Digital Scholarship in the Humanities Advance Access published July 25, 2015

Polysemy and Synonymy in Syntactic Dependency Networks

Radek Čech
Department of Czech Language, University of Ostrava, Ostrava, Czech Republic

Ján Mačutek
Department of Applied Mathematics and Statistics, Comenius University, Bratislava, Slovakia

Zdeněk Žabokrtský
Institute of Formal and Applied Linguistics, Charles University in Prague, Prague, Czech Republic

Aleš Horák
Natural Language Processing Centre, Masaryk University, Brno, Czech Republic

Correspondence: Radek Čech, Department of Czech Language, University of Ostrava, Reální 5, Ostrava 701 03, Czech Republic.
E-mail: cechradek@gmail.com

Abstract

The relationship between two important semantic properties (polysemy and synonymy) of language and one of the most fundamental syntactic network properties (a degree of the node) is observed. Based on the synergetic theory of language, it is hypothesized that a word which occurs in more syntactic contexts, i.e. it has a higher degree, should be more polysemous and have more synonyms than a word which occurs in less syntactic contexts, i.e. it has a lesser degree. Six languages are used for hypotheses testing and, tentatively, the hypotheses are corroborated. The analysis of syntactic dependency networks presented in this study brings a new interpretation of the well-known relationship between frequency and polysemy (or synonymy).

1 Introduction

A complex network (cf. Newman 2011) is a model of a system. It contains sets of nodes, representing entities, and links, representing relations among them. Syntactic properties of language can be described, inter alia, as a system which contains both words and syntactically motivated relations between pairs of words (cf. Mel'čuk, 1988; Hudson, 2007). Consequently, it is not surprising that complex networks have been used for an analysis of syntax almost since the complex network theory emerged (Barabási and Albert, 1999; Barabási, 2002). The analyses of syntactic complex networks opened new insights into a language functioning in the last decade. Specifically, new models of language acquisition were proposed (cf. Ninio, 2006, 2011; Ke and Yao, 2008; Corominas-Murtra et al., 2010), the relation between syntax and communication needs was analysed (cf. Ferrer i Cancho et al., 2005; Ferrer i Cancho, 2006a), differences of statistical properties of syntactic networks were used for typological

doi:10.1093/llc/fqv028

studies (cf. Čech and Mačutek, 2009; Liu and Li, 2010; Abramov and Mehler, 2011; Liu and Xu, 2011), the origin of projectivity, i.e. the fact that syntactic dependency crossing occurs but very rarely, was inquired (cf. Ferrer i Cancho, 2006b, 2008), as well as the origin of syntax and its role in syntactic complex networks (cf. Ferrer i Cancho et al., 2005; Solé, 2005; Liu and Hu, 2008; Liu et al., 2010; Čech et al., 2011). Despite these achievements, a linguistic explanation of syntactic network properties is strongly needed, especially because the majority of language network analyses were merely descriptive, and not explanative (cf. Ferrer i Cancho, 2010), and focused on global network characteristics (Cong and Liu, 2014a, b). On the other hand, network analyses can bring a finer-grained interpretation of some traditional linguistic problems (Liu, 2011; Gao et al., 2014), as is also presented in this study in case of polysemy and synonymy functioning, and the complex network approach have a potential to be inspiring for a theory of language, according to Ferrer i Cancho (2014).

One possible way how to explore a syntactic network from a linguistic point of view is to find those language characteristics which should correlate with syntactic network properties because of some theoretical reasons. Generally, the approach of this kind was developed by synergetic linguistics in Köhler (1986, 2005a). It was used to analyse phonology (Coloma, 2014), lexical properties (Köhler, 2005b; Köhler and Altmann, 1993; Wang, 2014), syntax (Köhler, 1999, 2007, 2012; Čech and Mačutek, 2010; Čech et al., 2010, 2011; Liu, 2011; Gao et al., 2014), and morphology (Prün and Steiner, 2005). The present study follows this approach, and its aim is to scrutinize the relationship between syntax properties modelled by complex network (specifically, a number of syntactically motivated lexical contexts of a given word are taken into account; in the complex network, it is determined by a degree of the node representing the word; see Section 3 for more details), on the one hand, and semantic properties (word polysemy and synonymy), on the other. The degree of the word is determined by the number of syntactic dependencies in all syntactic trees in a corpus; an aggregation of these trees makes the syntactic complex network.

We hypothesize that a word—actually, we considered a lemma, i.e. a canonical word form[1]—which occurs in more syntactically motivated lexical contexts, i.e. it has a higher degree, should be more polysemous and have more synonyms[2] than a word which occurs in less contexts, i.e. it has a lesser degree. Based on this deduction, we put forth hypotheses as follows:

(1) the higher the out-degree/in-degree of a word, the more meanings it has;

(2) the higher the out-degree/in-degree of a word, the more synonyms it has.

The out-degree of a word expresses the total number of words which are connected to it as its modifiers, i.e. the number of its syntactically dependent words, while the in-degree of the word expresses the total number of words which the word under consideration is connected to, i.e. the number of its heads, in an observed sample of language (see Section 3).

Since it is assumed that these hypotheses are not language specific, six languages (Czech, Dutch, English, German, Italian, and Spanish) are used for their testing. Admittedly, they do not cover all types of languages (all of them belong to the Indo-European family), but we consider them as a sufficient sample for the first preliminary analysis in this field.

The article is organized as follows: the status of polysemy and synonymy in language is presented in Section 2; a language material together with the methodology used is introduced in Section 3; Section 4 is focused on results; and the article is ended by Conclusion (Section 5).

2 Polysemy and synonymy of the word

At least since Zipf (1935), it is well known that the semantic aspect of language (i.e. a meaning of any language unit) is closely connected to other language properties (e.g. relative frequency, degree of intensity of accent, and degree of crystallization of the configuration). Probably the most systematical incorporation of certain semantic characteristics of language into a language model was presented by Köhler (1986, 2005b, 2012). Specifically, in the synergetic model of language, Köhler focuses on two semantic properties of language units, viz. polysemy and synonymy. He hypothesizes particular relationships among them and other language characteristics, such as frequency, word length, inventory sizes of language units, and functional loads of language units, and tests all hypotheses experimentally. Consequently, both polysemy and synonymy can be viewed as language properties which are ruled by a complex mechanism emerging as a result of intricate interrelations among so-called communication requirements (cf. Köhler, 2005b). The system of these requirements represents a novel way of developing an approach originally proposed by Zipf (1949). According to Zipf, the properties of language are determined by a principle which should have a crucial impact on the human behaviour in general; he calls it the principle of least effort.

