

In Search of a Bridge Between Network Analysis in Computational Linguistics and Computational Biology – A Conceptual Note

Alexander Mehler

Department of Computational Linguistics & Text Technology
Bielefeld University
D-33615 Bielefeld, Germany
Email: Alexander.Mehler@uni-bielefeld.de

Abstract—Recently, the inference of biological networks has been studied whose vertices represent proteins and recurrent sequential patterns – called domain types – thereof; cf., for example, [1]. What makes this an outstanding research object from the point of view of data mining is the explorative analysis of large networks whose emergence is simulated in order to get insights into the dynamics of the focal area. This research program is connected to analyzing informational and, especially, textual networks as explored in the area of text mining. Consequently, the question is put forward which graph representation model is common to both areas of investigation. This paper addresses this question from the perspective of network motifs [2]. It reconstructs this notion from the point of view of syntagmatic and paradigmatic learning in computational linguistics. As an epiphenomenon, the paper sheds light on the applicability of text mining procedures in the area of bioinformatics.

I. INTRODUCTION

Complex network analysis is a prominent topic not only in computational sociology [3] but also in computational biology [2] and computational linguistics [4]. This concerns *synthetical* aspects as the inference of large gene networks in computational biology [1] and, alternatively, of web pages in computer science [5] or of lexical networks in computational linguistics [6]. From an *analytical* point of view, this common interest regards small world characteristics [3] as well as the decomposition and classification of large networks [7] – see, for example, the review in [2]. What makes network analysis an outstanding research object from the point of view of data mining is the explorative analysis of large networks whose emergence is simulated in order to get insights into the dynamics of the focal area. This research program is connected with analyzing informational and, especially, textual networks as explored in the area of text mining. From this perspective, there are two candidates which allow bridging network analysis in information science and computational biology:

- Large document networks of digital libraries in the area of biology and related disciplines can be made input to text mining [8] in order to explore heretofore unknown information about networking of biological units [9], [10].

- Alternatively, we may build on the homology of informational and biological network analyses. Barabasi & Oltvai [2, p.101] claim, for example, that “the architectural features of molecular interaction networks within a cell are shared to a large degree by other complex systems, such as the Internet [...] and society.” Consequently, the apparatus of complex network analysis is seen to be common to its different fields of application, irrespective of whether we deal with biological, technical or linguistic units.

This paper is in the line of the second of these bridges. It aims at contributing to answering the question which graph model is common to the different areas of complex network analysis. Evidently, at least for the time being, this is only a methodical but not an ontological bridge. That is, we expect that biological and informational networks vary significantly with respect to their topological characteristics. Newman & Park [11] show, for example, that social networks are different from technical ones in the sense that the former, but not the latter tend to show assortative mixing. That is, nodes (actors, authors, etc.) in social networks are more probably linked with likewise linked nodes. Mehler [4] shows that Wiki-based text networks are characterized by disassortative mixing and, thus, behave more like technical networks. Beyond such indices of structure formation in networks, community building is also a common topic in network analysis. These communities are, amongst others, described in terms of recurrent sequences, motifs and large subgraphs [3], [12], [13] as candidate manifestations of complex functional units. From this perspective, the following questions arise:

- *What does an appropriate representation model look like which allows integrating the different fields of network analysis?*
- *What field-specific models have to be included?*
- *To what extent can computational linguistics, information science and computational biology share methods?*

This paper addresses these questions from a conceptual perspective. It builds on the hypothesis that because of the homology of network analyses in the different areas con-

cerned, structural linguistics comes into play as a reference point of extending the notion of a motif. The basic idea is to look not only for small, recurrent subgraphs composed of a couple of nodes, but to explore a hierarchical graph model in which nodes on a higher level may be composed of lower level motifs. This conception is motivated by the graph model of Ravasz et al. [14] which allows identifying the formation of hierarchical structures in large networks. In this sense, the present paper is in search of functional network units based on the hypothesis that these units are recurrently and reliably manifested by patterns of subgraphs. As an epiphenomenon, the paper sheds light on the applicability of some methods of computational linguistics to the area of bioinformatics.

The paper is organized as follows: Section II motivates the concept of a motif as a bridge between linguistic and biological network analysis. Sections II.A and II.B describe syntagmatic and paradigmatic learning as a reference point of reconstructing motifs from the point of view of computational linguistics. Further, Section II.C introduces an algorithm for the detection of syntagmatic and paradigmatic regularities of motifs. Finally, Section III gives a conclusion and a prospect of future work.

