Syntactic Complex Networks and Their Applications

Radek Čech, Ján Mačutek, and Haitao Liu

Abstract. We present a review of the development and the state of the art of syntactic complex network analysis. Some characteristics of such networks and problems connected with their construction are mentioned. Relations between global network indicators and specific language properties are discussed. Applications of syntactic networks (language acquisition, language typology) are described.

1 Introduction

Syntax is considered to be a key component of human language. Its properties, origin, cognitive status etc. have been discussed intensively for decades by researchers from different branches of science and it has caused tough controversies among them. Despite a huge number of arguments it has been difficult, or still impossible, to find a generally acceptable criterion or method which can help to solve fundamental problems considering syntax, especially its origin.

A theory of complex networks emerged at the turn of the millennium (Barabási and Albert 1999; Barabási 2002) and its rapid and successful development appeared to be a useful tool for an explanation of system properties in many branches of science. It is not surprising that a use of this theory for an explanation of some fundamentals of syntax was tempting. And indeed, just after the beginning of the endeavor

Radek Čech
Department of Czech Language, Faculty of Arts,
University of Ostrava, Reální 5, Ostrava, CZ-70103, Czech Republic
e-mail: cechradek@gmail.com

Ján Mačutek
Department of Applied Mathematics and Statistics,
Comenius University, Mlynská dolina, Bratislava, SK-84248, Slovakia
e-mail: jmacutek@yahoo.com

Haitao Liu
Department of Linguistics, Zhejiang University, Hangzhou, CN-310058, China
e-mail: lhtzju@gmail.com

A. Mehler et al. (eds.), *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*,
Understanding Complex Systems, DOI: 10.1007/978-3-662-47238-5_8

to study syntax properties by methods of complex network analysis Ferrer i Cancho et al. (2005) brought promising explanation considering syntax origin. Specifically, they introduced complex network based model of language which takes into account (1) relationships between words and objects, (2) relationships among words related to the same object, and (3) Zipf's law; a property of the model (namely, connectedness) represents a precondition for syntax evolving, according to the authors (for more details see Sect. 3). Afterwards, Solé (2005) presented the approach of Ferrer i Cancho et al. in slightly changed form in popular science article in *Nature*, one of the most prestigious scientific journals. Especially the article by Solé represents some kind of "great expectations" (cf. its title: "Syntax for free?") which could bring the use of complex network analysis in language analysis.

After almost a decade, it seems reasonable to critically review the development of syntactic complex network analysis and to try to answer the following questions: What are the results of the application of complex network theory to syntax analysis? Has the application met the expectations? What kind of explanation has complex network analysis of syntax brought? Which new problems have emerged? What is the actual scope of syntax network analysis now? What are the perspectives? In this paper, we attempt to track main aspects of the development of syntactic network analysis and to summarize the results of this scientific endeavor. Our article follows a review presented by Mehler (2007). It can be considered as a complement to a more general overview on network analysis (Baronchelli et al. 2013).

The review is organized as follows: first, main characteristics of syntactic networks are introduced in Sect. 2; then, an early development of syntactic network analysis is presented (Sect. 3) and an impact of syntax on network properties is discussed (Sect. 4); further, important problems related to data preprocessing (e.g., coordination and lemmatization) are discussed in Sect. 5; next, Sect. 6 is dedicated to applications of syntactic networks in language typology and language acquisition; and the article is finalized by Conclusions (Sect. 7).

2 Basic Characteristics of Syntactic Networks

A network is a set of nodes and links. Nodes represent some entities while links represent relationships among nodes. As for syntactic network, nodes usually denote either particular wordforms (e.g., *sing, sings, sang, sung*) or lemmas (in this case all word forms are represented by the canonical form, e.g. *sing, sings, sang, sung* are represented by the single lemma *sing*) (cf. Čech and Mačutek 2009).

Links denote the so-called dependencies, i.e., syntactic relationships between pairs of words. For instance, there are four syntactic relationships in the sentence

(1) *Peter gave Mary the pen,*

specifically, between pairs of words *Peter – gave, gave – Mary, gave – pen,* and *the – pen*. The notion of dependency expresses the fact that

> one wordform must depend on another for its linear position and its grammatical form,

cf. Mel'čuk (2003, p. 188). This approach to syntax is called the dependency grammar formalism Mel'čuk (2003) and Hudson (2007). It is the only syntactic formalism which has been so far exploited for syntactic network analysis. Other ones (e.g., phrase structure or construction grammar) cannot be excluded, in principle; however, due to the lack of linguistic reasoning or interpretation they have not been used, to our knowledge.

A sentence structure reflecting syntactic dependencies between pairs of words can be described by a tree graph (all nodes in the tree graph must be connected and no cycles are allowed in this type of graph, which is in accordance with the syntactic dependency formalism), see Fig. 1.

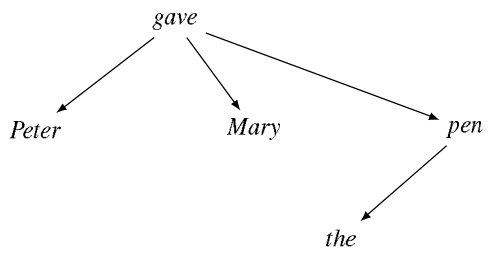

Fig. 1 The structure of sentence (1). Links between words represent the syntactic dependency relationships, arrows express the direction of the dependency. However, there is no general agreement among linguists regarding the direction; thus, one can find dependency formalisms using opposite direction (from modifier to head).

A syntactic dependency complex network is constructed by accumulating sentence structures and, thus, the network is an emergent property of these structures. Specifically, the network contains all words which occur in a text corpus and words are linked if the words appear syntactically linked at least once in a sentence of the corpus (Ferrer i Cancho et al. 2004; Ferrer i Cancho 2005a). Figure 2 shows an example of a small syntactic network containing 50 lemmas.

