On the bursty evolution of blogspace. In *Proc. of the 12th international conference on World Wide Web (WWW'03)*, pages 568–576, New York. ACM Press.

Kumar, R., Novak, J., Raghavan, P., and Tomkins, A. (2004). Structure and evolution of blogspace. *Communications of the ACM*, 47(12):35–39.

Kuperman, V. (2005). Productivity in the internet mailing lists: A bibliometric analysis. *Journal of the American Society for Information Science and Technology*, 57(1):51–59.

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

Larson, R. R. (1996). Bibliometrics of the world-wide web: An exploratory analysis of the intellectual structure of cyberspace. In *Proceedings of the Annual Meeting of the American Society for Information Science, Baltimore, Maryland*.

Leicht, E. A., Holme, P., and Newman, M. E. J. (2006). Vertex similarity in networks. *Physical Review E*, 73:026120.

Leopold, E. (2005). On semantic spaces. *LDV Forum*, 20(1):63–86.

Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187, New York. ACM Press.

Leydesdorff, L. (2001). *The Challenge of Scientometrics. The Development, Measurement, and Self-Organization of Scientific Communications*. Universal Publishers, USA.

Li, W.-S., Candan, K. S., Vu, Q., and Agrawal, D. (2002). Query relaxation by structure and semantics for retrieval of logical web documents. *IEEE Transactions on Knowledge and Data Engineering*, 14(4):768–791.

Li, W.-S., Kolak, O., Vu, Q., and Takano, H. (2000). Defining logical domains in a web site. In *Proc. of the 11th ACM on Hypertext and Hypermedia*, pages 123–132.

Li, X., Thelwall, M., Wilkinson, D., and Musgrove, P. (2005a). National and international university departmental web site interlinking. part 1: Validation of departmental link analysis. *Scientometrics*, 64(2):151–185.

Li, X., Thelwall, M., Wilkinson, D., and Musgrove, P. (2005b). National and international university departmental web site interlinking. part 2: Link patterns. *Scientometrics*, 64(2):187–208.

Lin, J. and Halavais, A. (2004). Mapping the blogosphere in america. In *Proceedings of the Workshop on the Weblogging Ecosystem at the 13th international conference on World Wide Web (WWW'04)*, New York.

Lueg, C. and Fisher, D. (2003). *From Usenet to CoWebs. Interacting with Social Information Spaces*. Springer, London.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, Massachusetts.

Martin, J. R. (1992). *English Text. System and Structure*. John Benjamins, Philadelphia.

Mehler, A. (2005). Zur textlinguistischen Fundierung der Text- und Korpuskonversion. *Sprache und Datenverarbeitung*, 1:29–53.

Mehler, A. (2006). Text linkage in the wiki medium — a comparative study. In Karlgren, J., editor, *Proceedings of the EACL Workshop on New Text — Wikis and blogs and other dynamic text sources, April 3-7, 2006, Trento, Italy*, pages 1–8.

Mehler, A. (2007). Compositionality in quantitative semantics. A theoretical perspective on text mining. In Mehler, A. and Köhler, R., editors, *Aspects of Automatic Text Analysis*, Studies in Fuzziness and Soft Computing, pages 139–167. Springer, Berlin/New York.

Mehler, A. and Gleim, R. (2005). Polymorphism in generic web units. A corpus linguistic study. In *Proceedings of Corpus Linguistics '05, July 14-17, 2005, University of Birmingham, Great Britian*, volume Corpus Linguistics Conference Series 1(1).

Mehler, A. and Gleim, R. (2006). The net for the graphs – towards webgenre representation for corpus linguistic studies. In Baroni, M. and Bernardini, S., editors, *WaCky! Working Papers on the Web as Corpus*, pages 191–224. Gedit, Bologna.

Mehler, A. and Wolff, C., editors (2005). *Text Mining*, volume 20(1) of *LDV Forum – Zeitschrift für Computerlinguistik und Sprachtechnologie*.

Meinel, C. and Sack, H. (2004). *WWW*. Springer, Berlin.

Melnikov, O., Sarvanov, V., Tyshkevich, R., and Yemelichev, V. (1998). *Exercises in Graph Theory*. Kluwer, Dordrecht.

Menczer, F. (2004). Lexical and semantic clustering by web links. *Journal of the American Society for Information Science and Technology*, 55(14):1261–1269.

Mika, P. (2005). Ontologies are us: A unified model of social networks and semantics. In *Proceedings of the 4th International Semantic Web Conference (ISWC 2005)*, LNCS 3729, pages 1–18. Springer.

Milgram, S. (1967). The small-world problem. *Psychology Today*, 2:60–67.

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. J. (1990). Introduction to wordnet: an on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., and Alon, D. C. U. (2002). Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827.

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

Mukherjea, S. (2000). Organizing topic-specific web information. In *Proc. of the 11th ACM Conference on Hypertext and Hypermedia*, pages 133–141. ACM.