As for the relationship between the principle of least effort and polysemy, Zipf (1949: 20–1) pointed out that

> [f]rom the viewpoint of speaker (*the speaker's economy*) who has the job selecting not only the meanings to be conveyed but also the words that will convey them, there would doubtless exist an important latent economy in a vocabulary that consisted exclusively of one single word—a single word that would mean whatever the speaker wanted it to mean. (...) But from the viewpoint of the auditor, who has the job of deciphering the speaker's meanings, the important internal economy of speech would be found rather in a vocabulary of such size that is possessed a distinctly different word for each different meaning to be verbalized.

As a result of these opposite communication strategies, an equilibrium regarding the number of meanings to be conveyed by particular words should emerge; this assumption was corroborated in Zipf (1949), Tuldava (1998), Ferrer i Cancho and Solé (2003), Ferrer i Cancho (2005a), and Kelih (2008). It was shown that the distribution of the number of meanings follows certain regularities.

Inspired by these findings, in the following paragraphs we hypothesize some relationships between both polysemy and synonymy, on the one hand, and degrees of nodes representing canonical word forms (lemmas) in a syntactic complex network, on the other.

First, it is necessary to realize that the meaning of a word is strongly influenced by words which are in a syntactic relation with it in an actual language usage. Obviously, the presence of syntactically related words makes it possible to express meaning refinements; thus, syntactically related words take part in conveying the meaning by the word under the consideration. Consequently, it is reasonable to assume that, with an increasing variability of syntactic contexts of the word, the variability of subtle meaning differences of the word also increases. If the differences of meaning are 'striking enough', they are detected by a lexicographer and they are recorded in dictionaries or databases such as the WordNet (see below) which are usually used for the determination of the number of meanings of the word (i.e. its polysemy).

Further, from the speaker's point of view, the presence of syntactically related words enables the speaker to use a unique word for conveying more meanings because syntactically related words take part in its meaning expression and its differentiation. This is in a clear agreement with the speaker's economy—the speaker tends to use as few words as possible for as many meanings as possible (see above); obviously, the 'cost' of the increasing polysemy of the word is the length (or complexity) of the syntactic constructions needed for the expression of the meaning (cf. Köhler, 2012). From the auditor's point of view, the presence of syntactically related words makes the determination of the meaning of the word easier, cf. the priming effect (Hoey, 2005). So, the auditor's economy 'presses' the speaker to use words with more meanings in a syntactically more complex environment because it makes it for him easier to determine the meaning of the entire expression. If the speaker would ignore the auditor's economy, the probability of auditor's misunderstanding increases, and, consequently, the probability that the auditor asks the speaker for a new, better explanation increases. From the

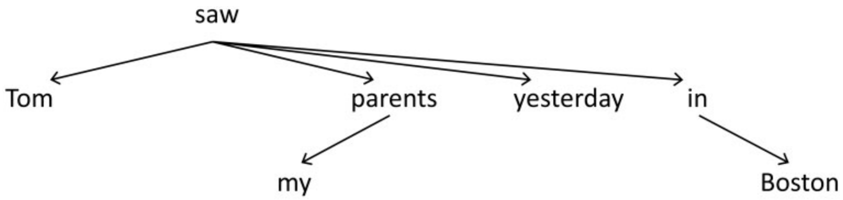

Fig. 1 The tree graph expressing the structure of the sentence 'Tom saw my parents yesterday in Boston' based on the dependency grammar formalism. Links between words represent the syntactic dependency relations, the arrows express the direction of the dependency

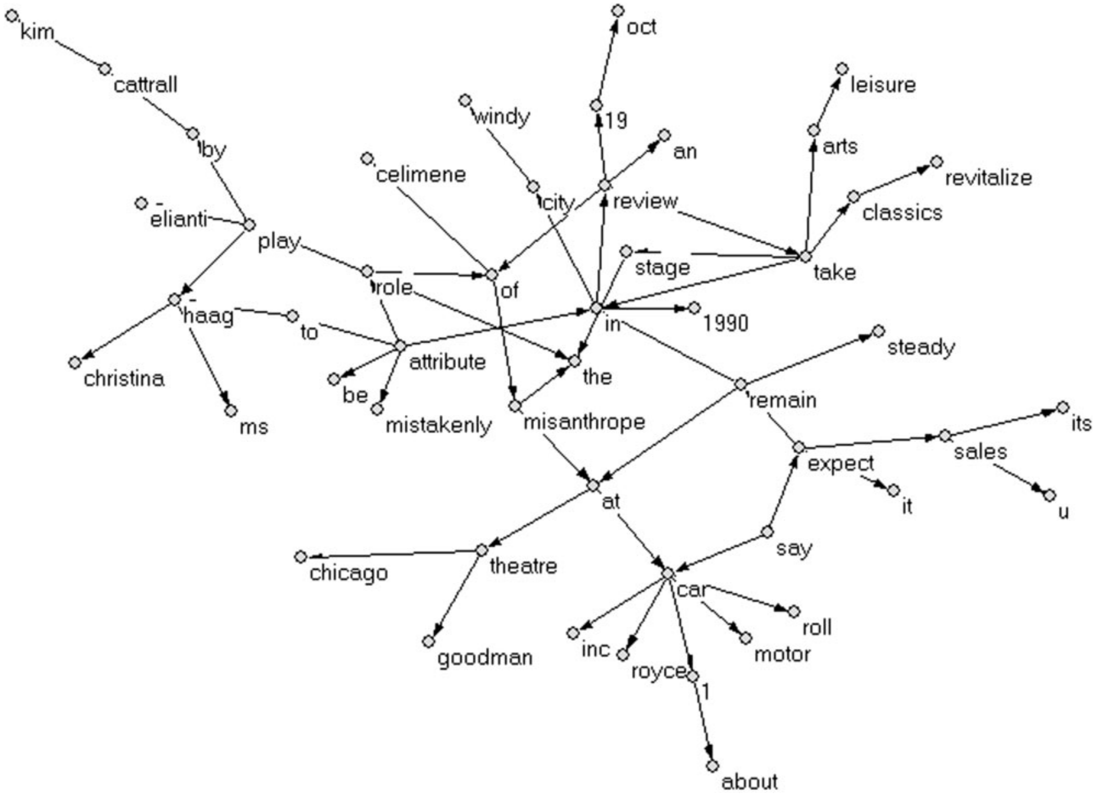

Fig. 2 The network containing the first fifty lemmas from the English treebank

speaker's point of view, the repetition of the expression conveying the same meaning is in a clear contradiction to his communication strategy (i.e. the speaker's economy).