3 Early Development of Syntactic Complex Network Analysis

The early development of syntactic complex network analysis can be characterized as an endeavor to show that (1) syntactic properties of human language follow some universal patterns that are not observable by an analysis of particular sentences (Ferrer i Cancho et al. 2004; Ferrer i Cancho 2005a; Solé et al. 2010) and (2) that some language properties which are modelled well by complex network represent a necessary precondition for development of full syntax, cf. Ferrer i Cancho et al. (2005) and Solé (2005). As for the former, the patterns emerge only if the language is analyzed from a global point of view as a complex system containing huge number of language units and interrelations among them. Of course, it is nothing new to see the language as a complex system – F. de Saussure (de Saussure 1979) was probably

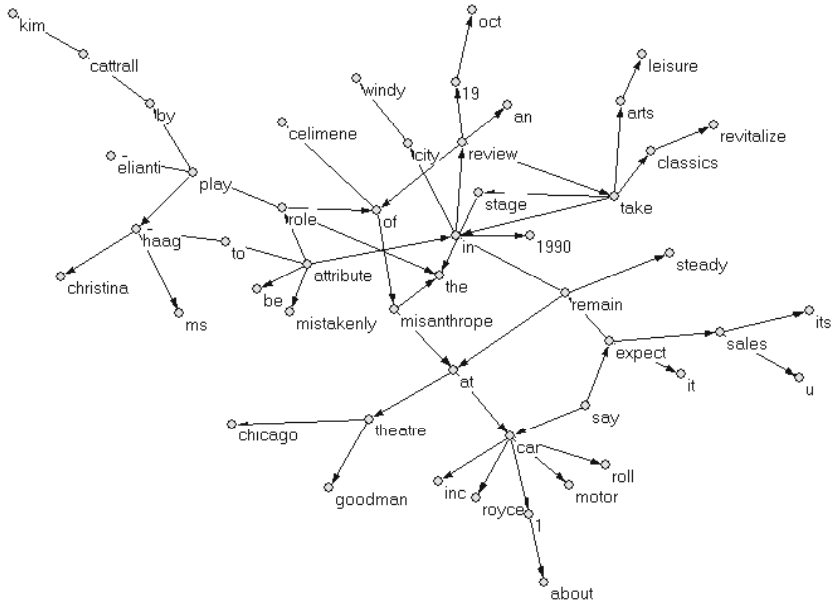

Fig. 2 The network containing the first 50 lemmas from the Penn TreeBank (Surdeanu et al. 2008).

the first one to stress this aspect. What is new, indeed, is the global point of view which is represented by the complex network theory (cf. Newman 2011). In other words, applying network models to language reveals that language networks share the same statistical characteristics, for instance, small world structure, degree distribution, betweenness centrality etc. Moreover, these statistical characteristics appear in networks based on different language units (phonemes, syllables, words) and different kind of relationships among the units (co-occurrence, collocation, syntactic dependency, semantic relationships). Therefore, these statistical characteristics are considered to be universal patterns which could be candidates for new linguistics universals (Ferrer i Cancho et al. 2004; Ferrer i Cancho 2005a; Solé et al. 2010; Choudhury and Mukherjee 2009). Focusing on syntax, this line of thinking brings an important implication. If syntactic networks display the same global characteristics as other linguistic networks, a uniqueness or specificity of syntax is cast into doubt (cf. Nowak et al. 2000, Hauser et al. 2002, and Fitch and Hauser 2004). Consequently, syntax should be ruled by the same (or similar) general principles as other language properties. Further, these principles should be rooted in a general cognitive faculty of human mind (Ferrer i Cancho 2005a, p. 68):

> the structure of global syntactic dependency networks mirrors the structure of brain. It is obvious that the brain is made of millions of neurons connected through synapses but the similarities go beyond mere physical resemblance. The activation of different brain areas shows the small-world phenomenon and a power degree distribution (Eguíluz et al.

> 2005; Grinstein and Linsker 2005). (...) While no one has ever found a rewriting rule in the brain of a human, the web organization of the brain at many levels, with linguistic networks on top, cannot be denied.

Naturally, this kind of findings opens questions on the origin of syntax. Complex network analyses of many different systems (World Wide Web, social networks, biochemical networks, ecological networks, neural networks etc.) reveal that a complex structure of these systems is a result of self-organization which is based on relatively simple principles, e.g., continuous growing of the system and preferential attachment (Barabási and Albert 1999). Specifically, the outcome of self-organized phenomena is a scale-free power-law distribution of degrees (i.e., numbers of links connected to each node). Analogously, the origin of syntax can be considered to be an outcome of the same principles.

As we noted in Sect. 1, Ferrer i Cancho et al. (2005) used the complex network as a model of a certain combinatorial property of words, namely, their connectedness, which is considered to be a necessary precondition for full syntax. According to Ferrer i Cancho et al. (2005), the connectedness arises naturally from the Zipf's law, independently of details of the linguistics setting. Even though it is stated that the model does not correspond to the complexity of human language, only a "small step" from the model to full syntax and full human language is supposed, cf. Ferrer i Cancho et al. (2005, p. 562)

> For various reasons, our grammar is not a grammar in the strict sense, but rather a protogrammar, from which full human language can *easily*[1] evolve (...) although our model is obviously much simpler than present-day languages, it provides a basis for the astronomically large number of sentences that human speakers can produce and process.

This aspect of the approach of Ferrer i Cancho et al. (2005) is strongly emphasized by Solé who links the property of the model to emerging of *syntactic rules*, cf. Solé (2005)

> ... sometimes illogical and quirky appearance of syntactic rules might be nothing but a by-product of scale-free network architecture. (...) ... Zipf's law could have been a precondition for syntax and symbolic communication.

According to us, a difference between 1) a conception of the complex network model as a necessary *precondition* for syntax, which is the most important outcome of the study (Ferrer i Cancho et al. 2005), and 2) a direct relationship between complex network properties and full syntax claimed by Solé (2005), is fundamental. In other words, the impact of complex network properties to syntax evolving radically differs, if one (1) consider these properties of complex network to be necessary, but not sufficient condition for full syntax, or (2) consider them to be both necessary and sufficient condition.

Regardless of the scope of the interpretation, the reference to the Zipf's law represents one of the most important aspects of the attempt to explain the origin of syntax by network analysis. The Zipf's law states that the relationship between the

[1] Emphasized by the authors of this review.

frequency of a word in a text and its rank is approximately linear when plotted on the double logarithmic scale, which means that the word frequency distribution is a power law. In Ferrer i Cancho et al. (2004) it is shown that the relationship between word frequency and word degree in all observed syntactic networks is approximately linear and, more interestingly, both distributions, those of word frequencies as well as of word degrees, have approximately the same exponent. Based on these observations, Ferrer i Cancho concludes that word degree distribution could be a consequence of word frequency and asks a fundamental question (Ferrer i Cancho 2005a, p. 66):

> If word degree is a consequence of the Zipf's law for word frequencies, a pressing question is: what is the origin of that law?

