Quantifying Semantics Using Complex Network Analysis

Chris Biemann *Stefanie Roos* *Karsten Weihe*
Computer Science Department
Technische Universität Darmstadt
Hochschulstr. 10
64289 Darmstadt, Germany
{biem,weihe}@cs.tu-darmstadt.de, stefanie.roos@cased.de

ABSTRACT
Though it is generally accepted that language models do not capture all aspects of real language, no adequate measures to quantify their shortcomings have been proposed until now. We will use n-gram models as workhorses to demonstrate that the differences between natural and generated language are indeed quantifiable. More specifically, for two algorithmic approaches, we demonstrate that each of them can be used to distinguish real text from generated text accurately and to quantify the difference. Therefore, we obtain a coherent indication how far a language model is from naturalness.
Both methods are based on the analysis of co-occurrence networks: a specific *graph cluster measure*, the *transitivity*, and a specific kind of *motif analysis*, where the frequencies of selected *motifs* are compared. In our study, artificial texts are generated by n-gram models, for $n = 2, 3, 4$. We found that, the larger n is chosen, the narrower the distance between generated and natural text is. However, even for $n = 4$, the distance is still large enough to allow an accurate distinction.
The motif approach even allows a deeper insight into those semantic properties of natural language that evidently cause these differences: polysemy and synonymy.
To complete the picture, we show that another motif-based approach by Milo et al. (2004) does not allow such a distinction.
Using our method, it becomes possible for the first time to measure generative language models deficiencies with regard to semantics of natural language.

KEYWORDS: quantitative linguistics, network analysis, motif analysis, co-occurrence networks, language models.

Proceedings of COLING 2012: Technical Papers, pages 263–278,
COLING 2012, Mumbai, December 2012.

1 Introduction

Language models are used in many text processing systems (e.g. machine translation, document classification, language generation etc.) and are undoubtedly a standard building block of natural language processing. However, there exist hardly any methods that characterize language models quantitatively, in order to measure their deficiencies with respect to real language: the commonly used perplexity measure is known to be insensitive to semantic aspects (Chang et al., 2009). To this end, we propose an automatic approach that not only can distinguish language-model-generated text from real text, it also quantifies their distance. Moreover, we can quantify language model shortcomings with respect to two semantic phenomena of natural language, namely polysemy and synonymy. The ability to measure, to what extent language models in fact model these and other characteristics of real language, is a prerequisite for improving language models to closer adhere to the many-layered structure of natural language.

Before laying out the details of our approach, we would like to briefly sketch the general idea: If we accept that different language models capture natural language semantics to different extents, and if we had a measure that indicates and quantifies this extent, then this measure enables us to drive the development of language models towards a better reflection of semantics. We propose such a measure, which is based on the assumption that the structure of the co-occurrence graph of a (real or generated by a language model) text reflects semantic adequateness. Computing and comparing different characteristics of these co-occurrence graphs allows us to quantify the differences.

This paper summarizes a series of computational studies to characterize the uniqueness of co-occurrence networks from real natural language, as opposed to co-occurrence networks from artificial "natural" language. We developed two testing methods to decide whether a corpus of text was written by humans or generated by a language model. For that, we analyze the structural difference of the co-occurrence networks induced by real text and generated text, respectively. As generative language models, we chose the 2-gram, 3-gram, and 4-gram models.

We examine co-occurrence graphs.[1] In that, we consider two types of co-occurrence: sequential occurrence between neighbors and co-occurrence within a whole sentence. Our measures are based on two network metrics. Traditionally, the structures of language networks are analyzed either on the global level or on the level of single nodes and edges (e.g. the degree distribution). Our first method is based on a global clustering metric, the *transitivity*.

Our second method is based on *motif analysis*. This approach addresses an intermediate level, where a structural entity is composed of a small number of nodes and edges. Networks are compared by counting the number of times a certain *k-node motif*, i.e. a small graph of k nodes, appears as an induced subgraph. In case of the directed sequential co-occurrence graphs, 3-node motifs are analyzed (see Fig. 1), whereas the bidirectional sentence co-occurrence networks are compared based on their 4-node motifs (see Fig. 2).

In case of sentence-based co-occurrence, each of these methods enables us to distinguish natural from generated text and to quantify the differences in a reasonable way. Such a quantification can be viewed as an indicator how far a language model is from naturalness. In fact, it turns out that this indicator conforms well to what one would expect: higher-order n-gram models generate a better approximation of real texts. However, no n-gram model is able to capture long-

[1] In accordance with the established terminology, we will use the terms *graph* and *network* synonymously, and choose the appropriate word in view of the respective context.

range semantic dependencies well, which we will exploit in our analysis techniques. Moreover, we will show that motifs can be related to specific semantic properties of natural languages that do not occur in *n*-gram generated text and hence explain the observed differences to a large extent. Here, the above-mentioned domain-specific phenomena are of a semantic kind. In particular, we show that two specific phenomena, *polysemy* and *synonymy*, are reflected by the counts of two motifs, the *chain motif* and the *box motif* (#2 and #4 in Fig. 2).

In summary, the presented work follows a new, successful, path and opens new, promising, perspectives on the analysis of language models. This is the first application of motif analysis to language networks and their underlying semantics. So, besides the specific computational results and the novel method to obtain them, the presented work is also relevant due to this general methodological innovation.

The paper is organized as follows. In Section 2, we briefly review the most relevant related work. Then, in Section 3, we will describe our approach in full detail. In Section 4, we present and discuss the conjectures that structure our research. In Section 5, we present our results, and finally discuss future perspectives in Section 6.