In sum, because both the out-degree and the in-degree of the word represent the number of syntactically related words in a corpus, we hypothesize that the higher out-degree (or in-degree) of the word, the more meanings the word has.

As for the relationship between synonymy and the degree of the word, it should be viewed as a consequence of the relationship between polysemy and synonymy. The more meanings a word obtains, the more semantic domains of other words it penetrates and, consequently, these words become its synonyms (cf. Köhler, 1986; Wimmer and Altmann, 2001). Hence, because the relationship between the degree and polysemy is assumed, also

Table 1 Ranked distribution of out-degrees in the Czech treebank

Rank	Out-degree
1	7,441
2	3,489
3	2,112
4	1,333
5	914
...	...
11,939	0
11,940	0

the relationship between the degree and synonymy should be hypothesized, i.e. the higher out-degree/in-degree of the word, the more synonyms it has.

3 Language material and methodology

For empirical studies on the relation between variability of the syntactic context and polysemy of words, we naturally need two types of data resources: (1) syntactically annotated data, and (2) dictionary data with enumerated words' meanings. In our experiments, the role of the former type is played by dependency treebanks, while wordnets are used as the latter type. We managed to interlink the two types of information for six languages so far: Czech, Dutch, English, German, Italian, and Spanish. In spite of additional treebanks and wordnets being simultaneously available for several other languages, we were not able to include them into our experiments due to various technical obstacles, such as an insufficient size, a missing lemmatization, or an incompatible tokenization.

We did not use the original treebank shapes for building dependency networks, but we used their 'Prague dependency treebank—harmonized' forms, as introduced in Zeman et al. (2012), where many treebanks were transformed in order to maximize the compatibility of the resulting dependency trees with annotation guidelines for so-called analytical trees (surface-syntactic dependency trees) of the Prague Dependency Treebank (Hajič et al., 2006). For instance, subordinating conjunctions are heads of subordinating clauses according to the Prague dependency treebank style, analogously to prepositions being heads of prepositional groups. We do not claim that the Prague Dependency Treebank style is superior to other treebanks' conventions when it comes to building syntactic networks, but we prefer to treat all the languages as uniformly as possible, and the Prague Dependency Treebank harmonization was the only normalization that was readily available to us.

We used data originating from the treebanks listed below (cf. Zeman et al., 2012, for more details on the respective resources and associated normalization procedures):

Czech: Prague dependency treebank 2.0 (Hajič et al., 2006),
Dutch: Alpino treebank (van der Beek et al., 2002),
English: Penn treeBank 2 (Surdeanu et al., 2008),
German: Tiger treebank (Brants et al., 2002),
Italian: Italian syntactic-semantic treebank (Montemagni et al., 2003),
Spanish: AnCora (Taulé et al., 2012).

The first WordNet database was published by Miller et al. (1993) at the University of Princeton. Since then, more than seventy national wordnets were created following the same principles as the original Princeton WordNet for English (cf. Horák et al., 2008).

The data in wordnet databases are organized as networks of basic entities called 'synsets', synonym sets. Each synset corresponds to one meaning of a word or a collocation. The synonymy relation in wordnet is not the standard 'strict' synonymy, where synonymous words are simply identical (or nearly identical) in meaning (e.g. 'pretty' and 'handsome'). The word meanings in synsets are 'in a near synonymy' relation—they are synonymous in the sense that they can be exchanged in the 'same contexts'. For example, the synset 'exist:1, be:4' in wordnet relates the fourth meaning (generally called a 'literal') of 'to be' with the first meaning of 'to exist' with the definition of 'to have an existence'.

For the purpose of the current experiment, we used the data of six languages in the form resulting

Table 2 Mean values of out-degrees/in-degrees, the mean number of synonyms, and the mean number of meanings in the Czech treebank

Out-degree	Synonyms	Polysemy	In-degree	Synonyms	Polysemy
507.94	8.59	5.61	313.95	3.70	3.71
214.34	6.21	4.27	137.93	4.92	3.74
167.74	5.26	3.47	96.63	4.27	3.40
137.61	4.59	3.42	76.24	3.46	2.95
116.47	4.64	3.38	63.45	5.60	3.78
100.53	5.18	3.32	54.50	3.68	2.89
88.95	4.15	2.92	47.97	3.28	2.74
78.83	3.40	2.54	42.27	2.96	2.54
70.18	3.96	3.31	37.44	3.83	2.84
62.61	3.18	2.56	33.36	2.95	2.63
56.80	3.37	2.52	30.03	3.08	2.48
51.96	3.18	2.32	26.91	3.31	2.44
47.47	3.10	2.50	24.24	2.51	2.18
43.41	3.32	2.59	22.51	3.18	2.43
39.97	3.07	2.39	20.54	2.75	2.21
37.06	2.67	2.23	18.50	2.66	2.10
33.94	3.11	2.33	16.43	2.34	1.97
31.03	2.33	2.11	14.45	2.40	2.16
28.46	2.89	2.11	13.00	2.43	1.97
26.51	2.75	2.11	12.00	2.11	2.02
24.49	2.91	2.28	11.00	1.82	1.74
22.50	2.99	2.10	10.00	2.51	1.97
20.48	2.50	2.16	9.00	2.10	1.85
18.49	2.19	1.85	8.00	2.28	1.79
17.00	2.41	1.79	7.00	1.95	1.69
16.00	2.35	2.07	6.00	2.10	1.74
15.00	2.16	1.83	5.00	1.87	1.63
14.00	2.47	1.90	4.00	1.94	1.59
13.00	2.26	1.94	3.00	1.82	1.54
12.00	2.11	1.84	2.00	1.64	1.43
11.00	2.10	1.65	1.00	1.64	1.35
10.00	2.31	1.75	0.00	1.86	1.30
9.00	2.28	1.79			
8.00	2.15	1.67			
7.00	1.92	1.76			
6.00	1.97	1.62			
5.00	1.96	1.54			
4.00	1.76	1.54			
3.00	1.83	1.52			
2.00	1.66	1.41			
1.00	1.56	1.42			
0.00	1.21	1.34			

from the EuroWordNet and BalkaNet EU projects. The studied wordnets have the following statistics:

Czech WordNet: 32,116 words and collocations, 28,448 synsets, 43,958 literals;
Dutch WordNet: 56,329 words and collocations, 44,128 synsets, 70,356 literals;
English WordNet: 156,588 words and collocations, 117,597 synsets, 207,018 literals;
German WordNet: 17,098 words and collocations, 15,132 synsets, 20,453 literals;
Italian WordNet: 32,964 words and collocations, 40,406 synsets, 48,475 literals;