Since Zipf (1949) it has been known that power law distribution of word frequencies could be explained by general communication principles, such as the principle of least effort (Zipf 1949) or communication requirements (Köhler 1986; Köhler 2005). Roughly speaking, these principles are based on the idea that a general communication strategy is to minimize the cost of word usage, on the side of speaker, and the cost of word perception, on the side of hearer. These competitive strategies lead to an equilibrium that has the main impact on the form of the language system in general. The Zipf's law represents one kind of the equilibrium, cf. Ferrer i Cancho (2005b) and Ferrer i Cancho and Solé (2003).

To sum up, the early development of syntactic network analysis reveals a relationship between frequency of word and its syntactical properties expressed by the degree distribution. However, the mere relationship between these two phenomena does not explain the emerging of syntactic rules, i.e., if there is a linear correlation between the word frequency and word degree in a syntactic complex network, another question appears: what is the role of syntax in syntactic networks, if the word degree can be seen as a consequence of word frequency? Attempts to answer this question are presented in the next section.

4 Role of Syntax in Syntactic Dependency Complex Networks

Traditionally, syntax is considered to be, roughly speaking, a set of rules which govern the behavior of words in a sentence. We emphasize that the rules should be understood as probabilistic, not deterministic. The aim of syntax as a science is to describe a sentence structure, the character of rules and, sometimes, to explain why both, the structure and rules, are as they are. Regardless of different approaches to syntax, there is a general agreement among linguists about the hierarchical sentence structure – it means that "behind" the linearity of a sentence there are grammatical relationships between words. This fact is clearly illustrated on the example of highly inflected languages which have a flexible word order, cf. six grammatically well-formed Czech sentences expressing the English sentence *Peter beats Paul*:

(2) *Petr bije Pavla* [$\text{Petr}_{\text{noun-subject-nominative}}$ – $\text{bije}_{\text{verb}}$ – $\text{Pavla}_{\text{noun-object-accusative}}$]

(3) *Pavla bije Petr* [$\text{Pavla}_{\text{noun-object-accusative}}$ – $\text{bije}_{\text{verb}}$ – $\text{Petr}_{\text{noun-subject-nominative}}$]

(4) *Petr Pavla bije* [$\text{Petr}_{\text{noun-subject-nominative}}$ – $\text{Pavla}_{\text{noun-object-accusative}}$ – $\text{bije}_{\text{verb}}$]

(5) *Pavla Petr bije* [$\text{Pavla}_{\text{noun-object-accusative}}$ – $\text{Petr}_{\text{noun-subject-nominative}}$ – $\text{bije}_{\text{verb}}$]

(6) *Bije Petr Pavla* [$\text{Bije}_{\text{verb}}$ – $\text{Petr}_{\text{noun-subject-nominative}}$ – $\text{Pavla}_{\text{noun-object-accusative}}$]

(7) *Bije Pavla Petr* [$\text{Bije}_{\text{verb}}$ – $\text{Pavla}_{\text{noun-object-accusative}}$ – $\text{Petr}_{\text{noun-subject-nominative}}$]

In all these instances, the object of the sentence *Pavla (Paul)* is determined by its accusative form which is a result of the syntactic rule; the object of the sentence is not determined by the word order. Similarly, syntactic rules also determine the dependency of *Pavla (Paul)* on the verb. Consequently, according to the dependency grammar formalism, the structure of sentences (2)–(7), i.e. the relationships between each pair of words, is expressed by the graph in Fig. 3, in accordance with dependency grammar formalism.

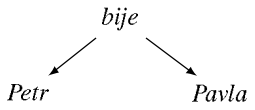

Fig. 3 The hierarchical structure of sentences (2)–(7).

but not by graphs in Fig. 4 (of course, there are more possibilities of non-correct graphs).

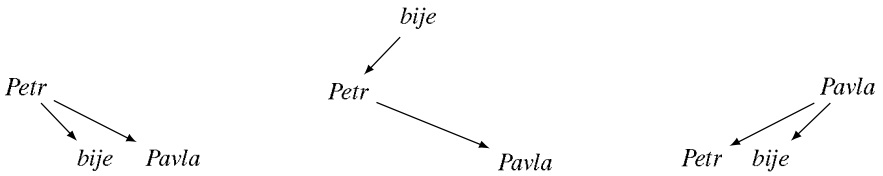

Fig. 4 Examples of non-correct hierarchic representations (i.e. representations violating syntactic rules) of the structure of sentence *Petr bije Pavla.*

Remember, a syntactic dependency complex network is constructed by accumulating sentence structures (cf. Sect. 2) and it was assumed (Solé 2005) that

> sometimes illogical and quirky appearance of syntactic rules might be nothing but a by-product of scale-free network architecture.

If Solé's statements were true, one should find radical differences between a network based on "proper" syntactic rules, on the one hand, and a network based on relationships between words which are not ruled by syntax (e.g, if relationships are generated randomly), on the other.

The idea to investigate the role of syntax in syntactic dependency networks was introduced by Liu and Hu (2008). According to them, the fact that language networks built on different principles are small-world and scale-free just like other real networks opens a problem with respect to a relation between these global network indicators and specific language properties. They focused on syntax and asked the following questions:

> if all language networks have properties such as small-world and scale-free: Could they be viewed as a general feature of a language network? What role does syntax play in such a syntactic (language) network? If dependencies are built by randomly linking words in the same sentence, would the network still follow the properties similar to the syntactic one? Can the local (micro) syntactic analysis in a sentence be reflected in the global (macro) properties of a language network?

As an attempt to answer these questions, Liu and Hu (2008) com-pared properties of three networks: 1) a syntactic network generated from a treebank based on the dependency grammar formalism; 2) a network based on randomly generated relationships between each pair of words within a sentence (sentences were taken from the treebank): from each sentence one word was randomly selected as a root (in accordance with the dependency grammar formalism) and for each of the remaining words a word of the same sentence was randomly selected as a governor; 3) a network based on randomly generated relationships between each pair of words from a sentence which, however, respect the principle of continuity (also called projectivity) of a tree representing a sentence structure; this condition is added because discontinuous (or non-projective) syntactic trees are rather exceptional in the natural language (Ferrer i Cancho 2006), so, this network should express more similar properties to a syntactic network than the totaly random network.