2 Related Work

2.1 Network Analysis

There is a large scientific body of methods and applications of network analysis (Aggarwal and Wang, 2010; Aggarwal, 2011). Graph mining – the art of detecting and analyzing patterns and structures in graphs – is the specific focus of the surveys (Cook and Holder, 2006; Fortunato, 2010).

It seems reasonable to classify network analysis techniques by the level of granularity they address. Elementary statistical measures such as the node degree distribution operate on the level of single nodes and edges. In the opposite extreme case, on the global level, the structure of a network is captured in a single (scalar) numerical value. Examples for global measures are the average shortest path length, the diameter, as well as simple characteristics such as node and edge count. See the above-mentioned surveys for a systematic discussion.

For our analysis, a global clustering metric, the *transitivity*, is considered, however, our main focus is on *motif analysis*. *Motif analysis* addresses an intermediate level: local structures consisting of a small number of nodes and edges. Networks are compared by comparing the number of occurrences of selected motifs.

Motif analysis has first been investigated in computational biology (Shen-Orr et al., 2002) and has since been applied to a variety of network types in biology and biochemistry (Schreiber and Schwöbbermeyer, 2010). The underlying insight is that biological and biochemical dynamics are statistically related to the occurrence of small functional blocks, which have specific structures. This insight is well captured by motif signatures, and in fact, many computational studies reveal significant relations. Due to this success, it did not take long time until this technique has been applied to networks from other domains. For example, (Milo et al., 2002, 2004) compare networks from biology, electrical engineering, natural language and computer science and find that the motif signatures from different domains are so different that they may serve as a "fingerprint" of the respective domain.

The idea of functional blocks applies in domains beyond biology and biochemistry as well, surprisingly, even in social networks. In (Krumov et al., 2011), we analyzed citation networks,

which we modeled as undirected graphs on the authors. An edge indicates at least one joint publication. In a sense, the citation numbers of individual publications within an occurrence of a motif can be aggregated to a citation number of the entire occurrence. We considered four natural ways for aggregation. Roughly speaking, the main result of (Krumov et al., 2011) is this: the average citation number of the box motif, taken over all occurrences, is statistically significantly larger than expected. This effect occurs for all four ways of aggregation. A deeper look revealed that certain occurrences of the *box motif* (#4 in Fig. 2) explain this result: two "seniors," A and B, have jointly published, A has published with a "junior" C, B with a junior D, and C and D have joint publications as well, but neither A with D nor B with C. Among these occurrences, the ones that serve as "bridges" in the network in a certain sense are particularly responsible for the observed effect.

In view of the outlook (Section 6), we further mention recent work that uses the concept of motifs for other purposes than network analysis. (Krumov et al., 2010a,b) developed an algorithm to optimize the structure of peer-to-peer networks based on local operations only. Each node manipulates the local structure in its vicinity in order to thrive the local motif signature towards the average local motif signature of an optimal network.

2.2 Complex Networks of Natural Language

The structure of natural language networks has been extensively investigated, see e.g. (Masucci and Rodgers, 2006) and references therein.

(Ferrer-i-Cancho and Solé, 2001) have shown that co-occurrence networks of natural language are scale-free small world graphs. Whereas scale-freeness seems to be a consequence of the Zipfian word-frequency distribution (Biemann, 2007), Steyvers and Tenenbaum (2005) find the small-world property in co-occurrence networks and lexical-semantic resources, which indicates that co-occurrence networks reflect semantic properties.

There is only very little work on operationalizing complex network analysis for natural language processing applications. (Pardo et al., 2006) evaluate the quality of automatic summaries by analyzing the degree distributions of networks generated from words at different fixed offsets in the text, and (Amancio et al., 2012) characterize texts for authorship attribution by quantifying their consistency index, which is measured by the number of authors that use content words in a sequence. A related work is (Köhler and Naumann, 2010), where segments of words with increasing length and frequency are used to characterize texts of different authors.[2] We are not aware of any other research that uses network analysis to assess the quality of language models trained from real text.

3 Methodology

Our results have been generated in a three-step process: First, the text needs to be selected, respectively generated, before the graphs can be derived from the texts according to a parameterizable strategy. In the last step, our proposed metrics are evaluated on these graphs. These three steps are explained in detail in Sects. 3.1–3.3, respectively.

[2]**Note:** (Köhler and Naumann, 2010) also use the term 'motifs', but they refer to the aforementioned sequences of words, not to subgraphs as in our work.

3.1 Text Basis

Text corpora For our experiments, we use corpora of different languages of one million sentences each, provided by LCC[3] (Biemann et al., 2007). We use the same corpus of real language for training the n-gram model and for comparison. For comparison between real and generated language, we generate text according to the same sentence length (number of tokens) distribution as found in the respective real language corpus, since we have found in preliminary experiments that co-occurrence network structure is dependent on the sentence length distribution. We have found in further experiments, that the general picture of results is stable for corpora of different sizes, starting from about ten thousand sentences.