II. THE SYSTEMIC PERSPECTIVE: SYNTAGMATICS VS. PARADIGMATICS

The majority of network models presented in the literature on information networks are positivistic in the sense that they do not rely on functionally or topically explored compound nodes (e.g. functional devices or text patterns) and their edges (indicating processing circuits or information flow etc.), but start from formally demarcated elementary units (e.g. genes or web pages) and their untyped structural links forming, e.g., sequences. In other words, they miss a notion of structure formation based on recurrent patterns by analogy with the notion of motifs [12] in complex networks. As will be explained in this section, such a notion naturally arises from the linguistic opposition of syntagmatic and paradigmatic learning. Its relevancy is, amongst others, connected to the concept of *duplication*. This concept is referred to in order to explain the scale-free topology of gene networks [2]. It allows bridging biological and informational network analysis as follows: Interactions of textual units are partly due to their typological intertextuality, that is, to the fact that they replicate, so to speak, the same or related text patterns and, thus, are interrelated from the point of view of their form, function or content (due to the functional, topical or simply formal nature of the patterns being replicated).¹ In other words, interactions in text networks are partly due to the *replication of textual patterns* where the emergence of these patterns is, in turn, related to the contiguity and similarity patterns of linguistic units manifested by text constituents down to the level of tokens. This connection between text linkage on the one hand and contiguity as well as similarity relations of textual

constituents on the other hand can be utilized as a starting point for deriving analogical models of structure formation in biological networks. It opens a perspective on exploring network patterns as it demands to distinguish between (i) code systematic patterns and (ii) their text-internal manifestations, i.e. replications. In this section, we shed light on modeling syntagmatic and paradigmatic learning and thereby stress the need for more elaborated network models which include this two-stage model of structure formation. This is done by outlining the development of the linguistic model in question, secondly, by describing its operationalization in machine learning and, thirdly, by extending it to the area of learning patterns in graphs.

A. A brief account of syntagmatic and paradigmatic learning

In order to give a brief account of syntagmatic and paradigmatic learning we start with de Saussure [16] who adopted the notion of association by *similarity* and of association by spatio-temporal *contiguity*. More specifically, he described *syntagmatic* or conjunctive relations between constituents of the same compound unit in opposition to *disjunctive* relations of substitutability. This concept was specified in glossematics [17] where syntagmatic relations are seen to hold between (groups of) linguistic items co-occurring in instances of the same (syntactic) context type, though not necessarily side by side. In contrast to this, associative relations are seen to hold between items which are substitutable for each other within the same contexts (of a certain type under consideration) without changing their focal syntactical, semantical or pragmatical properties (e.g. grammaticality, well-formedness, acceptability etc.). What makes these notions relevant from the point of view of machine learning is that de Saussure identified syntagmatic relations as a source of the formation of similarity associations which were later called *paradigmatic* and reconstructed in terms of substitutability under invariance of certain linguistic features [17]. Starting from this hypothesis, it is evident to claim that the similarities of the syntagmatic relations, into which linguistic items enter, contribute to their paradigmatic similarity.

The further development of de Saussure's model focused solely on structural considerations. This was done by Hjelmslev [17] who opposed *syntagmatic*, i.e. *text*-based relations by *paradigmatic*, i.e. *system*-based relations. In order to do that he invented the glossematic notion of a *function*: Starting from the concept of a functor as an argument of a linguistic function, Hjelmslev defined *constants* to be functors whose presence is a necessary condition of the presence of those functors, with which they enter into the same function [17]. In contrast to this, he defined *variables* to be functors whose presence is not a sufficient condition in this sense. Utilizing the notion of a constant and variable, Hjelmslev distinguished three types of syntagmatic (conjunctive) and paradigmatic (disjunctive) functions in order to define *paradigms as classes of units which enter into homogeneous functions*. Consequently, he distinguished *language systematic* paradigms from *text internal* syntagmata and, thus, separated two reference points of

¹In linguistics, this notion is, amongst others, known under the heading of *typological intertextuality* – cf. [15].

structure formation: within the language system and within its textual instances.