A comparison of the networks was focused on five global network characteristics, namely, the average path length, which is defined as the shortest distance between a pair of vertices; the diameter, which is defined as the longest shortest path in a network; the average degree, i.e. the average number of links of a node; the clustering coefficient, i.e. the probability that two nodes which are neighbors of a given node are neighbors themselves; and the degree distribution. The analysis revealed that all networks displayed very similar global network characteristics and all of them were scale-free and small-world. Consequently, the fact that a network is small-world or scale-free cannot alone explain the role of syntax in the network. Obviously, this result seriously undermines the Solé's statement which considers the emergence of syntactic rules to be a by-product of the scale-free network architecture (Solé 2005, similarly also Solé et al. 2010). The properties of small-worldness and scale-freeness in random syntactic networks can be explained by frequency characteristics of words (the Zipf's law), not by syntax. A mere presence of these statistical indicators in a syntactic network is not a consequence of syntactic properties of language and, therefore, it cannot be considered to be a candidate for a *syntactic* language universal. Results of Liu and Hu (2008) reveal that network properties can be considered, at most, to be just the necessary *precondition* for the emergence of syntax (in the sense of a set of rules). Consequently, it is not

appropriate to suppose that full syntax can "easily evolve" as soon as the precondition is satisfied (Ferrer i Cancho et al. 2005; Solé 2005).

Interestingly, a similar attempt to build a random network (i.e., without predefined syntactic rules) was presented by Corominas-Murtra et al.[2] (2009, 2010). The creation of a random network was based on two principles: 1) the frequency of words followed the Zipf's law, and 2) the length of a sentence corresponded to real language data. The relationships between words were generated randomly. The resulting random network had similar typical global network properties, such as the clustering coefficient, the degree distribution or the average path lengths, like real language syntactic networks. This fact was interpreted as a proof of emerging syntax properties in a language evolution "for free" and led the authors to more general arguments on the language evolution: specifically, its non-adaptive nature and innateness of syntax.

However, there are several problems in the analysis. First, to interpret the emergence of network properties as a result of innateness of syntax is not appropriate for the following reasons: 1) there is not a clear and empirically proved connection between the emergence of network properties and innateness; 2) both small number of children and only one language (English) is analyzed; the results from so tiny a sample cannot be interpreted in such a general way, for obvious reasons.

Further, the authors do not take into account a potential impact of the length of the samples on network properties; in other words, the emergence of observed network characteristics could be a side-effect of the fact that as time goes on, children speak more and the length of their transcripts increases. Next, the model is based on the assumption that the exponent of the Zipf's law is equal one and that it remains constant as a child evolves. However, this assumption is not correct (cf. Baixeries et al. 2013); the change of the exponent can lead to different network characteristics. Yet another problem is the absence of statistical testing.

Finally, in the light of the study Liu and Hu (2008), conclusions presented in Corominas-Murtra et al. (2009) and Corominas-Murtra et al. (2010) seem to be not acceptable because the global network properties emerge even if syntactic rules are "deleted" by randomness involved in the process of the network creation. In other words, the emergence of these global network properties is a result of frequency characteristics of language, and does not have to do anything with syntax.

On the other hand, Liu and Hu (2008) obtained results which show that there are some differences between syntactic and random networks (e.g., a syntactic network has a lower average degree and clustering coefficient). Even though these differences are too small for classifying syntactic and random networks as different types of complex networks, their existence indicates some influence of syntax on these indicators. This means that a use of complex networks for a syntactic analysis has to be focused on more fine-grained network properties in order to be useful. Moreover, it should not be forgotten that a syntactic network analysis (as well as any language network analysis) should be based linguistically which means that it is necessary to

[2] Both papers, i.e., Corominas-Murtra et al. (2009) and Corominas-Murtra et al. (2010), contain the same data and introduce the same procedure; actually, in Corominas-Murtra et al. (2009) one can find a more thorough discussion and a more general explanation.

explain observed network properties with regard to certain language characteristics or (in the better case) a linguistic theory. In other words, to say that, for instance, "syntax is small-world" without linguistic grounds can be misleading, as the analysis by Liu and Hu (2008) uncovered. The same conclusion was presented in the analysis by Liu et al. (2010) focused on an impact of different annotation schemes used for capturing syntactic coordination on global network properties.

In accordance with the approach from Liu and Hu (2008) and Liu (2008), which emphasizes the need for a linguistic explanation of network properties, the analysis of the role of syntax in syntactic complex networks in Čech et al. (2011) focused on verb characteristics. Contradicting previous studies that revealed the linear relationship between frequency and degree in a syntactic network in general, it was hypothesized that verbs should play a central role (expressed by the degree of node) in a syntactic network not due to their frequency but because of their syntactic property called valency (Allerton 2005; Liu 2008).

The paper Čech et al. (2011) starts with the well-known idea on the relationship between the shape of a network (its topological properties) and its functionality (Caldarelli 2007) and, further, it was deduced how a function of a verb in a sentence should influence network characteristics. Specifically, verb valency determines, besides other things, the number of words obligatory dependent on the verb in a sentence and it plays a decisive role in the sentence structure. Moreover, a verb is present at least once in each sentence due to its syntactic function to be the predicate of the sentence – this fact guarantees a relatively high frequency of verbs in any language sample. However, it should be emphasized that verbs are not the most frequent part of speech (nouns have the highest frequency, at least in languages used for the analysis). Based on these properties of verbs (i.e., valency and a relative high frequency caused by its function in a sentence), it was predicted (Čech et al. 2011, p. 3616) that

> verbs should play an important role in the network expressing syntactic relationships in the language. In other words, it is predicted that verbs will occur among the most important elements of the network.

The importance of an element was determined by its degree.

Six languages (Catalan, Czech, Dutch, Hungarian, Italian, and Portuguese) were used for testing the hypothesis. The results reveal that proportions of verbs (with regard to other parts of speech) in histogram bins of the ranked distributions of degrees tend to decrease while the proportions of verbs (again with regard to other parts of speech) in histogram bins of the rank-frequency distribution of lemmas are more or less constant and clearly tend to attain lower values than verb proportions in the case of degrees. Differences between rank-degree and rank-frequency distributions are statistically significant. Thus, the results do not falsify the hypothesis and allow to state that the topology of a syntactic dependency network is significantly affected by syntax of the language, at least in the case of verb valency.