Text generation with n-gram models For the scope of this work, we chose n-gram models, which are the standard workhorses of language modeling. A language model assigns a probability to a sequence of words, based on a probabilistic model of language. This can be used to pick the most probable/fluent amongst several alternatives, e.g. in a statistical machine translation system (Koehn, 2010). An n-gram language model (cf. (Manning and Schütze, 1999)) over sequences of words is defined by a Markov chain of order $(n-1)$, where the probability of the next word only depends on the $(n-1)$ previous words, and the probability of a sentence is defined as $P(w_1...w_k) = \prod_{i=1..k} P(w_i|w_{i-1}..w_{i-n+1})$. We add special symbols, BoS and EoS, to indicate sentence beginning and end. Then we generate sentences word by word, starting from a sequence of $(n-1)$ BoS-symbols, according to the probability distribution over the vocabulary. As soon as the EoS symbol is generated, we generate the next sentence. Probabilities are initialized by training on the respective corpus of real text (see above) from the relative counts, i.e. $P(w_i|w_{i-1}..w_{i-n+1}) = count(w_i..w_{i-n+1})/count(w_{i-1}..w_{i-n+1})$. In our study, we used n-gram models with $n \in \{2,3,4\}$ (in some contexts, we additionally consider $n = 1$ for completion).

Shortcomings of n-gram models are obvious: no long-range relations are modeled explicitly, thus n-gram models produce locally readable but semantically incoherent text. This study is, to our knowledge, the first attempt to quantify this phenomenon. Despite their simplicity, n-gram models still excel in NLP applications (cf. (Ramabhadran et al., 2012)).

In NLP applications, n-gram models are usually subject to smoothing and back-off techniques (cf. (Manning and Schütze, 1999)). Smoothing is necessary to account for unseen words, which is not an issue for generation. We only present results for language models without back-off in this work, although we did some experiments with texts generated from n-gram models with back-off estimated through deleted estimation. Note that we found no substantial differences to text generated without back-off.

3.2 Network Construction

Next, we describe the construction of a complex network from a text corpus of (real or generated) language. The nodes of the derived graphs correspond to the m most frequent words in the considered text. An edge from node A to B exists if the word corresponding to A *co-occurs*, i.e. occurs together in a well-defined context, with the word corresponding to B significantly often. Different kinds of co-occurrence contexts are considered, as well as significance thresholds and graph sizes m.

[3] see `http://corpora.informatik.uni-leipzig.de/`

Network size The number of nodes m in the graph, corresponding to the most frequent words in the considered text, was set to be $5,000$, as to match the commonly assumed size of the core vocabulary of a language (Dorogovtsev and Mendes, 2001). In preliminary experiments, we have verified that, qualitatively, our results are stable across vocabulary sizes between $1,000$ and $20,000$, as long as the *most frequent* words are considered.

Co-occurrence contexts We consider two different kinds of contexts: co-occurrence as immediate neighbors in a sequence, and co-occurrence within a sentence (sequences as limited by BoS and EoS). Thus, for each corpus of text, composed of sentences, we can compute the co-occurrence graph by connecting word nodes with edges, if words co-occur. Edges are directed in the case of neighbor-based co-occurrence, and undirected for the sentence-based case. It is known (Biemann et al., 2004) that sentence-based co-occurrences, besides capturing collocations, often reflect semantic relations and capture topical dependencies between words.

Significance threshold Since mere co-occurrence results in a large number of edges and very dense networks, we apply a significance test that measures the deviation of the actual co-occurrence frequency from the co-occurrence frequency that would have been observed if the two co-occurring words would be distributed independently. Here, we use the log-likelihood test (Dunning, 1993) to prune the network: We only draw edges between word nodes, if the words co-occur with a significance value above a certain threshold. For our experiments, we used a threshold of 10.83[4] . During preliminary experiments, we have found the reported results to be stable across a wide range of significance thresholds. See (Biemann and Quasthoff, 2009) for an analysis of global properties of significant co-occurrence graphs of natural language. The co-occurrence graph was computed using the TinyCC[5] corpus production engine (Quasthoff et al., 2006).

3.3 Network Analysis

Transitivity Let $G = (V, E)$ be an undirected graph. A *closed triangle* is a set of three nodes such that all three possible edges do exist. On the other hand, a *triple* is any set of three nodes and two edges (in other words, a chain of two edges). The *transitivity* of $T(G)$ of G is three times the total number of closed triangles divided by the total number of triples, as defined by (Newman et al., 2002). This can be calculated by iterating over every node v and counting the triangles and triples in which v is incident to two edges:

$$T(G) = \frac{\sum_{v \in V} \delta(v)}{\sum_{v \in V} \binom{k(v)}{2}} \tag{1}$$

with $\delta(v) = |\{u, v, w\} \in V, \{\{u, w\} \in E \text{ and } \{v, u\} \in E \text{ and } \{v, w\} \in E\}|$, and $k(v)$ the degree of v.

Motif analysis A k-node motif is a small connected graph of k nodes. An *occurrence* of a motif M in a network $G = (V, E)$ is a set $V' \subseteq V$ of nodes such that the subgraph of G induced by V' is isomorphic to M.[6] For a set of motifs, the *motif signature* is the vector of number of instances

[4] which corresponds to an error level of 0.1% of falsely rejecting the hypothesis that words co-occur independently

[5] available for download at `http://wortschatz.uni-leipzig.de/~cbiemann/software/TinyCC2.html`

[6] For a graph $G = (V, E)$ and a node set $V' \subseteq V$, the *subgraph* of G *induced* by V' is the unique graph (V', E'), where $E' \subseteq E$ is the set of all edges of E with both endnodes in V'. Note that this really means *all* edges. In fact, if E' is only required to contain *some* of the edges with both endnodes in V', (V', E') is usually called just a subgraph, not *the*

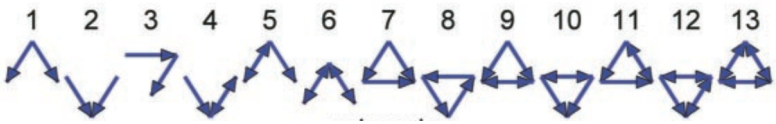

Figure 1: Directed motifs of size 3 as used in (Milo et al., 2004)