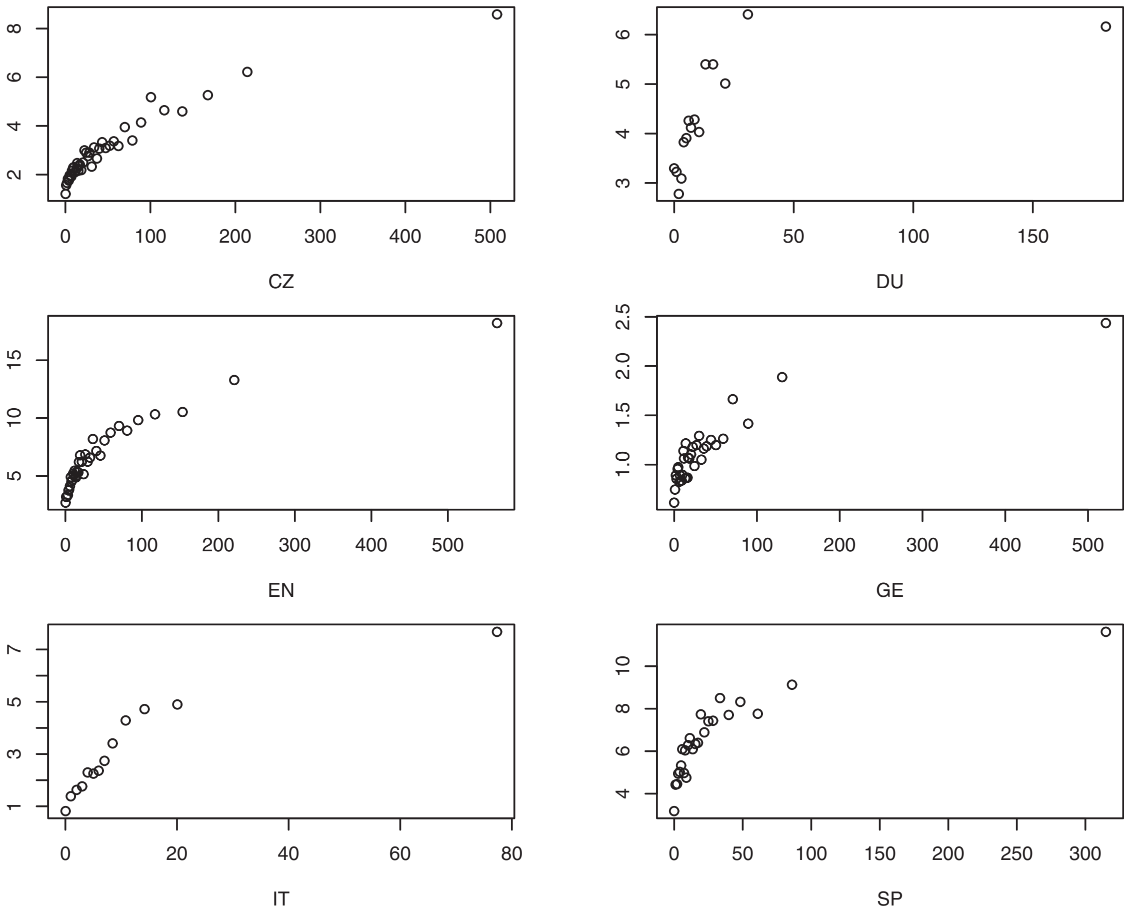

Fig. 3 Relation between the mean out-degree (x-axis) and the mean number of synonyms (y-axis)

Spanish WordNet: 32,606 words and collocations, 30,350 synsets, 52,362 literals.

Using these wordnet data, the nodes of the treebank-based syntactic networks were labelled with two values:

Number of synonyms—the number of other literals in all synsets for the lemma,

Number of meanings—the number of different literals/synsets for the lemma.

For example, the verb 'intend' has four meanings/synsets in the English WordNet:

(1) intend: 1, mean: 4, think: 7;
(2) intend: 2, destine: 2, designate: 4, specify: 6;
(3) mean: 1, intend: 3;
(4) mean: 3, intend: 4, signify: 1, stand for: 2.

The node 'intend' thus received nine synonyms and four meanings.

In constructing the syntactic dependency networks, the methods developed by Ferrer i Cancho et al. (2004) and Liu (2008) were followed. Each node of the complex network represents a particular lemma. Two nodes are linked if there is a dependency relation between the respective lemmas in the treebank. The links are directed; they go from the head to the modifier (see Fig. 1).

A single graph is used in our analysis, i.e. only unique connections between particular lemmas are counted. Thus, a global syntactic dependency network is constructed by accumulating sentence structures, and the network should be viewed as an emergent property of sentence structures (Ferrer i