To sum up, studies focused on the analysis of the role of syntax show that it is not acceptable to interpret and explain syntax properties of human language by global network properties, such as the small-worldness and scale-freeness. Further,

a network should be used as a tool for a linguistically well-grounded research in order to avoid some mistakes which can be caused, for instance, by a non-proper analogy with another kind of network analyses. However, one has to bear in mind that all studies focused on this topic represent only first steps in an unexplored area.

5 Preprocessing of Data for a Syntactic Complex Network Analysis – Pitfalls to be Avoided

Any complex network represents a relatively simple model of an observed system. In many cases, it is not difficult to determine both units, which are represented by nodes of the network, and relationships, which are represented by links connecting the nodes: e.g., World Wide Web, sexual relationships among people, a co-occurrence network of word forms. However, the analysis of syntactic networks is not the case. Even though the dependency grammar formalism brings general principles for sentence parsing, the variability of particular parsing systems is rather large. Further, the majority of syntactic networks analyses uses syntactic treebanks as the source of language data; the treebanks are language corpora containing a syntactic annotation which is usually processed automatically. One should keep in mind that there is not a unique annotation scheme for automatic parsing and that different annotation schemes lead to different results (Boyd and Meurers 2008). To illustrate that these differences are not negligible for network analysis, we present various approaches to coordination (Liu et al. 2010) and the problem of lemmatization.

Coordination (Crysmann 2006, p. 183) is

> a combination of like or similar syntactic units into some larger group of the same category or status, typically involving the use of a coordinating conjunction, such as *and* or *or*, to name just two. The units grouped together by means of a coordinating conjunction are usually referred to as conjuncts (or conjoints).

This phenomenon is a difficult point especially for dependency syntax, in which binary asymmetrical relations are basic elements (Lobin 1993; Osborne 2003; Temperley 2005). The problem is that all members of a coordination group fill one syntactic "slot", in fact. For instance, there is one accusative object (*Mary*) in the sentence (8)

(8) *I see Mary,*

while in the sentence (9)

(9) *I see Peter, Paul, and Mary,*

the accusative object – again one(!) object as well as in sentence (9), from the syntactic point of view – is represented by three members (*Peter, Paul, Mary*). As the development of syntactic studies reveals, it is impossible to find "the best" annotation scheme of this phenomenon which would be broadly accepted among linguists.

Following Liu et al. (2010), we can parse sentence (9) in three different ways, as is presented in Figure 5. It should be emphasized that each parsing is linguistically well grounded (Tesnière 1959; Schubert 1987; Mel'čuk 1988; Liu and Huang 2006; Hudson 2007).

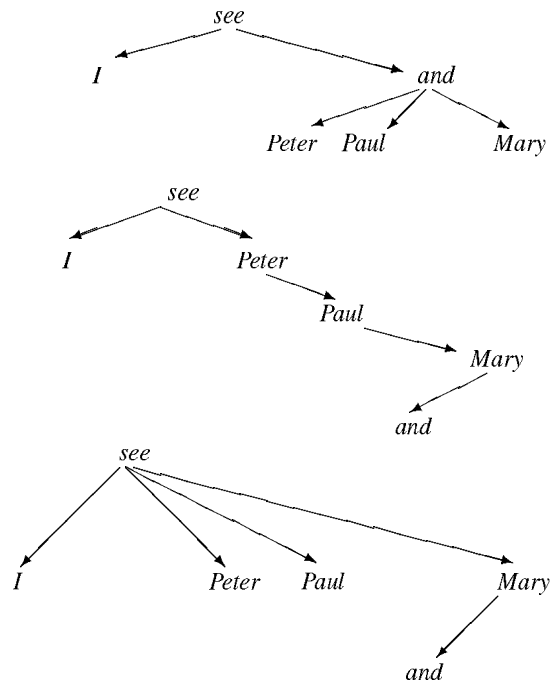

Fig. 5 Three different representations of the structure of (9).

We remind again that a syntactic dependency network is constructed by accumulating sentence structures, such as are in Fig. 5. At the first sight it is obvious that syntactic networks based on different annotation schemes will have different structures. Specifically, the first approach (Fig. 5 on the top) will "favor" (with regard to the node degree) the conjunction, the approach from Fig. 5 in the middle will lead to a more uniform distribution of links in the network, while graph in the bottom will "favor" verbs." Naturally, the choice of an annotation scheme can strongly influence tests of hypotheses, for instance the analysis of verbs presented in Sect. 4.

Not only differences among approaches to syntactic relationships can significantly influence a syntactic network analysis. In addition, the annotation of linguistic units, which are represented by nodes in a syntactic network, should always be clearly presented. There are relatively few problems if one uses word forms. However, often word forms are not suitable units for the analysis, for many reasons. For instance, if one aims at working with word as a semantic unit, the use of lemmas is a

more reasonable approach. At the first sight, the use of lemmas is unproblematic; a lemma represents the canonical form for all word forms, so, for instance, the lemma *sing* represents the word forms *sing, sings, sang, sung*. However, the situation becomes more problematic when polysemy enters into the play; cf. different meanings of the word *school* in sentence (10), (11), and (12):

(10) *I visited her school*

(11) *This linguistic school has influenced the history of science in the U.S.A.*

(12) *My experience is drawn from the school of life*

Can all these occurrences of the word *school* be represented by the unique lemma? Or by two lemmas, one denoting a building and the other denoting an abstract notion? Or even by three different lemmas because the meaning of *school* seems not to be the same in sentences (11) and (12)? There are good linguistic reasons to follow all of these three approaches. This fact is reflected by different methods of lemmatization: some corpora annotate differences caused by polysemy (but the question is how fine-grained the annotation is), while others not. As in the case of coordination, it is obvious that different lemmatizations lead to different network characteristics.

Coordination and lemmatization are obviously not the only phenomena which can be (and actually they are) annotated differently in particular corpora – in fact, they were used just for an illustration of a more general problem. We emphasize the impact of the annotation scheme because this factor seems to be, in our opinion, neglected in language complex network analyses. To avoid shortcomings of this kind, it is necessary to know details of language data used for the analysis, especially if one's goal is to use complex networks for comparative studies (e.g., language typology). The cases of coordination and lemmatization show that the problem of preprocessing data for the network analysis is not trivial and that it needs to be taken seriously. In the ideal case, in any analysis of this kind a technical report should be enclosed which would provide an opportunity to critically analyze presented results.