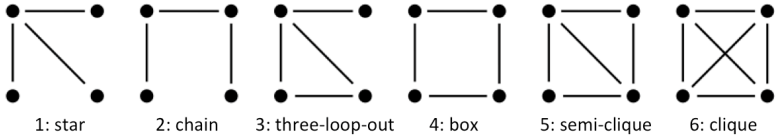

Figure 2: Undirected motifs of size 4 with names used throughout this paper

of each motif in G (typically, but not exclusively, all motifs in the set have the same number of nodes). Alternatively, the fraction of each motif to the total number of motifs is frequently considered, i.e. for m motifs M_1,...,M_m with counts $c(M_1)$,...,$c(M_m)$, the *motif signature* is the vector $(s(M_1), ..., s(M_m)$ with

$$s(M_i) = \frac{c(M_i)}{\sum_{j=1}^{m} c(M_j)}$$

To compare graphs of different sizes and edge counts, we generally present fractions instead of absolute counts. Our results are mainly presented in the form of xy-diagrams, mapping the motif to the corresponding frequency. Due to the high diversity in frequency, a logarithmic scale is used on the y axis. For presentational purposes and in accordance with the literature, we connect the dots in the plots, although they represent discrete values.

Throughout this paper, we consider two kinds of motifs, directed 3-node motifs (see Figure 1) and undirected 4-node motifs (see Figure 2). For directed and undirected graphs, respectively, these are the smallest meaningful motifs for our purposes. In fact, undirected 3-node motifs are triangles and triples and, hence, implicitly covered by our transitivity analysis. The motif counts were computed efficiently with the MotifAnalysis[7] software.

4 Conjectures

In the following, the term *(network) parameter* can refer to either the transitivity or a component of the motif signature.

Distinction We conjecture that a significant difference between the network parameters from real and generated text can be observed. Informally speaking, this conjecture is affirmatively substantiated within the scope of our computational studies if there is at least one network parameter such that for each language, the three values from generated text are on the same side of the value from natural text (either all smaller or all larger), and the distance between the former three values and the latter value is substantial.

induced subgraph.

[7] available for download at `https://github.com/stef-roos/MotifAnalysis`

Quantification We conjecture that the network parameters induce a reasonable quantitative measure how far a language model is from naturalness. In our studies, the investigated language models are the 2-gram, 3-gram, and 4-gram models. It is quite reasonable to say, the larger n is, the closer the n-gram model is to naturalness. Therefore, the conjecture is affirmatively substantiated within the scope of our computational studies, if we find at least one network parameter whose values for the n-gram models show a strictly monotonous convergence behavior towards the value for natural language, and this behavior is consistent throughout all languages.

Relation to semantics We conjecture that some of the motifs substantiate the first two conjectures, and that these motifs also allow a deeper insight into the semantic reasons for the observed differences. This conjecture is affirmatively substantiated within the scope of our computational studies, if we can identify semantic phenomena that (1) occur in natural text more often than in generated text and (2) significantly increase respectively decrease the number of occurrences of some motifs.

Relation to local syntax We conjecture that the motif profile of the neighbor-based graph does reflect local syntactic dependencies. A comparison of the motif profiles of the neighbor-based graphs should quantify the extent, to which language models capture local syntactic dependencies between adjacent words. Since n-gram models are trained on short word sequences, we conjecture that there is no difference between real and n-gram generated text with respect to local syntax for $n > 1$.

5 Results

With regard to sentence co-occurrence, the computational results in Sections 5.1 and 5.2 confirm that transitivity fulfills the first and second conjecture, and motif analysis fulfills all three conjectures. The local syntactic conjecture, based on neighborhood co-occurrence, is proven valid in our computational studies, as detailed in Section 5.3.

5.1 Distinction and Quantification

Transitivity Table 1 shows the transitivity values of the sentence-based co-occurrence networks for six languages, in each case for the 2-gram, 3-gram, and 4-gram models and for real natural language. The gap between natural text and any generated text is nowhere smaller than 15%. This substantiates the first conjecture. Evidently, the values for the n-gram models converge strictly monotonously towards the value of natural language in each case, which substantiates the second conjecture.

As an explanation, we attribute this to missing links in n-gram networks that result from the deficiency of such models to capture semantic coherence. The linguistic interpretation of transitivity is the following: if two words A and B co-occur significantly, and A occurs significantly with a third word C, what is the probability that B and C also co-occur significantly? Semantic cohesion (Halliday and Hasan, 1976) means that a text, thus a sentence, is about a certain topic, and there are several sentences that refer to the same topic in corpora. Topics manifest themselves in a certain set of words that will be used frequently together to describe this topic, which results in cliques in the co-occurrence network. While n-gram models capture topical relations between words if they co-occur within a short distance, they miss semantic relations between words that occur at long distances.

Clique motif The differences in the relative shares of the clique motif (#6 in Fig. 2) are even stronger. In fact, Table 2 shows the relative number of cliques in n-gram generated text normalized by the relative number of cliques in real text. The gap between natural text and any generated text is dramatic for $n = 2, 3$ and still always greater than 38% for $n = 4$. The Lithuanian language is the only exception to monotonous convergence. However, even for this language, the discrepancy is greater than 35% for $n = 3$. Except for the Lithuanian language, strictly monotonous convergence is evident, and most of the convergence steps are quite large. In summary, the first conjecture is completely substantiated by this particular motif as well, and the second conjecture is substantiated to a very large extent.