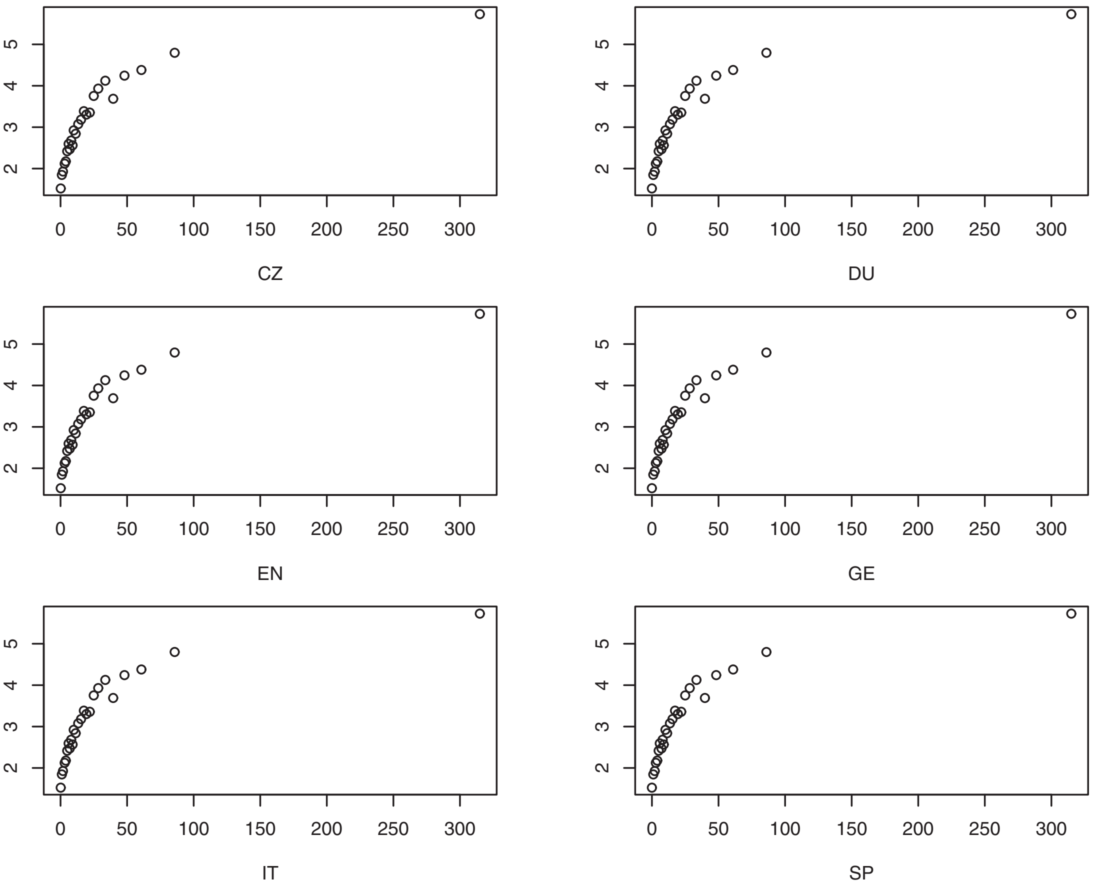

Fig. 4 Relation between the mean out-degree (x-axis) and the mean number of meanings (y-axis)

Cancho et al., 2004; Ferrer i Cancho, 2005b). The free software *Pajek 2.05* (cf. de Nooy et al., 2005) was used for network creation and computing. For an illustration, Fig. 2 shows the syntactic dependency complex network containing the first fifty lemmas from the English Treebank by Surdeanu et al. (2008).

4 Statistical Methodology

We follow the procedures described in Čech et al. (2011), which we remind here briefly. First, words which have zero meanings (it is determined by the absence of synsets in the WordNet) were omitted from our analyses. In the following step, rank-frequency distributions of out-degrees were constructed (for each language separately). The rank-frequency distribution is then exploited in the process of construction of the figures in Section 5 (ranks were used in the binning procedure; see below). Values of out-degrees are ordered from the highest to the lowest, the highest value receives the rank 1, the second highest one the rank 2, etc. Thereafter, we consider the ranks to be values of an auxiliary random variable and the out-degrees their frequencies. This auxiliary variable was used because the data are highly skewed—histograms constructed directly from both out- and in-degree values contain huge numbers of empty bins.

In the example from the Czech language, cf. Table 1, in the rank-frequency distribution we

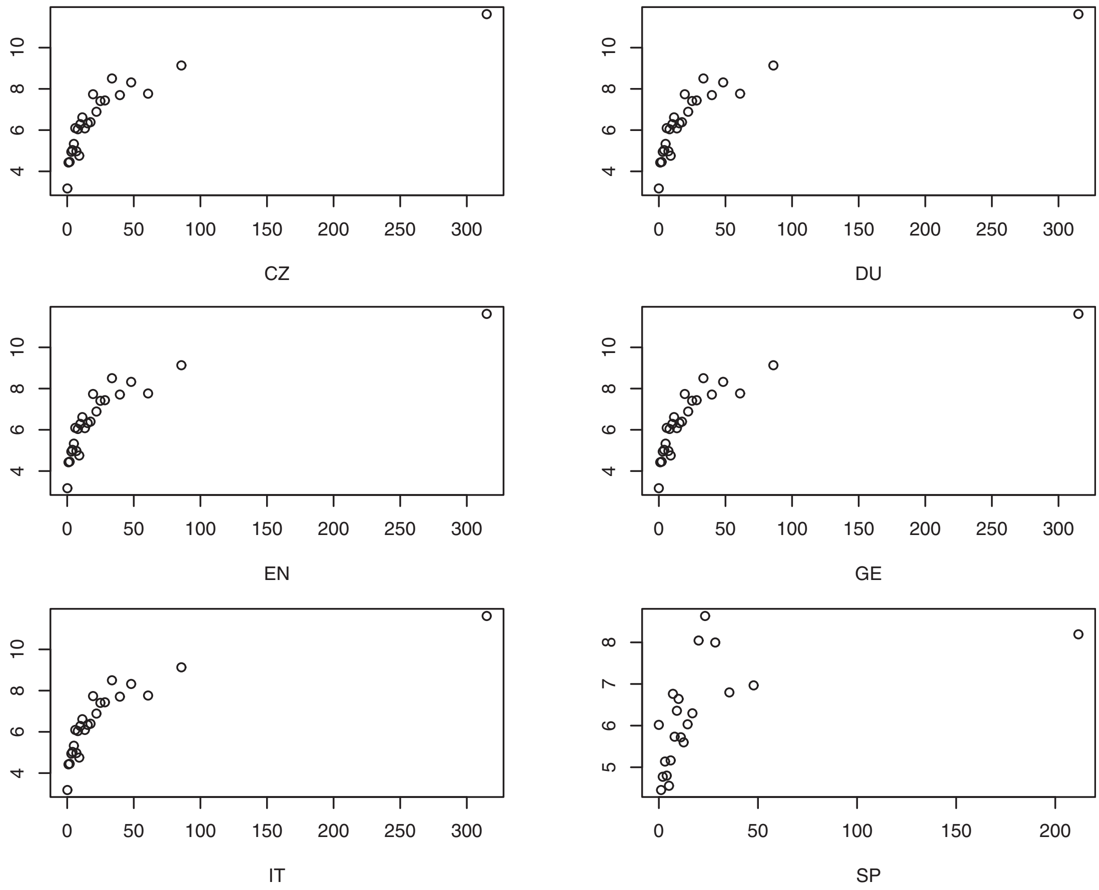

Fig. 5 Relation between the mean in-degree (x-axis) and the mean number of synonyms (y-axis)

assume to have 7,441-times value 1, 3,489-times value 2, etc; value 11,940 does not occur (it has frequency 0).

Next, a histogram was created from the rank-frequency distribution. Its bin width was calculated first according to Scott (1979). Then, the bins boundaries were adjusted so that equal values (i.e. equal out-degrees in the original data) are kept in one bin (e.g. all ranks corresponding to out-degree 1 belong to the same bin). The reason for adjusting the bin widths (which results in a histogram with different bin widths) is that there is no reasonable hierarchy of word lemmas which share the same out-degree in a network—they appear in the same order as they were entered into the treebanks (or then found in the process of network creation). Therefore, we assign all lemmas with the same out-degree to the same histogram bin.