6 Applications of Syntactic Complex Networks to Language Typology and Acquisition

Syntactic networks are not applied only to syntax studies. The complex network theory (Newman 2011) offers many ways how to analyze global network properties (not only small-world and scale-freeness) which can be used to compare individual systems modeled by the network. There are currently two main directions in the applications of syntactic network analysis: language typology and language acquisition.

6.1 Language Typology

Despite the problems related to the role of syntax (we mean full syntax properties of present languages) in syntactic complex networks presented in Sect. 4, an application of global network characteristics to language typology seems to be fruitful. It is no surprise, if one realizes that some global network characteristics are connected to word frequency and the Zipf's law (see Sect. 4); many indexes based on word frequency were used for language typological studies (Popescu, Altmann, et al. 2009; Popescu, Mačutek, and Altmann 2009; Popescu et al. 2010; Popescu et al. 2011). However, syntactic networks express more than frequency characteristics and, consequently, their analysis could enhance typological characteristics of languages.

Studies focused on a typological interpretation of syntactic network properties take global characteristics of syntactic networks and exploit a cluster analysis to evaluate (dis)similarities among observed languages (Liu and Li 2010; Liu and Xu 2012). Satisfactory results of these studies justify this approach, even though some linguistic reasons (e.g., the impact of word frequencies vs. syntax properties) causing the differences remain unclear. From a linguistic point of view, a more interesting approach is represented by studies which add an explanation of a linguistic meaning of particular global network characteristics. Specifically, Čech and Mačutek (2009) analyze in detail a potential impact of grammar on differences between lemma and word form syntactic networks (the average degree and clustering coefficient are scrutinized) and try to determine which properties should be influenced by a typological character of language, on the one hand, and which by language usage (e.g., genre differences), on the other. Further, Liu and Xu (2011) compare lemma and word form syntactic networks of 15 languages and show how differences among global network characteristics reflect morphological variation degrees and a morphological complexity. Finally, Abramov and Mehler (2011) explain and discuss the linguistic meaning of particular global network characteristics thoroughly whenever it is possible and, consequently, offer a deeper insight to the issue (11 languages were used for the analysis).

It is important to note that there are great similarities between language typology (i.e., language classification) and text classification; one conducts text classification when the texts in question are from different genres of the same language and language classification or typological identification of languages when the texts in question are from different languages. This fact can have a significant impact on the results of a complex network based language typology. Methodologically, there is a fundamental problem related to a usage-based language typology analysis in general, in the previous studies (Liu and Li 2010; Liu and Xu 2011; Liu and Xu 2012; Abramov and Mehler 2011). Specifically, the syntactic dependency networks in these studies are based on language data which are not necessarily consistent in semantic content and genre. The basic assumption of language classification based on syntactic dependency networks is that the topological similarities and differences of these networks (manifested by their complex network parameters) reflect the similarities and differences of the corresponding languages. However, heterogeneity in

the semantic content and genre of the language data selected, which is independent of the similarities and differences of the languages, may also contribute to the topological similarities and differences of the corresponding syntactic dependency networks and thus may affect the results of language classification. Therefore, a more desirable type of language data for complex network based language classification are parallel texts (i.e., a collection of texts with the same semantic content but in different languages, e.g., a novel plus its translations in different languages), which are consistent in both semantic content and genre (c.f. Kelih 2009). However, a requirement to analyze parallel texts is constrained by technical reasons, up to now; even though it is possible to annotate language data automatically, tools for the annotation are usually not easily available and a manual annotation of syntax for a network analysis is almost impossible because a huge amount of data is needed. In addition, an existence of parallel treebanks (e.g., the Prague Czech-English Dependency Treebank 2.0; SMULTRON - Stockholm MULtilingual Treebank; GRUG Parallel Treebank) is rather exceptional and their range (expressed by the number of languages) is too small.

6.2 Language Acquisition

The complex network theory enables not only to model and interpret real systems from a global point of view but also to analyze a global dynamic behavior of the systems. Consequently, the use of network analysis for modeling language acquisition should be no surprise, because language acquisition is nothing else than a dynamic process. What is more surprising is a relatively rare application of the network analysis in this branch of science; one could expect that a successful application of this methodology to modeling other dynamic systems should trigger its usage in this research area as well. Moreover, existing results are promising (Ninio 2006; Ke and Yao 2008; Corominas-Murtra et al. 2009; Corominas-Murtra et al. 2010; Hills et al. 2009). According to us, both an unfamiliarity of scientists focused on the language acquisition with the network theory and a relative technical difficulty (especially in the case of syntactic networks) are the main reasons of this state.

Ninio (2006) was the first who tried to use the network theory for modeling language acquisition focusing on syntax, to our knowledge. Despite an interest of syntax, her approach cannot be interpreted as an analysis of syntactic networks in the sense as is presented here (cf. Sect. 2). Specifically, she uses bipartite networks (this kind of network contains different classes of nodes, here one class represents speakers while the other words) to model language behavior of both mothers and children and then observes distributions of words, verbs etc. Discovered similarities of distributions (all are power-law) lead her to a conclusion as follows (Ninio 2006, p. 141):

> Application of Complexity Theory to language development sheds new light on the stance of the learner vis-à-vis the linguistic environment. It sees language as a network of speakers and the speech items they produce which children join when they, too, start to produce similar items. Developmental data shows that children act just like

> Google when it searches the Web: they pick popular items, but only if their content is relevant for them. The results support a view of children as free agents exercising Preferential Attachment when they develop their minds and acquire knowledge in a social environment.