Language	**Real**		**2-gram**		**3-gram**		**4-gram**	
	$T(G)$	rel. $T(G)$	$T(G)$	rel. $T(G)$	$T(G)$	rel. $T(G)$	$T(G)$	rel. $T(G)$
English	0.1533	1.0	0.0729	0.4757	0.0886	0.5781	0.0937	0.6111
German	0.1255	1.0	0.0700	0.5573	0.0841	0.6701	0.1057	0.8420
French	0.1468	1.0	0.0652	0.4440	0.0773	0.5263	0.1047	0.7133
Indonesian	0.1789	1.0	0.0883	0.4936	0.1227	0.6858	0.1479	0.8263
Farsi	0.2143	1.0	0.0764	0.3565	0.1116	0.5207	0.1557	0.7265
Lithuanian	0.1615	1.0	0.0893	0.5530	0.1216	0.7529	0.1289	0.7981

Table 1: Transitivity $T(G)$ in absolute and relative terms to real language networks for six languages, comparing networks from real text with networks from n-gram generated text

To exemplify this, the closed neighborhood graph of "Monday" for its 20 most significant co-occurrences, which is the subgraph consisting of all edges involving "Monday" and the edges between all involved nodes, is depicted in Figure 3 for our real and n-gram networks of English. While collocations like "Monday evening" are present in all graphs, words like "Football" and "Saturday" do not get connected in the 3-gram graph: although they are generated significantly frequently with "Monday", this happens in different generated sentences, whereas they significantly co-occur in real language. Further, the density of these graphs is monotonically increasing with n and highest for real language.

5.2 Semantic Conjecture

Recall the concept of *functional blocks* from Section 2. Next we will show that, in our context here, the chain and the box motif are functional blocks in quite an analogous sense.

Figure 4 shows the motif profiles of networks on a log-scale. In Fig. 4 (left upper), we depict the

Language	**Real**		**2-gram**		**3-gram**		**4-gram**	
	abs	rel	abs	rel	abs	rel	abs	rel
English	0.1090	1.0	0.0339	0.3113	0.0387	0.3548	0.0407	0.3734
German	0.0735	1.0	0.0321	0.4364	0.0342	0.4659	0.0437	0.5944
French	0.1002	1.0	0.0192	0.1902	0.0284	0.2841	0.0484	0.4838
Indonesian	0.1706	1.0	0.0336	0.1971	0.0672	0.3939	0.1041	0.6104
Farsi	0.2668	1.0	0.0250	0.0937	0.0492	0.1843	0.1107	0.4149
Lithuanian	0.1755	1.0	0.0474	0.2699	0.1139	0.6487	0.1078	0.6143

Table 2: Percentage of clique motifs (#6) in absolute (abs) and relative (rel) terms to real language for six languages, comparing networks from real text with networks from n-gram generated text

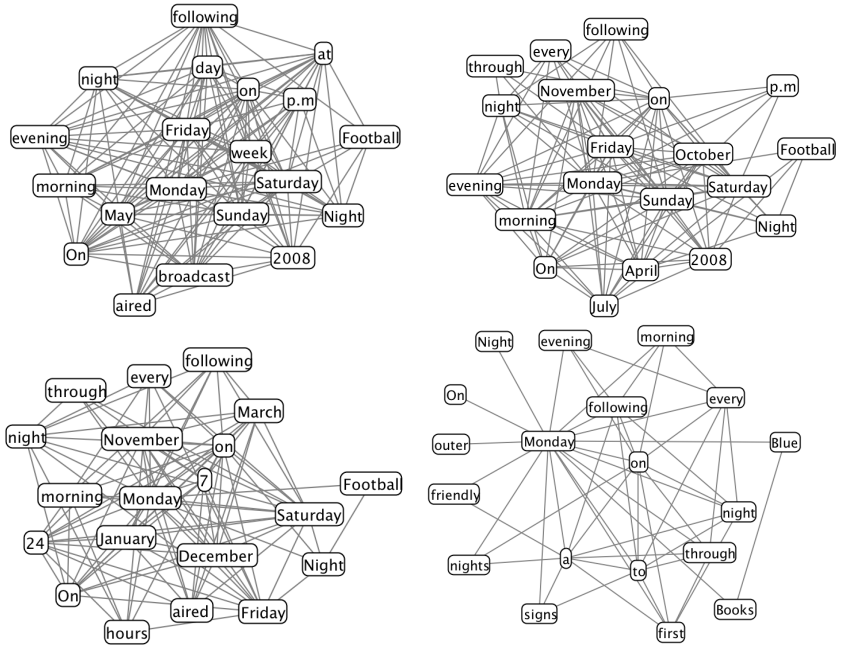

Figure 3: Neighborhood graphs of "Monday" in the English networks for real (upper left), 4-gram (upper right), 3-gram (lower left) and 2-gram (lower right) language, which exemplify the deficiency of n-gram models to capture long-range semantic relations

motif profiles for English networks of real and generated language for $n = 1, 2, 3, 4$; The other plots in Fig. 4 shows the profiles for all other languages for $n = 2, 3, 4$. It is clearly visible that real language networks exhibit fewer star (#1) motifs and a higher amount of all other motifs. Differences for the chain (#2) and the box (#4) motifs are especially pronounced. Examining instances of these motifs more closely, we are able to link these differences to properties of natural language semantics, which will be explained more thoroughly in the remainder.