Then, for each histogram bin we computed the mean out-degree, the mean number of synonyms, and the mean number of meanings (e.g. the mean polysemy) of the lemmas represented by ranks belonging to the bin. The same procedure was applied also to in-degrees. We thus obtained six data sets for out-degrees and six data sets for in-degrees. This approach (ranked frequencies) serves as a tool for a better visualization; the tests (cf. Section 5) were performed on the original data.

The two hypotheses from Section 1 (the higher out-degree/in-degree of a word, the more meanings it has; the higher out-degree/in-degree of a word, the more synonyms it has) were tested using the

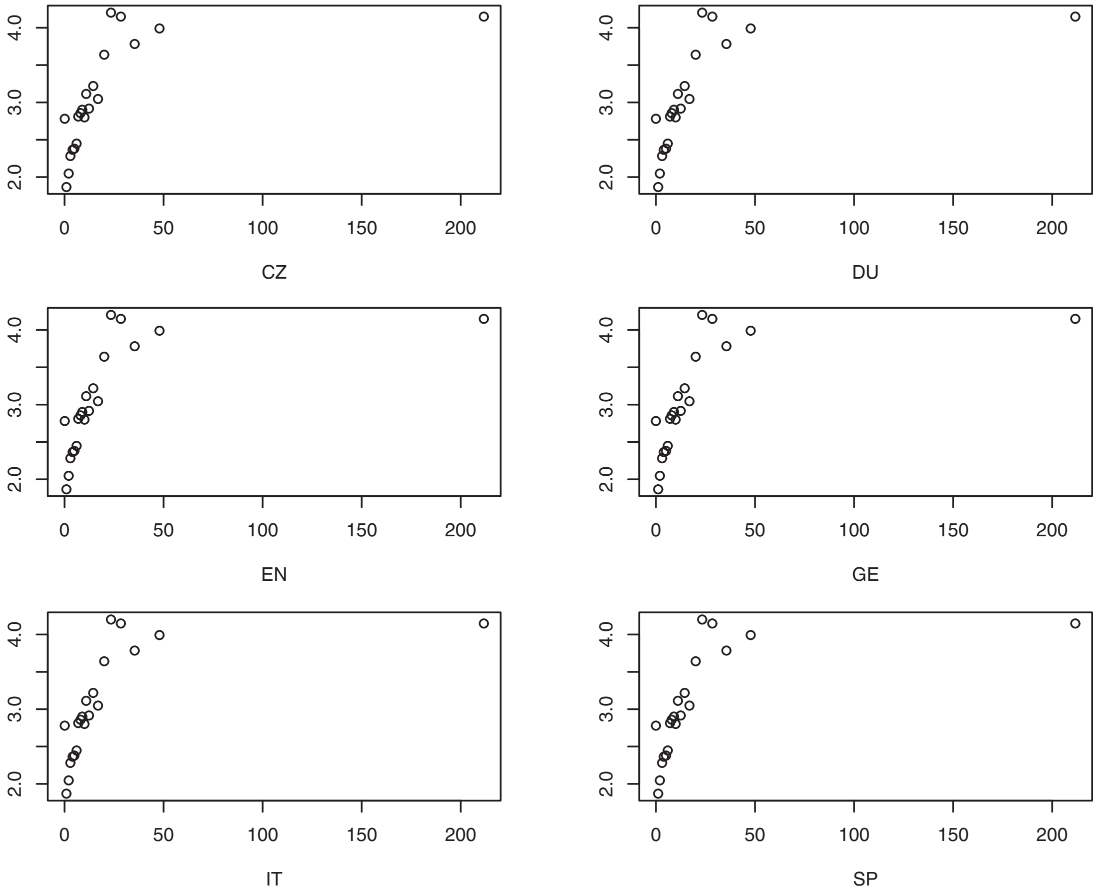

Fig. 6 Relation between the mean in-degree (x-axis) and the mean number of meanings (y-axis)

Kendall correlation coefficient (cf. Hollander and Wolfe, 1999: 382); see Section 5.

5 Results

The mean numbers of synonyms and meanings in the histogram bins tend to increase quite clearly with the increasing out-degree/in-degree. The values for the Czech Treebank[3] can be found in Table 2. Figures 3–6 present the values graphically. We use the following abbreviations for the languages: CZ—Czech, DU—Dutch, EN—English, GE—German, IT—Italian, SP—Spanish.

In addition to the figures, the association between the variables (out-degree/in-degree and the number of synonyms, out-degree/in-degree and the number of meanings) was measured by the Kendall correlation coefficient[4] (cf. Hollander and Wolfe, 1999). The values of the correlation coefficient can be found in Table 3. As the original data without any smoothing were used, the coefficients are relatively small; nevertheless, the correlation is positive and statistically significant in all cases (all *P*-values are below 0.001), which, together with the trends for binned data in Figs 1–6 corroborates the hypotheses from Section 1 (the higher out-degree/in-degree of the word, the more synonyms it has; the higher out-degree/in-degree of the word, the more meanings it has).

Given that sample sizes in Table 3 are quite large, the (very) small *P*-values themselves do not say too much—it is well known that, for large samples, virtually all null hypotheses are rejected (cf., e.g.

Table 3 Kendall correlation coefficients

	CZ	DU	EN	GE	IT	SP
Number of lemmas	11,940	6,582	14,554	6,571	5,261	8,277
Synonyms, out-degree	0.209	0.110	0.229	0.092	0.270	0.180
Meanings, out-degree	0.271	0.238	0.322	0.322	0.286	0.285
Synonyms, in-degree	0.070	0.121	0.188	0.065	0.137	0.093
Meanings, in-degree	0.141	0.212	0.295	0.236	0.179	0.211

Mačutek and Wimmer, 2013, and references therein). Therefore, Figs 1–6 provide perhaps a more reliable evidence.