Nardi, B. A., Schiano, D. J., Gumbrecht, M., and Swartz, L. (2004). Why we blog. *Communications of the ACM*, 47(12):41–46.

Newman, M. E. J. (2000). Models of the small world. *Journal of Statistical Physics*, 101:819–841.

Newman, M. E. J. (2002). Assortative mixing in networks. *Physical Review Letters*, 89(20):208701.

Newman, M. E. J. (2003a). Mixing patterns in networks. *Physical Review E*, 67:026126.

Newman, M. E. J. (2003b). The structure and function of complex networks. *SIAM Review*, 45:167–256.

Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46:323–351.

Newman, M. E. J. and Park, J. (2003). Why social networks are different from other types of networks. *Physical Review E*, 68:036122.

Newman, M. E. J., Watts, D. J., and Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 99(1):2566–2572.

O'Reilly, T. (2005). What is Web 2.0? Design patterns and business models for the next generation of software. `http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html`.

Orlov, J. K. (1982). Dynamik der Häufigkeitsstrukturen. In Orlov, J. K., Boroda, M. G., and Š. Nadarejšvili, I., editors, *Sprache, Text, Kunst. Quantitative Analysen*, pages 82–117. Brockmeyer, Bochum.

Otte, E. and Rousseau, R. (2002). Social network analysis: a powerful strategy, also for the information sciences. *Journal of Information Science*, 28(6):441–454.

Park, H. W. (2003). Hyperlink network analysis: A new method for the study of social structure on the web. *Connections*, 25(1):49–61.

Pennock, D. M., Flake, G. W., Lawrence, S., Glover, E. J., and Giles, C. L. (2002). Winners don't take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(8):5207–5211.

Pinski, G. and Narin, F. (1976). Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing and Management*, 12:297–312.

Polanyi, L. (1988). A formal model of discourse structure. *Journal of Pragmatics*, 12:601–638.

Prime, C., Bassecoulard, E., and Zitt, M. (2002). Co-citations and co-sitations: a cautionary view on an analogy. *Scientometrics*, 54(2):291–308.

Raible, W. (1995). Arten des Kommentierens – Arten der Sinnbildung – Arten des Verstehens. Spielarten der generischen Intertextualität. In Assmann, J. and Gladigow, B., editors, *Text und Kommentar*, pages 51–73. Fink, München.

Rapoport, A. (1953). Spread of information through a population with sociostructural basis: I. Assumption of transitivity. *Bulletin of Mathematical Biophysics*, 15:523–543.

Rapoport, A. (1982). Zipf's law re-visited. In Guiter, H. and Arapov, M. V., editors, *Studies on Zipf's Law*, pages 1–28. Brockmeyer, Bochum.

Ravasz, E. and Barabási, A.-L. (2003). Hierarchical organization in complex networks. *Physical Review E*, 67:026112.

Ravasz, E., Somera, A. L., Mongru, D. A., Oltvai, Z. N., and Barabási, A.-L. (2002). Hierarchical organization of modularity in metabolic networks. *Science*, 297:1551–1555.

Ravichandra Rao, I. K. (1996). Methodological and conceptual questions of bibliometric standards. *Scientometrics*, 35(2):265–270.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal*, B(4):131–134.

Rehm, G. (2002). Towards automatic web genre identification – a corpus-based approach in the domain of academia by example of the academic's personal homepage. In *Proc. of the Hawaii Internat. Conf. on System Sciences*.

Resnik, P. and Smith, N. A. (2003). The web as a parallel corpus. *Computational Linguistics*, 29(3):349–380.

Rousseau, B. and Rousseau, R. (2000). LOTKA: A program to fit a power law distribution to observed frequency data. *Cybermetrics*, 4(1).

Rousseau, R. (1997). Sitations: an exploratory study. *Cybermetrics*, 1(1).

Santamaría, C., Gonzalo, J., and Verdejo, F. (2003). Automatic association of web directories to word senses. *Computational Linguistics*, 29(3):485–502.

Schenker, A., Bunke, H., Last, M., and Kandel, A. (2005). *Graph-Theoretic Techniques for Web Content Mining*. World Scientific, New Jersey/London.

Schmidt, J. (2006). *Weblogs. Eine kommunikationssoziologische Studie*. UVK, Konstanz.

Schmidt, J., Schönberger, K., and Stegbauer, C. (2005). Erkundungen von Weblog-Nutzungen. Anmerkungen zum Stand der Forschung. *kommunikation@gesellschaft*, 6.

Schummer, J. (2004). Multidisciplinarity, interdisciplinarity, and patterns of research collaboration in nanoscience and nanotechnology. *Scientometrics*, 59(3):425–465.

Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science (JASIS)*, 43(9):628–638.

Sengupta, I. N. and Kumari, L. (1991). Bibliometric analysis of AIDS literature. *Scientometrics*, 20(1):297–315.