Polysemy and chain motif *Semantic polysemy* refers to the phenomenon that a word, denoted as a string of characters, can have different denotations in different contexts, e.g. "board" as an assembly or a piece of wood. In real sentences, words are not co-occurring at random, but usually revolve around a certain topic. Thus, it is not likely to find the word "wood" in a sentence that talks about a "board of directors", and sentences about wooden planks usually do not contain the word "chairman".

In co-occurrence networks, polysemy leads to chains: ambiguous words connect words that are not connected to each other, and act as a bridge between different topical word clusters. In a chain of length four, one more word from a topical cluster is observed, which does not connect to the polysemous word since it seems that their occurrences are deemed rather independent

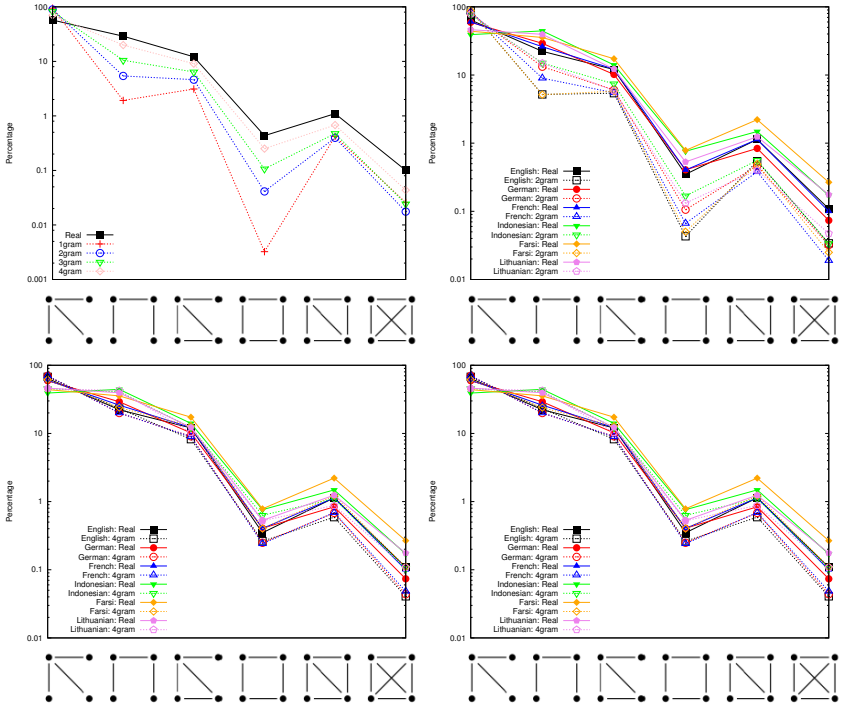

Figure 4: Motif profiles for real and generated text networks based on sentence co-occurrence. Upper Left: motif signature for English, comparing real language to $n = 1, 2, 3, 4$. Upper Right: Comparison Real vs. 2-gram for six languages. Bottom Left: Comparison Real vs. 3-gram for six languages. Bottom Right: Comparison Real vs. 4-gram for six languages. High congruence across languages of different language families demonstrates language independence of our analysis method

by the significance measure.

Enumerating the chain motif instances of the English real network, we exemplify this point with the following chains (the ambiguous word is emphasized in each line):

- total - km^2 - **square** - root
- Democrats - **Social** - Sciences - Arts
- Number - **One** - Formula - Championship
- Abraham - **Lincoln** - Nebraska - Iowa

N-gram models are oblivious to these sense distinctions. Thus, nothing prevents e.g. a 3-gram model from generating e.g. a sequence "Abraham Lincoln , Nebraska" with high likelihood,

confusing the two senses of "Lincoln" as a last name and a city. In the co-occurrence network, this can result in a connection between "Abraham" and "Nebraska", which decreases the chain motif count. The remaining chains of *n*-gram networks, on the other hand, consist mostly of highly frequent words that occur next to each other, e.g. "slowly started on finals", "personal taste good advice". These are also present in the real language network. We observe a much smaller number of chains formed of words of lower frequencies in *n*-gram generated text. Note that it is neither the case that all polysemous words cause chains, nor do all words in the central positions of a chain exhibit lexical ambiguity – differences in chain motif counts rather quantitatively measure the amount of such polysemy than qualify as an instrument to find single instances.

Hence, the lower amount of chain motifs can be explained by the creation of links that are not present in real language. On the first glance, this should lead to a higher clustering, contradictory to the results for motifs #3,5,6. However, as explained in Section 5.1, the clustering of *n*-grams is drastically lower than for real language. Although some of the 4-nodes sets that represent boxes or chains in real languages form (semi-)cliques in *n*-grams, instances of motif #3,5,6 in real languages are replaced by stars in *n*-gram graphs more frequently.

From these observations it becomes very clear that chain motif counts reflect polysemy. The lower *n* is chosen in the generating *n*-gram model, the smaller is the modelling context for ambiguities, resulting in lower chain motif counts.

Synonymy and box motif *Synonymy* means that different words refer to the same concept. Two words are synonyms if they can be used interchangeably without changing the meaning, but there are also rather syntactic variants of words that refer to the same concept, such as nominalizations of adjectives or verb forms of different inflections.

In natural language, the principle of *parsimony* leads to the effect that the same concept is rarely referred to several times in the same sentence. In fact, synonyms usually do not co-occur, but they share a large number of significant co-occurrences – an observation that leads to the operationalization of the distributional hypothesis (Miller and Charles, 1991). When two such concepts are mentioned together frequently since they belong to the same topic, this leads to box motifs, as the following examples from the English real language network illustrate:

- - Ancient - Greek - ancient - Greece -
- - winning - award - won - price -
- - Ph.D - his - doctorate - University -
- - said - interview - stated - " -
- - wrote - articles - published - poems -

We observe different kinds of word pairs for the same concept: synonyms like (award, price), same word stem within or across word classes like (winning, won) or (Greek, Greece), and artifacts of punctuation or spelling (ancient, Ancient) or (interview, "), Thus, box motifs capture a very loose notion of synonymy: "interview" and the double quote " e.g. both refer to a (indirect or direct) speech act.