6 Conclusions

The results presented in the study brought two main findings. First, we found out that one of the most fundamental syntactic network properties, the degree of the node, significantly correlates with some important semantic properties (polysemy and synonymy) of language. Moreover, the hypothesis concerning the relationships between degrees and polysemy (or synonymy) is based on a theoretical linguistic reasoning. Consequently, our findings, perhaps, advocate the usage of complex network in linguistic research and they can be viewed as a constructive response to the call for linguistic explanation of syntactic network properties (cf. Cong and Liu, 2014a, b; Ferrer i Cancho, 2014). Second, the empirical corroboration of the hypothesis concerning the relationship between the degree and polysemy can be interpreted as a deeper insight into the well-known relationship between frequency and polysemy (e.g. Zipf, 1945; Baayen, and Moscoso del Prado, 2005; Ilgen and Karaoglan, 2007). In other words, the original hypothesis concerning the impact of the word (or lemma) frequency on polysemy presupposes ‘implicitly’ that frequent words occur in more contexts and, consequently, this fact leads to an increase of polysemy. In our study, contexts have been ‘explicitly’ operationalized—the degrees express the number of syntactic contexts, in fact. Further, the use of single graphs diminishes the impact of the frequency as much as possible in studies of this kind and it allows observing the impact of a purely syntactic property (i.e. the degree of node in a syntactic complex network) of the lemma on its polysemy. To sum up, our approach experimentally proves the implicit assumption and reveals more detailed characteristics of the relationship between these important language characteristics.

Acknowledgements

This work was supported by the by the Czech Science Foundation [grant no. P406/11/0268 to R.Č.]; the Scientific Grant Agency of the Slovak Republic [grant no.VEGA 2/0047/15 to J.M.]; the Ministry of Education of the Czech Republic [LINDAT-Clarin project LM2010013 to A.H.]; the Ministry of the Interior of the Czech Republic [project VF20102014003 to A.H.]; and by the Czech Science Foundation [project P401/10/0792 to A.H.].

References

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

Baayen, H. and Moscoso del Prado, M. (2005). Semantic density and past tense formation in three Germanic languages. *Language*, 81: 666–98.

Barabási, A. (2002). *The New Science of Networks*. Cambridge, MA: Perseus Publishing.

Barabási, A. and Albert, R. (1999). Emergence of scaling in random networks, *Science*, 286: 509–12.

Brants, S., Dipper, S., Hansen, S., Lezius, W. and Smith, G. (2002). The TIGER Treebank. In Hinrichs, E. and Simov, K. (eds), *Proceedings of the First Workshop on Treebanks and Linguistic Theories*. Sozopol: Bulgarian Academy of Sciences, pp. 24–41.

Coloma, G. (2014). Towards a synergetic statistical model of language phonology. *Journal of Quantitative Linguistics*, 21: 100–22.

Cong, J. and Liu, H. (2014a). Approaching human language with complex networks. *Physics of Life Reviews*, 11: 598–618.

Cong, J. and Liu, H. (2014b). Linguistic complex networks: Rationale, application, interpretation, and directions: Reply to comments on "Approaching human language with complex networks". *Physics of Life Reviews*, 11: 644–9.

Corominas-Murtra, B., Valverde, S. and Solé. R. V. (2010). Emergence of Scale-free Syntax Networks. In Nolfi, S. and Mirolli, M. (eds), *Evolution of Communication and Language in Embodied Agents*. Berlin: Springer, pp. 83–99.

Čech, R. and Mačutek, J. (2009). Word form and lemma syntactic dependency networks in Czech: A comparative study. *Glottometrics*, 19: 85–98.

Čech, R. and Mačutek, J. (2010). On Quantitative Analysis of Valency in Czech. In Grzybek, P., Kelih, E. and Mačutek, J. (eds), *Text and Language. Structures, Functions, Interrelations, Quantitative Perspectives.* Wien: Praesens Verlag, pp. 21–9.

Čech, R., Mačutek, J. and Žabokrtský, Z. (2011). 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: 3614–23.

Čech, R., Pajas, P. and Mačutek, J. (2010). Full valency. Verb valency without distinguishing complements and adjuncts. *Journal of Quantitative Linguistics*, 17: 291–302.

de Nooy, W., Mrvar, A. and Batagelj, V. (2005). *Exploratory Social Network Analysis with Pajek*. New York: Cambridge University Press.

Ferrer i Cancho, R. (2005a). Zipf's law from a communicative phase transition. *European Physical Journal B*, 47: 449–57.

Ferrer i Cancho, R. (2005b). The Structure of Syntactic Dependency Networks: Insights from Recent Advances in Network Theory. In Altmann, G., Levickij, V. and Perebyinis, V. (eds), *Problems of Quantitative Linguistics*. Chernivtsi: Ruta, pp. 60–75.

Ferrer i Cancho, R. (2006a). When language breaks into pieces. A conflict between communication through isolated signals and language. *Biosystems*, 84: 242–53.

Ferrer i Cancho, R. (2006b). Why do syntactic links not cross? *Europhysics Letters*, 76: 1228–35.

Ferrer i Cancho, R. (2008). Some word order biases from limited brain resources. A mathematical approach. *Advances in Complex Systems*, 11: 394–14.

Ferrer i Cancho, R. (2010). Network Theory. In Hogan, P. C. (ed.), *The Cambridge Encyclopedia of the Language Sciences*. Cambridge: Cambridge University Press, pp 555–7.

Ferrer i Cancho, R. (2014). Beyond description. Comment on "Approaching human language with complex networks" by Cong & Liu. *Physics of Life Reviews*, 11(4): 621–3.

Ferrer i Cancho, R. and Solé, R. (2003). Least effort and the origins of scaling in human language. *Proceedings of the National Academy of Sciences USA*, 100: 788–91.

Ferrer i Cancho, R., Solé, R. V. and Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E*, 69: article no. 051915.

Ferrer i Cancho, R., Riordan, O. and Bollobás, B. (2005). The consequences of Zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society of London Series B*, 272: 561–5.

Gao, S., Zhang, H., Liu, H. (2014). Synergetic properties of Chinese verb valency. *Journal of Quantitative Linguistics*, 21: 1–21.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas P., Štěpánek, J., Havelka, J., Mikulová, M., Žabokrtský, Z. and Ševčíková-Razímová, M. (2006). *Prague Dependency Treebank 2.0* (CD-ROM). Philadelphia: Linguistic Data Consortium.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. New York: Routledge.

Hollander, M. and Wolfe, D. A. (1999). *Nonparametric Statistical Methods*. Hoboken: Wiley.

Horák, A., Pala, K. and Rambousek, A. (2008). The Global WordNet grid Software Design. In Tanács, A., Csendes, D., Vincze, V., Fellbaum, C. and Vossen, P. (eds), *Fourth Global WordNet Conference Proceedings*. Szeged: University of Szeged, pp. 194–9.

Hudson, R. (2007). *Language Networks. The New Word Grammar*. Oxford: Oxford University Press.

Ilgen, B. and Karaoglan, B. (2007). Investigation of Zipf's law-of-meaning on Turkish corpora. *22nd International Symposium on Computer and Information Sciences (ISCIS 2007)*, Ankara: IEEE, pp. 1–6.