This is what we believe to be a *conditio sine qua non* for seeking a bridge between the different branches of complex network analysis, that is, looking for paradigmatic patterns within the focal code of functional, topical or structural types subject to syntagmatic patterns within the observable networks instantiating this code.

It is also worth noting that Hjelmslev left the narrow limitation of syntagmatics to sub-sentential units and focused instead on whole texts and, thus, hinted on applying syntagmatics and paradigmatics on whatever stratum of linguistic resolution. Nevertheless, Hjelmslev did not operationalize his concept of syntagmatics and paradigmatics in terms of machine learning.

B. A machine learning perspective on syntagmatic and paradigmatic learning

Such a machine learning-oriented reconstruction was done in *computational linguistics*—cf. [18], [19] and [20]. It is based on the hypothesis that the similarity relations of units result from a *two-level* process of inductive learning starting from the units' contiguity relations. In the case of lexical items, Landauer & Dumais [20], for example, equate these contiguity relations with co-occurrence relations. More specifically, their approach to *Latent Semantic Analysis* (LSA) describes the learning of similarity relations as a process of dimensionality reduction in terms of singular value decomposition [21]. This process allows detecting similarities of cognitive items even if they rarely co-occur within the input stream of the learning system.

Landauer & Dumais describe similarity associations of linguistic items as functions of their contiguity associations. According to this model, inductive learning of similarity relations of linguistic items results from exploiting the similarities of their usage contexts according to the *weak contextual hypothesis* [22] which says that the similarity of the contextual representations of words contributes to their semantic and, as we may add, functional similarity. A central merit of this model is that it allows learning *indirect* and, thus, generalized similarity relations:

- It allows to interrelate elementary items even if they do not or only rarely co-occur, but tend to be used in similar contexts.
- Further, it allows to interrelate compound units even if they do not share any or only few constituents, but whose constituents are recursively similar according to the present or, in the case that they are elementary, the latter claim.

For the time being, this model disregards typing of contiguity and similarity relations. This is the place where Hjelmslev's notion of glossematic functions comes into play as a way out of this indifference – cf. [23]. Nevertheless, although this model proved to be highly successful with respect to learning lexical patterns, it does not account for learning syntagmatic patterns above the level of elementary units up to the level of whole texts. Actually, these more complex patterns underly

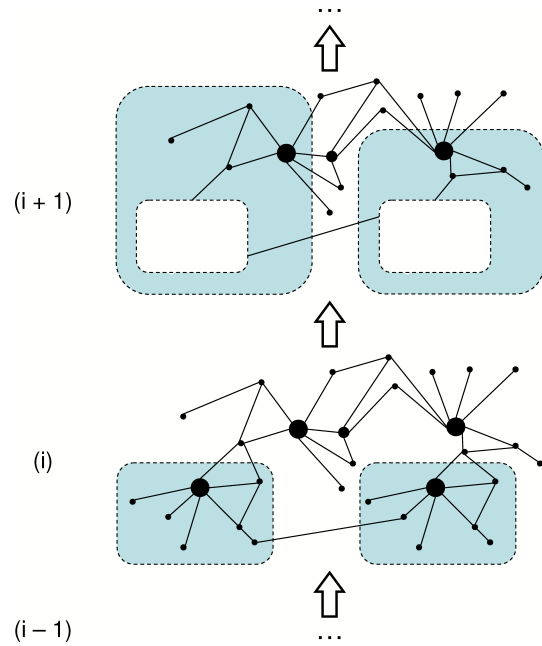


Fig. 1. Schematic representation of two passes of an algorithm of recursive motif induction.

what has been called typological intertextuality as a result of pattern reduplication. Although Landauer & Dumais [20] apply their model to learning lexical associations and text similarities only, it is evident how to extend it for learning higher level syntactic patterns. This is outlined in the following section.

C. Motifs in a paradigmatic perspective

A more general account of syntagmatics and paradigmatics above the level of lexical items was proposed by Solan et al. [24] who build on Harris' distributional hypothesis in order to develop an algorithm of grammar induction. This hypothesis claims that distributional regularities correlate with some aspects of their meaning in a way that semantic or, as we add again, functional differences are reflected by dissimilar distributions and vice versa. Solan et al. recursively apply this distributional hypothesis by attributing units which occur in the same or alike contexts to belong to the same category. Starting from a stream of input sentences, their algorithm recursively learns distributional categories, their contiguity and similarity relations based on its output in the preceding step. Solan et al. propose a two-level algorithm in which, firstly, a graph representation model is build of the input stream whose constituents are, secondly, learned by means of a *Pattern Acquisition Algorithm* (PAA). The PAA essentially detects recurrent sequences within the input graph which are subsequently classified into equivalence classes, i.e. paradigms of alike constituents within the generalization step. This algorithm bootstraps more and more complex paradigms as nodes of a graph whose edges denote syntagmatic relations which generalize over more and more relations of elementary vertices.