Again, *n*-gram models are not aware of concepts and references to them, so there is no mechanism that prevents the *n*-gram model from generating sentences that refer to the same concept several times or even use the same word repeatedly. This possibly results in a connection

between those pairs, reducing the box motif count. Box motifs that can be found in both real and n-gram language are again resulting from local sequences of highly frequent words that are possibly circular, not necessarily from the same contexts. Examples include "desktop cover art background", "hall nearby on church" and "these will ask why".

These observations lead to the conclusion that synonymy of natural language leads to box motifs in sentence-based co-occurrence networks, and the difference in the box motif count quantifies the amount of capability of the language model to inhibit the generation of words that refer to concepts which already have been mentioned. N-gram models have this capability only for a very limited context, which again increases with higher n.

5.3 Local Syntactic Conjecture

To assess whether n-gram models really grasp local syntactic dependencies, we examine the motif profile of directed motifs of size 3 on the neighbor-based co-occurrence graph as described in Section 3. In this, we follow Milo et al. (2004), who unfortunately do not specify their procedure of graph construction from language in detail. Figure 5 shows the directed motif profiles of real text and 2-gram-generated text for four languages, and for English for $n = 2, 3, 4$. As Milo et al. (2004) observe, different languages have a very similar motif profile, and the corresponding graphs belong to the same superfamily of networks, i.e. the mostly bipartite networks. We also observe that there is no systematic difference between the neighbor-based networks of real language and generated language, even for $n = 2$. This substantiates our fourth conjecture: n-gram models do in fact grasp local syntactic dependencies very well.

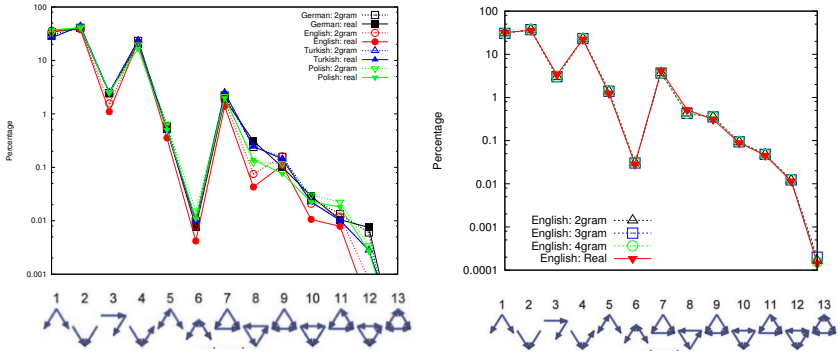

Figure 5: Motif profile of the directed neighbor-based co-occurrence graphs for real and 2-gram generated language (left), respectively for English with n-gram generated text for $n = 2,3,4$ (right), showing a high inter-language agreement and the inability of this measure to distinguish real from generated language

This renders the result of Milo et al. (2004) as not being very influential for natural language research, since the neighbor-based motif profile of real language can be generated by a highly deficient 2-gram language model.

6 Conclusion and Outlook

The methods presented in this paper open new ways to the assessment of the quality of language models, as reflected in the first and second conjecture. We showed that both a global graph metric, transitivity, and an intermediate metric, motif signatures, quantify the difference between natural and generated language. With our third conjecture, we go even one step further by presenting a way to relate semantics to network structure. Looking at the motif signature of real and generated language, we can identify differences due to polysemy and synonymy, which are not adequately modeled by *n*-grams.

Our analysis builds on the fact that generation with language models is not tied to any target application, and generative language models that do not have mechanisms to ensure cohesion will fail to show the same patterns as real language, especially regarding semantic properties such as synonymy and polysemy. In applications like Machine Translation, where language models are used to rank alternatives rather than free generation and are thus bound to the cohesive structure of the source language text, the shortcomings discussed in this paper do not necessarily impede the performance on the task. In fact, preliminary experiments involving comparisons of real translations with automatic translations of the same text did not result in motif profile differences. However, as e.g. (Tan et al., 2012) point out, there is a need for more coherent language models even in these applications: e.g. speech-to-text in noisy environments might greatly benefit from better language models.

Our computational studies with regard to co-occurrence graphs based on sentences and neighboring words indicate that language models based on *n*-grams reflect local syntax well, but fail to model semantic cohesion and topicality. Further, these language models do not have means of regulation for referring to the same concept several times. While these results confirm the common intuition of *n*-grams, we present the first study to actually quantify this deficiency.

The presented series of experiments are but a first step towards a more systematic analysis of the relation between the global characteristics of language and the structure of co-occurrence networks. Varying the notion of co-occurrence, for example, to involve positional offsets, could possibly unveil grammatical differences for a quantitative typology of languages. Further work should include more sophisticated language models such as the syntactic topic model (Boyd-Graber and Blei, 2008), which explicitly models topicality. The restriction to a subset of word classes, for example, nouns or verbs, or to class-based *n*-gram models (Brown et al., 1992) may also be interesting.

All of these ideas still address the *analysis* of language models. As mentioned in Section 2, the concept of motifs has recently been used for a constructive purpose. We anticipate that this change of perspective is also promising in the realm of language networks and may well guide the design of new, semantically more adequate, language models.