Corominas-Murtra et al. (2009, 2010) focused on dynamics of a large-scale organization of the use of syntax. Particularly, they used language data from child's corpora, parsed them and, finally, modeled a development of syntactic networks based on this data. As a result, they observed dynamics of the syntax behavior of children between theirs 22 and 28 months. Their analysis reveals both two different regimes of children's syntactic behavior and a sharp transition between these regimes. Specifically, a tree-like organization of a syntactic network before the transition (around 24 month) is suddenly replaced by a much larger, heterogeneous network which has global properties similar to adults networks. Further, the transition is accompanied by a strong reorganization of the network; in the pre-transition stage degenerated lexical items, such as *it*, are words with the highest degree while after the transition functional items, such as *a* or *the*, replaced them. Even though author's conclusions regarding the innateness are not convincing, according to us, and despite some methodological problems of the analysis (cf. Sect. 4), the study shows that syntactic network analysis can bring interesting findings which contribute to a better understanding of the process of language acquisition, at least heuristically.

7 Conclusion

Syntax is one of the main components of the human language system. However, due to the lack of means, the emergence of syntax was difficult to study in the past. Nowadays, complex networks provide a feasible tool. Therefore, in recent years, the construction of syntactic networks and investigations of their properties have become important and interesting fields in language research. Promising results on syntactic global network characteristics were achieved in some branches of linguistics, especially in language typology and language acquisition. On the other hand, the development revealed also several pitfalls (some of which are, admittedly, already at least partially solved):

1. There are some exaggerated interpretations of (almost ubiquitous) small-worldliness and scale-freeness of language networks (cf. Solé 2005, Solé et al. 2010, Corominas-Murtra et al. 2009, and Corominas-Murtra et al. 2010).

   The problem is more of a historical than a scientific character, as studies of random language complex network clearly show (Liu and Hu 2008) that the global properties of language complex networks are a consequence of word frequencies rather than syntax. Therefore, the complex network properties are a necessary, not a sufficient condition for syntax. However, given that the paper Solé (2005) was published in one of the most prestigious scientific journals (*Nature*), it remains to be influential.

2. Automatic parsing is problematic in case of coordination – there are several linguistically substantiated possibilities which result in dramatically different representations, cf. Sect. 5.
3. Similarly, semantics also has an impact on syntactic complex networks (cf. the example of polysemy in Sect. 5).

Consequently, we allow ourselves to summarize some challenges which a researcher using syntactic complex network faces:

1. Particular characteristics of syntactic networks need more in-depth linguistic interpretations.
2. Properties of full syntax of present-day language and their impact on network characteristics should be analyzed in detail.
3. Either a universal syntactic dependency-based parsing formalism, which would be a basis for a more detailed study on syntactic networks, must be searched for, or, at least, one should take into account several possibilities of parsing and then compare results.
4. The relationship between syntactic and cognitive networks should be investigated.

These open problems are symptoms of the development, rather than indications that complex networks should not be used to study the human language syntax. We are convinced that the analysis of global syntactic dependency networks is a helpful tool in language research. It can contribute – under the condition that the results obtained be linguistically interpretable and interpreted – to a deeper understanding of basic and important issues of the human language and cognition, such as, for example, the emergence of syntax, syntactic relations, and connections between syntax and cognition.

Acknowledgements. The authors are thankful to Ramon Ferrer-i-Cancho and to the anonymous referees. Their comments significantly improved the article.

Radek Čech, Ján Mačutek and Haitao Liu were supported by the Czech Science Foundation (grant no. P406/11/0268 Historical semantics), by the VEGA grant agency (grant no. 2/0036/12) and by the National Social Science Foundation of China (grant no. 11&ZD188), respectively.

References

Abramov, O., Mehler, A.: Automatic language classification by means of syntactic dependency networks. Journal of Quantitative Linguistics 18, 291–336 (2011)

Allerton, D.J.: Valency grammar. In: Brown, K. (ed.) The Encyclopedia of Language and Linguistics, pp. 4878–4886. Elsevier (2005)

Baixeries, J., Elvevåg, B., Ferrer i Cancho, R.: The evolution of the exponent of Zipf's law in language ontogeny. PLoS ONE 8(3), e53227 (2013)

Barabási, L.-A.: Linked: The New Science of Networks. Perseus, Cambridge (2002)

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

Baronchelli, A., Ferrer i Cancho, R., Pastor-Satorras, R., Chater, N., Christiansen, M.H.: Networks in cognitive science. Trends in Cognitive Sciences 17, 348–360 (2013)
Boyd, A., Meurers, D.: Revisiting the impact of different annotation schemes in PCFG parsing: A grammatical dependency evaluation. In: Proceedings of the Workshop on Parsing German, pp. 24–32 (2008)
Caldarelli, G.: Scale-Free Networks: Complex Webs in Nature and Technology. Oxford University Press, Oxford (2007)
Čech, R., Mačutek, J.: Word form and lemma syntactic dependency networks: a comparative study. Glottometrics 19, 85–98 (2009)
Čech, R., Mačutek, J., Žabokrtský, Z.: The role of syntax in complex networks: local and global importance of verbs in a syntactic dependency network. Physica A: Statistical Mechanics and its Applications 390, 3614–3623 (2011)
Choudhury, M., Mukherjee, A.: The structure and dynamics of linguistic networks. In: Ganguly, N., Deutsch, A., Mukherjee, A. (eds.) Dynamics on and of Complex Networks: Applications to Biology, Computer Science and the Social Sciences, pp. 145–166. Birkhäuser, Boston (2009)
Corominas-Murtra, B., Valverde, S., Solé, R.V.: The ontogeny of scale-free syntax networks: Phase transition in early language acquisition. Advances in Complex Systems 12, 371–392 (2009)
Corominas-Murtra, B., Valverde, S., Solé, R.V.: Emergence of scale-free syntax networks. In: Nolfi, S., Mirolli, M. (eds.) Evolution of Communication and Language in Embodied Agents, pp. 83–101. Springer, Heidelberg (2010)
Crysmann, B.: Coordination. In: Brown, K. (ed.) The Encyclopedia of Language and Linguistics, pp. 183–196. Elsevier, Oxford (2006)
de Saussure, F.: Cours de linguistique générale. Payot, Paris (1979)
Eguíluz, V.M., Chialvo, D.R., Cecchi, G.A., Baliki, M., Apkarian, A.V.: Scale-free brain functional networks. Physical Review Letters 94, 018102 (2005)
Ferrer i Cancho, R.: The structure of syntactic dependency networks: insights from recents advances in network theory. In: Altmann, G., Levickij, V., Perebyinis, V. (eds.) Problems of Quantitative Linguistics, pp. 60–75. Ruta, Chernivtsi (2005a)
Ferrer i Cancho, R.: Zipf's law from a communicative phase transition. European Physical Journal B 47, 449–457 (2005b)
Ferrer i Cancho, R.: Why do syntacitc links not cross? Europhysics Letters 76, 1228–1235 (2006)
Ferrer i Cancho, R., Riordan, O., Bollobás, B.: The consequences of Zipf's law for syntax and symbolic reference. Proceedings of the Royal Society of London Series B 272, 561–565 (2005)
Ferrer i Cancho, R., Solé, R.V.: Least effort principle and the origins of scaling in human language. Proceedings of the National Academy of Sciences of the USA 100, 788–791 (2003)
Ferrer i Cancho, R., Solé, R.V., Köhler, R.: Patterns in syntactic dependency networks. Physical Review E 69 (2004)
Fitch, W.T., Hauser, M.D.: Computational constraints on syntactic processing in a nonhuman primate. Science 303, 377–380 (2004)
Grinstein, G., Linsker, R.: Synchronous neural activity in scale-free network models versus random network models. Proceedings of the National Academy of Sciences of the USA 102, 9948–9953 (2005)
Hauser, M.D., Chomsky, N., Fitch, W.T.: The faculty of language: what is it, who has it and how did it evolve? Science 298, 1569–1579 (2002)