Ke, J. and Yao, Y. (2008). Analysing language development from a network approach. *Journal of Quantitative Linguistics*, 15: 70–99.

Kelih, E. (2008). Modelling polysemy in different languages: A continuous approach. *Glottometrics*, 16: 46–56.

Köhler, R. (1986). *Zur linguistischen Synergetik: Struktur und Dynamik der Lexik*. Bochum: Brockmeyer.

Köhler, R. (1999). Syntactic structures. Properties and interrelations. *Journal of Quantitative Linguistics*, 6: 46–57.

Köhler, R. (2005a). Synergetic Linguistics. In Köhler, R., Altmann, G. and Piotrowski, R. G. (eds), *Quantitative Linguistics. An International Handbook*. Berlin; New York: de Gruyter, pp. 760–74.

Köhler, R. (2005b). Properties of Lexical Units and Systems. In Köhler, R., Altmann, G. and Piotrowski, R. G. (eds), *Quantitative Linguistics. An International Handbook*. Berlin; New York: de Gruyter, pp. 305–13.

Köhler, R. (2007). Quantitative Analysis of Syntactic Structures in the Framework of Synergetic Linguistics. In Mehler, A. and Köhler, R. (eds), *Aspects of Automatic Text Analysis*. Berlin; Heidelberg; New York: Springer, pp. 191–210.

Köhler, R. 2012. *Quantitative Syntax Analysis*. Berlin; Boston: de Gruyter.

Köhler, R. and Altmann, G. (1993). Begriffsdynamik und Lexikonstruktur. In Beckmann, F. and Heyer, G. (eds), *Theorie und Praxis des Lexikons*. Berlin: de Gruyter, pp. 173–90.

Liu, H. (2008). The complexity of Chinese syntactic dependency networks. *Physica A: Statistical Mechanics and Its Applications*, 387: 3048–58.

Liu, H. (2011). Quantitative properties of English verb valency. *Journal of Quantitative Linguistics*, 18: 207–33.

Liu, H. and Hu, F. (2008). What role does syntax play in a language network? *EPL*, 83: article no. 18002.

Liu, H. and Li, W. W. (2010). Language clusters based on linguistic complex networks. *Chinese Science Bulletin*, 55: 3458–65.

Liu, H. and Xu, C. (2011). Can syntactic networks indicate morphological complexity of a language? *EPL*, 93: article no. 28005.

Liu, H., Zhao, Y. and Huang, W. (2010). How do local syntactic structures influence global properties in language networks? *Glottometrics*, 20: 35–9.

Mačutek, J. and Wimmer, G. (2013). Evaluating goodness-of-fit of discrete distribution models in quantitative linguistics. *Journal of Quantitative Linguistics*, 20: 227–40.

Mel'čuk, I. (1988). *Dependency Syntax: Theory and Practice*. Albany: State University of New York Press

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. (1993). *Five Papers on WordNet (CSL Report 43)*. Princeton: Princeton University.

Montemagni, S., Barsotti, F., Battista, M., Calzolari, N., Corazzari, O., Lenci, A. and Zampolli, A. (2003). Building the Italian Syntactic-semantic Treebank. In Abeillé, A. (ed), *Treebanks: Building and Using Parsed Corpora*. Dordrecht: Kluwer, pp. 189–210.

Newman, M. E. J. (2011). *Networks. An Introduction*. Oxford: Oxford University Press.

Ninio, A. (2006). *Language and the Learning Curve: A New Theory of Syntactic Development*. Oxford: Oxford University Press.

Ninio, A. (2011). *Syntactic Development, Its Input and Output*. Oxford: Oxford University Press.

Prün, C. and Steiner, P. (2005). Quantitative Morphologie. Eigenschaften der Morphologischen Einheiten und Systeme. In Köhler, R., Altmann, G. and Piotrowski, R. G. (eds), *Quantitative Linguistics. An International Handbook*. Berlin; New York: de Gruyter, pp. 227–42.

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, 66: 605–10.

Solé, R. V. (2005). Syntax for free? *Nature*, 434: 289.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L. and Nivre, J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*. Manchester: ACL, pp. 159–77

Taulé, M., Martí, M. A. and Recasens, M. (2008). AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. Marrakech, pp. 96–101.

Tuldava, J. (1998). *Probleme und Methoden der Quantitativ-systemischen Lexikologie*. Trier: WVT.

van der Beek, L., Bouma, G., Daciuk, J., Gaustad, T., Malouf, R., van Noord, G., Prins, R. and Villada, B. (2002). *Algorithms for Linguistic Processing. NWO PIONIER Progress Report*. Groningen: Graduate School for Behavioral and Cognitive Neurosciences.

Wang, L. (2014). Synergetic studies on some properties of lexical structures in Chinese. *Journal of Quantitative Linguistics*, 21: 177–97.

Wimmer, G. and Altmann, G. (2001). Two Hypotheses on Synonymy. In Ondrejovič, S. and Považaj, M. (eds), *Lexicographica ́99*. Bratislava: Veda, pp. 218–25.

Zeman, D., Mareček, D., Popel, M., Ramasamy, L., Štěpánek, J., Žabokrtský, Z. and Hajič, J. (2012). HamleDT: To Parse or Not to Parse? In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*. Istanbul, pp. 2735–41.

Zipf, G. K. (1935). *The Psycho-biology of Language. An Introduction to Dynamic Philology*. Boston: Houghton-Mifflin.

Zipf, G. K. (1945). The meaning frequency relationship of words. *Journal of General Psychology*, 33: 251–66.

Zipf, G. K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge: Addison-Wesley.

Notes

[1] Canonical word form is the infinitive form of a verb and the nominative form of a substantive, adjective, pronoun, and so on (for example, the lemma of word forms 'go', 'goes', 'went', 'gone' is GO). In this analysis, simple lemmatization is used; differences among particular part-of-speeches are not taken into account, e.g. all forms of a word 'mean' fall to one lemma MEAN, without differentiation of verb, noun, or adjective.

[2] The terms polysemy, the number of meanings, and synonymy are used in accordance with the approach presented by the WordNet (see Section 3).

[3] On the webpage http://www.cechradek.cz/data/network_polysemy_synonymy_tables.pdf one can find values for all languages analysed in the article.

[4] The Kendall correlation is a measure of monotonous relation between two variables. It achieves its maximum value 1 if the relation 'the greater x, the greater y' holds for all pairs x,y; similarly, its minimum value −1 corresponds to the relation 'the greater x, the smaller y'.