References

Aggarwal, C. C. (2011). *Social Network Data Analytics*. Kluwer.

Aggarwal, C. C. and Wang, H. (2010). Managing and mining graph data. *Database*, 40:487–513.

Amancio, D., Oliveira Jr., O., and da Costa. L.F. (2012). Using complex networks to quantify consistency in the use of words. *Journal of Statistical Mechanics: Theory and Experiment*, 2012(01):P01004.

Biemann, C. (2007). A random text model for the generation of statistical language invariants. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 105–112, Rochester, New York.

Biemann, C., Bordag, S., and Quasthoff, U. (2004). Automatic acquisition of paradigmatic relations using iterated co-occurrences. In *Proceedings of LREC 2004*, Lisbon, Portugal.

Biemann, C., Heyer, G., Quasthoff, U., and Richter, M. (2007). The Leipzig Corpora Collection - monolingual corpora of standard size. In *Proceedings of Corpus Linguistic 2007*, Birmingham, UK.

Biemann, C. and Quasthoff, U. (2009). Networks generated from natural language text. In Ganguly, N., Deutsch, A., and Mukherjee, A., editors, *Dynamics On and Of Complex Networks*, Modeling and Simulation in Science, Engineering and Technology, pages 167–185. Birkhäuser Boston.

Boyd-Graber, J. and Blei, D. M. (2008). Syntactic topic models. In *Neural Information Processing Systems*.

Brown, P. F., Pietra, V. J. D., deSouza, P. V., Lai, J. C., and Mercer, R. L. (1992). Class-Based n-gram Models of Natural Language. *Computational Linguistics*, 18(4):467–479.

Chang, J., Boyd-Graber, J., Wang, C., Gerrish, S., and Blei, D. M. (2009). Reading tea leaves: How humans interpret topic models. In *Neural Information Processing Systems*.

Cook, D. J. and Holder, L. B. (2006). *MIning Graph Data*. Wiley.

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*.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

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.

Fortunato, S. (2010). Community detection in graphs. *Physics Reports*, 486:75–174.

Halliday, M. A. K. and Hasan, R. (1976). *Cohesion in English*, volume 1 of *English Language Series*. Longman.

Koehn, P. (2010). *Statistical Machine Translation*. Cambridge University Press, New York, NY, USA, 1st edition.

Köhler, R. and Naumann, S. (2010). A syntagmatic approach to automatic text classification. statistical properties of F- and L-motifs as text characteristics. In Grzybek, P., Kelih, E., and Mačutek, J., editors, *Text and Language*, pages 81–89. Praesens Verlag, Vienna.

Krumov, L., Andreeva, A., and Strufe, T. (2010a). Resilient peer-to-peer live-streaming using motifs. In *11th IEEE World of Wireless, Mobile and Multimedia Networks (WoWMoM)*, pages 1–8.

Krumov, L., Fretter, C., Müller-Hannemann, M., Weihe, K., and Hütt, M.-T. (2011). Motifs in co-authorship networks and their relation to the impact of scientific publications. *European Physical Journal B*, 84(4):535–540.

Krumov, L., Schweizer, I., Bradler, D., and Strufe, T. (2010b). Leveraging network motifs for the adaptation of structured peer-to-peer-networks. In *GLOBECOM*, pages 1–5.

Manning, C. D. and Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, USA.

Masucci, A. P. and Rodgers, G. J. (2006). Network properties of written human language. *Phys. Rev. E*, 74:026102.

Miller, G. A. and Charles, W. G. (1991). Contextual Correlates of Semantic Similarity. *Language and Cognitive Processes*, 6(1):1–28.

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M., and Alon, U. (2004). Superfamilies of evolved and designed networks. *Science*, 303(5663):1538–1542.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., and Alon, U. (2002). Network Motifs: Simple Building Blocks of Complex Networks. *Science*, 298(5594):824–827.

Newman, M. E. J., Watts, D. J., and Strogatz, S. H. (2002). Random graph models of social networks. *Proc. Natl. Acad. Sci. USA*, 99(suppl. 1):2566–2572.

Pardo, T. A. S., Antiqueira, L., Nunes, M. G. V., Oliveira Jr., O. N., and da F. Costa, L. (2006). Using complex networks for language processing: The case of summary evaluation. In *Proceedings of the International Conference on Communications, Circuits and Systems (ICCCAS'06) - Special Session on Complex Networks*, pages 2678–2682, Gui Lin, China. UESTC Press.

Quasthoff, U., Richter, M., and Biemann, C. (2006). Corpus portal for search in monolingual corpora. In *Proceedings of the fifth international conference on Language Resources and Evaluation, LREC*, pages 1799–1802, Genoa, Italy.

Ramabhadran, B., Khudanpur, S., and Arisoy, E., editors (2012). *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*. Montréal, Canada.

Schreiber, F. and Schwöbbermeyer, H. (2010). *Statistical and Evolutionary Analysis of Biological Network Data*, chapter Motifs in biological networks, pages 45–64. Imperial College Press/World Scientific.

Shen-Orr, S., Milo, R., Mangan, S., and Alon, U. (2002). Network motifs in the transcriptional regulation network of Escherichia coli. *Nature genetics*, 31(1):64–68.

Steyvers, M. and Tenenbaum, J. B. (2005). The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29(1):41–78.

Tan, M., Zhou, W., Zheng, L., and Wang, S. (2012). A scalable distributed syntactic, semantic, and lexical language model. *Computational Linguistics*, 38(3):631–671.