Sigman, M. and Cecchi, G. A. (2002). Global organization of the WordNet lexicon. *Proceedings of the National Academy of Sciences*, 99(3):1742–1747.

Sigogneau, A. (2000). An analysis of document types published in journals related to physics: Proceeding papers recorded in the science citation index database. *Scientometrics*, 47(3):589–604.

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

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press, Oxford.

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society of Information Science*, 24:265–269.

Small, H. (1999). A passage through science: Crossing disciplinary boundaries. *Library Trends*, 48(1):72–108.

Smith, M. A. (2003). Measures and maps of Usenet. In Lueg and Fisher (2003), pages 47–78.

Steyvers, M. and Tenenbaum, J. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Storrer, A. (2002). Coherence in text and hypertext. *Document Design*, 3(2):156–168.

Strogatz, S. H. (2001). Exploring complex networks. *Nature*, 410:268–276.

Stubbs, M. (1996). *Text and Corpus Analysis. Computer-Assisted Studies of Language and Culture*. Blackwell, Cambridge, Massachusetts.

Stubbs, M. (2001). *Words and Phrases. Corpus Studies of Lexical Semantics*. Blackwell, Oxford.

Stubbs, M. (2006). Inferring meaning: Text, technology and questions of induction. In Mehler, A. and Köhler, R., editors, *Aspects of Automatic Text Analysis*, volume 209 of *Studies in Fuzziness and Soft Computing*, chapter Part IV — Corpus Linguistic and Text Technological Modeling, pages 233–253. Springer, Berlin.

Tajima, K. and Tanaka, K. (1999). New techniques for the discovery of logical documents in web. In *Internat. Symposium on Database Applications in Non-Traditional Environments*, pages 125–132. IEEE.

Tang, R. and Thelwall, M. (2004). Patterns of national and international web inlinks to US academic departments: An analysis of disciplinary variations. *Scientometrics*, 60(3):475–485.

Thelwall, M. and Tang, R. (2003). Disciplinary and linguistic considerations for academic web linking: An exploratory hyperlink mediated study with mainland china and taiwan. *Scientometrics*, 58(1):155–181.

Thelwall, M., Vaughan, L., and Björneborn, L. (2006). Webometrics. *Annual Review of Information Science Technology*, 6(8).

Thelwall, M. and Wouters, P. (2005). What's the deal with the web/blogs/the next big technology: A key role for information science in e-social science research? In *Proc. of the 5th Int. Conf. on Conceptions of Library and Information Sciences (CoLIS'05)*, pages 187–199.

Tricas, F., Ruiz, V., and Merelo, J. J. (2004). Do we live in a small world? Measuring the Spanish-speaking blogosphere. In *Blogtalk 2.0, June 5-6, Wien*.

Tuldava, J. (1998). *Probleme und Methoden der quantitativ-systemischen Lexikologie*. Wissenschaftlicher Verlag, Trier.

van Dijk, T. A. and Kintsch, W. (1983). *Strategies of Discourse Comprehension*. Academic Press, New York.

van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, 59(3):467–472.

van Raan, A. F. J. (2005). Reference-based publication networks with episodic memories. *Scientometrics*, 63(3):549–566.

Ventola, E. (1987). *The Structure of Social Interaction: a Systemic Approach to the Semiotics of Service Encounters*. Pinter, London.

Voss, J. (2005). Measuring Wikipedia. In *Proceedings of the 10th International Conference of the International Society for Scientometrics and Informetrics*.

Wasserman, S. and Faust, K. (1999). *Social Network Analysis. Methods and Applications*. Cambridge University Press, Cambridge.

Watts, D. J. (1999). *Small Worlds. The Dynamics of Networks between Order and Randomness*. Princeton University Press, Princeton.

Watts, D. J. (2003). *Six Degrees. The Science of a Connected Age*. W. W. Norton & Company, New York/London.

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

White, H. D. and McCain, K. W. (1989). Bibliometrics. *Annual Review of Information Science Technology (ARIST)*, 24:119–165.

Widdows, D. and Dorow, B. (2002). A graph model for unsupervised lexical acquisition. In *19th International Conference on Computational Linguistics, August 24 – September 1, 2002, Taipeh, Taiwan*.

Wimmer, G. and Altmann, G. (1999a). Review article: On vocabulary richness. *Journal of Quantitative Linguistics*, 6(1):1–9.

Wimmer, G. and Altmann, G. (1999b). *Thesaurus of univariate discrete probability distributions*. Stamm Verlag, Essen.

Zelman, A. and Leydesdorff, L. (2000). Threaded email messages in self-organization and science & technology studies oriented mailing lists. *Scientometrics*, 48(3):361–380.

Zipf, G. K. (1972). *Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology*. Hafner Publishing Company, New York.

Zlatic, V., Bozicevic, M., Stefancic, H., and Domazet, M. (2006). Wikipedias: Collaborative web-based encyclopedias as complex networks. *Physical Review E*, 74:016115.