Hills, T.T., Maouene, M., Maouene, J., Sheya, A., Smith, L.: Longitudinal analysis of early semantic networks: preferential attachment or preferential acquisition? Psychological Science 20, 729–739 (2009)
Hudson, R.A.: Language Networks: The New Word Grammar. Oxford University Press, Oxford (2007)
Ke, J., Yao, Y.: Analyzing language development from a network approach. Journal of Quantitative Linguistics 15, 70–99 (2008)
Kelih, E.: Slawisches Parellel-Textkorpus: Projektvorstellung von "Kak zakaljalas' stal' (KZS)". In: Kelih, E., Levickij, V.V., Altmann, G. (eds.) Methods in Text Analysis, pp. 106–124. Ruta, Chernivtsi (2009)
Köhler, R.: Zur linguistischen Synergetik. Struktur und Dynamik der Lexik. Quantitative Linguistics, vol. 31. Brockmeyer, Bochum (1986)
Köhler, R.: Synergetic linguistics. In: Altmann, G., Köhler, R., Piotrowski, R.G. (eds.) Quantitative Linguistics. An International Handbook, pp. 760–775. de Gruyter, Berlin (2005)
Liu, H.: The complexity of Chinese syntactic dependency networks. Physica A: Statistical Mechanics and its Applications 387, 3048–3058 (2008)
Liu, H., Hu, F.: What role does syntax play in a language network? EPL 83 (2008)
Liu, H., Huang, W.: A Chinese Dependency Syntax for Treebanking. In: Proceedings of the 20th Pacific Asia Conference on Language, Information and Computation, pp. 126–133. Tsinghua University Press, Beijing (2006)
Liu, H., Li, W.: Language clusters based on linguistic complex networks. Chinese Science Bulletin 55, 3458–3465 (2010)
Liu, H., Xu, C.: Can syntactic network indicate morphological complexity of a language? EPL 93 (2011)
Liu, H., Xu, C.: Quantitative typology analysis of Romance languages. Poznań Studies in Contemporary Linguistics 48, 597–625 (2012)
Liu, H., Zhao, Y., Huang, W.: How do local syntactic structures influence global properties in language networks? Glottometrics 20, 38–58 (2010)
Lobin, H.: Koordinationssyntax als prozedurales Phänomen. Narr, Tübingen (1993)
Mehler, A.: Large text networks as an object of corpus linguistic studies. In: Lüdeling, A., Kytö, M. (eds.) Corpus Linguistics. An International Handbook, pp. 328–382. de Gruyter (2007)
Mel'čuk, I.A.: Dependency syntax: Theory and Practice. State University of New York Press, Albany (1988)
Mel'čuk, I.A.: Levels of Dependency in Linguistic Description: Concepts and Problems. In: Ágel, V., Eichinger, L.M., Eroms, H.W., Hellwig, P., Herringer, H.J., Lobin, H. (eds.) Dependency and Valency. An International Handbook of Contemporary Research, vol. 1, pp. 188–229. de Gruyter, Berlin (2003)
Newman, M.E.J.: Networks: An Introduction. Oxford University Press, Oxford (2011)
Ninio, A.: Language and the Learning Curve: A New Theory of Syntactic Development. Oxford University Press, Oxford (2006)
Nowak, M.A., Plotkin, J.B., Jansen, V.A.A.: The evolution of syntactic communication. Nature 404, 495 (2000)
Osborne, T.: The Third Dimension: A Dependency Grammar Theory of Coordination for English and German. PhD thesis. Pennsylvania State University (2003)
Popescu, I.-I., Altmann, G., Grzybek, P., Jayaram, B.D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., Vidya, M.N.: Word Frequency Studies. de Gruyter, Berlin (2009)
Popescu, I.-I., Čech, R., Altmann, G.: The Lambda-Structure of Texts. RAM-Verlag, Lüdenscheid (2011)

Popescu, I.-I., Mačutek, J., Altmann, G.: Aspects of Word Frequencies. RAM-Verlag, Lüdenscheid (2009)
Popescu, I.-I., Mačutek, J., Kelih, E., Čech, R., Best, K.-H., Altmann, G.: Vectors and Codes of Text. RAM-Verlag, Lüdenscheid (2010)
Schubert, K.: Metataxis: Contrastive Dependency Syntax for Machine Translation. Foris, Dordrecht (1987)
Solé, R.V.: Syntax for free? Nature 434, 289 (2005)
Solé, R.V., Corominas-Murtra, B., Valverde, S., Steels, L.: Language networks: Their structure, function and, evolution. Complexity 15, 20–26 (2010)
Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., Nivre, J.: The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In: CoNLL 2008: Proceedings of the 12th Conference on Computational Natural Language Learning, pp. 159–177 (2008)
Temperley, D.: The Dependency Structure of Coordinate Phrases: A Corpus Approach. Journal of Psycholinguistic Research 34, 577–601 (2005)
Tesnière, L.: Eléments de la syntax structurelle. Klincksieck, Paris (1959)
Zipf, G.K.: Human Behavior and the Principle of Least Effort. An Introduction to Human Ecology. Addison Wesley, Cambridge (1949)