Because of its generality, this algorithm should allow looking forward at learning equivalence classes of subgraphs in complex networks. As a starting point for extending this kind of syntagmatic and paradigmatic learning to graph pattern induction, we refer to the notion of a *motif*. Roughly spoken, a motif of a network G is a subgraph pattern which is instantiated in G more often than expected by chance, that is, than observable in the corresponding random network—cf. Milo et al. [12]. Motifs are, so to speak, basic building blocks of large networks, that is, patterns of node linkage which recurrently occur in many different parts of them [2], [13]. A basic assumption of Shen-Orr et al. [13] is that motifs can be functionally identified: they are seen to carry functions as “elementary computational circuits” [12]. In this sense, a motif is a class of candidate manifestations of a functional unit.

This conception can be made a starting point for reconstructing motifs in terms of syntagmatic and paradigmatic learning higher-order candidates of network devices. Generally speaking, this algorithm works iteratively in the sense that it tries to detect in each processing step i motifs of the present class of networks (by analogy with a corpus of texts) which are referred to as single nodes in the subsequent step $i + 1$. As a consequence, this pattern induction algorithm induces hierarchical graphs, that is, graphs whose nodes may be graphs on their own whereby a superordinate node inherits its links from its subordinate nodes. Figure (1) gives a schematic sketch of such an iteration.

In order to formally specify the latter algorithm we, first, define motifs and their instantiations as a prerequisite of a model of hierarchical graphs based thereon:²

Definition 1. Motifs and their instances. Let X be a set of $n \geq m > 2$ elements and \mathbb{M} the set of all simple directed graphs (MV, ME) such that $m \leq \text{Card}(MV) \leq n$ and $ME \subseteq MV^2 \subseteq X^2$. This allows to define $\mathcal{A} = \{|\mathcal{G}| \mid \mathcal{G} \in \mathbb{X}\}$ as the set of all equivalence classes $|\mathcal{G}|$ induced by the equivalence relation $R \subseteq \mathbb{X}^2$ where $(\mathcal{G}_i, \mathcal{G}_j) \in R$ iff there exists an isomorphism g between \mathcal{G}_i and $\mathcal{G}_j \in \mathbb{X}$. (The elements of \mathcal{A} will be referred to as denoting graph patterns manifested by subgraphs of the focal input networks.) Let now $G = (V, E)$ be a simple directed graph representing a complex network and $P(G) = (V, E)$ be a corresponding random graph according to some random graph model (e.g. according to [25]). Let further $M(G)$ be the set of all connected subgraphs $G' = (V', E'), m \leq V' \leq n$, of G such that there exists an isomorphism h interrelating G' with some element of \mathbb{M} – in order to guarantee uniqueness, this isomorphism will be noted as $h_{G'}$. For any subgraph pattern $|\mathcal{G}| \in \mathcal{A}$, this allows to define the number of its preimages in G as

$$\text{conf}(|\mathcal{G}|, G) = \text{Card}(\{G' \in M(G) \mid h_{G'}(G') = \mathcal{G}\})$$

We call $|\mathcal{G}|$ an $(m, n, P(G))$ -*motif* or simply a *motif* in G if $\text{conf}(|\mathcal{G}|)$ is higher than expected by chance, that is, if $\text{conf}(|\mathcal{G}|, G)$ is significantly higher than $\text{conf}(|\mathcal{G}|, P(G))$

²To keep things simple, we only deal with simple directed graphs.

according to a corresponding confidence interval. We define $\mathbb{M}(G)$ as the set of all such motifs of G . Finally, we call $G' \in M(G)$ with $h_{G'}(G') = |\mathcal{G}| \in \mathbb{M}(G)$ an *instance* of $|\mathcal{G}|$ in G . \square

Obviously, $M(G)$ is the set of all instances of motifs in G . Moreover, as any of these instances is uniquely mapped onto its corresponding motif $|\mathcal{G}| \in \mathbb{M}(G)$, $|\mathcal{G}|$ can be conceived as the paradigm of these instances within G .

Now, we can define a class of hierarchical graphs which are iteratively built by replacing in each iteration step the motifs of the corresponding input graph by means of single nodes and then repeating this procedure till no more motifs are found:

Definition 2. Let $G = (V, E)$ be a simple directed graph with the set of motifs $\mathbb{M}(G)$ and their instances $M(G)$ according to definition (1). Then we iteratively derive graphs thereof as follows:

$$G_0 = (V_0, E_0) = G; M(G_0) = M(G).$$

:

$$G_i = (V_i, E_i) \text{ with the vertex set}$$

$$V_i = V_{i-1} \setminus v(M(G_{i-1})),$$

$$v(M(G_{i-1})) = \cup_{(V', E') \in M(G_{i-1})} V', \text{ and the edge set}$$

$$\begin{aligned} E_i = & E_{i-1} \setminus \{(v, w) \mid \exists (V', E') \in M(G_{i-1}) : (v, w) \in E'\} \\ & \cup \{((V', E'), w) \mid (V', E') \in M(G_{i-1}) \wedge \\ & w \in V_{i-1} \wedge w \notin v(M(G_{i-1})) \wedge \\ & \exists v \in V' : (v, w) \in E_{i-1}\} \\ & \cup \{(v, (V', E')) \mid (V', E') \in M(G_{i-1}) \wedge \\ & v \in V_{i-1} \wedge v \notin v(M(G_{i-1})) \wedge \\ & \exists w \in V' : (v, w) \in E_{i-1}\} \\ & \cup \{((V', E'), (W, D)) \mid (V', E'), (W, D) \in M(G_{i-1}) \\ & \wedge \exists (v, w) \in E_{i-1} : v \in V' \wedge w \in W\} \end{aligned} \quad \square$$

These two preliminary definitions of recursive motif development in complex networks abstract from node and link types and, thus, underachieve structure formation at least in linguistic networks where typing is indispensable. Wiki-based networks as a kind of textual networks based on social tagging, for example, have a very rich system of node and link types [4]. Actually, the typological deficit of the present definitions is straightforwardly put aside by referring, for example, to the notion of function in terms of Hjelmlevian glossematics. But in order to make this a quantitative statement with respect to the amount such a typed dependency relation holds among elementary or complex nodes in hierarchical graphs described by definition (2), we need to refer to corpora of networks as the underlying data pool for estimating these quantities.³ Once more, this would be done in strict accordance of dealing with either linguistic or biological networks as in both cases motifs prove to be a valuable concept. How far this analogy actually reaches is a question of empirical estimation and will be part of future experiments in the field.

³As this would demand to deal with dependencies of graph patterns, it would naturally lead to some kind of data oriented parsing [26].

III. CONCLUSION

We described a generalization of the notion of a motif in complex networks. This was done by means of a hierarchical graph model. Its starting point is the hypothesis of a homology of linguistic and biological networks, at least in methodological terms. We referred to the linguistic notion of syntagmatic and paradigmatic learning in order to distinguish patterns of linkage and substitutability of motifs. As described by Ravasz et al. [14] by example of gene networks, hierarchical models of motifs are a valuable concept in computational biology. Thus, nested motifs can be seen to enrich the commonalities of network analyses in computational linguistics and biology. Future work aims at empirically testing the present model by means of informational and biological networks.

ACKNOWLEDGMENT

The author would like to thank Matthias Dehmer and Frank Emmert-Streib for motivating efforts in bridging biological and linguistic network analysis.

REFERENCES

- [1] I. Iossifov, M. Krauthammer, C. Friedman, V. Hatzivassiloglou, J. S. Bader, K. P. White, and A. Rzhetsky, "Probabilistic inference of molecular networks from noisy data sources," *Bioinformatics*, vol. 20, no. 8, pp. 1205–1213, 2004.
- [2] A.-L. Barabási and Z. N. Oltvai, "Network biology: Understanding the cell's functional organization," *Nature Reviews. Genetics*, vol. 5, no. 2, pp. 101–113, 2004.
- [3] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.
- [4] A. Mehler, "Text linkage in the wiki medium – a comparative study," in *Proceedings of the EACL Workshop on New Text – Wikis and blogs and other dynamic text sources, Trento, Italy, April 3-7, 2006*. [Online]. Available: http://www.sics.se/jussi/newtext/working_notes/01_mehler.pdf
- [5] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.
- [6] M. Steyvers and J. Tenenbaum, "The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth," *Cognitive Science*, vol. 29, no. 1, pp. 41–78, 2005.
- [7] F. Emmert-Streib and M. Dehmer, "A systems biology approach for the classification of dna microarray data," in *Proceedings of ICANN 2005, Poland/Torun*, 2006.
- [8] M. A. Hearst, "Untangling text data mining," in *Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics, University of Maryland, 1999*, 1999.
- [9] A. Rzhetsky, I. Iossifov, T. Koike, M. Krauthammer, P. Kra, M. Morris, H. Yu, P. A. Duboué, W. Weng, W. J. Wilbur, V. Hatzivassiloglou, and C. Friedmann, "Geneways: a system for extracting, analyzing, visualizing, and integrating molecular pathway data," *Journal of Biomedical Informatics*, vol. 37, pp. 43–53, 2004.
- [10] M. Cokol, I. Iossifov, C. Weinreb, and A. Rzhetsky, "Emergent behavior of growing knowledge about molecular interactions," *Nature Biotechnology*, vol. 23, no. 10, pp. 1243–1247, 2005.
- [11] M. E. J. Newman and J. Park, "Why social networks are different from other types of networks," *Physical Review E*, vol. 68, p. 036122, 2003.
- [12] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, and D. C. U. Alon, "Network motifs: simple building blocks of complex networks," *Science*, vol. 298, no. 5594, pp. 824–827, 2002.
- [13] S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network motifs in the transcriptional regulation network of escherichia coli," *Nature Genetics*, vol. 31, no. 1, pp. 64–68, 2002.
- [14] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási, "Hierarchical organization of modularity in metabolic networks," *Science*, vol. 297, pp. 1551–1555, 2002.
- [15] W. Heinemann, "Zur Eingrenzung des Intertextualitätsbegriffs aus textlinguistischer Sicht," in *Textbeziehungen: linguistische und literaturwissenschaftliche Beiträge zur Intertextualität*, J. Klein and U. Fix, Eds. Tübingen: Stauffenburg, 1997, pp. 21–37.
- [16] F. de Saussure, *Grundfragen der allgemeinen Sprachwissenschaft*. Berlin/New York: de Gruyter, 1967.
- [17] L. Hjelmslev, *Prolegomena to a Theory of Language*. Madison: University of Wisconsin Press, 1969.
- [18] B. Rieger, "Situation semantics and computational linguistics: towards informational ecology," in *Information. New Questions to a Multidisciplinary Concept*, K. Kornwachs and K. Jacoby, Eds. Berlin: Akademie-Verlag, 1995, pp. 285–315.
- [19] H. Schütze, *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*, ser. CSLI Lecture Notes. Stanford: CSLI Publications, 1997, vol. 71.
- [20] T. K. Landauer and S. T. Dumais, "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge," *Psychological Review*, vol. 104, no. 2, pp. 211–240, 1997.
- [21] E. Leopold, "On semantic spaces," *LDV Forum*, vol. 20, no. 1, pp. 63–86, 2005.
- [22] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and Cognitive Processes*, vol. 6, no. 1, pp. 1–28, 1991.
- [23] A. Mehler, "Preliminaries to an algebraic treatment of lexical associations," in *Proceedings of the Workshop Learning and Extending Lexical Ontologies at the 22nd International Conference on Machine Learning (ICML 2005), Bonn, Germany, August 7-11, 2005*.
- [24] Z. Solan, E. Ruppim, D. Horn, and S. Edelman, "Automatic acquisition and efficient representation of syntactic structures," in *Advances in Neural Information Processing*, S. Thrun, Ed. Cambridge: MIT Press, 2003, vol. 15.
- [25] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, and U. Alon, "Subgraphs in random networks," *Physical Review E*, vol. 68, p. 026127, 2003. [Online]. Available: <http://www.citebase.org/cgi-bin/citations?id=oai:arXiv.org:cond-mat/0302375>
- [26] R. Bod, *Beyond Grammar. An Experience-Based Theory of Language*. Stanford: CSLI Publications, 1998.