

Word form and lemma syntactic dependency networks in Czech: a comparative study

*Radek Čech¹, Ostrava
Ján Mačutek², Graz*

Abstract. We compare several parameters of word form and lemma syntactic dependency networks in Czech. Models for degree distributions are suggested.

Keywords: syntactic dependency network, word form, lemma, degree distribution.

1. Introduction

A complex network analysis (e.g., Caldarelli 2007) has been used for several studies of human language in recent years, and now it is well known that linguistic networks display some statistical properties of complex networks (Mehler 2007). These properties have been detected at different linguistic levels – there are analyses of *semantic* networks (Sigman, Cecchi 2002; Motter et al. 2002; Holanda et al. 2004; Steyvers, Tenenbaum 2005), *syntactic* networks (Ferrer i Cancho, Solé, Köhler 2004; Ferrer i Cancho 2005), *co-occurrence* networks (Ferrer i Cancho, Solé 2001, Dorogostev, Mendes 2001), and *syllabic* networks (Soares, Corso, Lucena 2005). The fact that complex network features have been found across languages as well as across linguistic levels seems to suggest a new type of human language universals (Ferrer i Cancho 2005). The network approach was also applied to language development analysis (Ninio 2006; Ke, Yao 2008).

The same statistical properties of networks have been found in non-linguistic systems – in biology, ecology, the Internet, social systems and so on (see Caldarelli 2007). So, from the perspective of network analysis, language might be under the same rules or laws as many social and biological systems.

Recent research in syntax based on the network analysis was brought in connection with some important grammatical features of languages, for example projectivity (Ferrer i Cancho, 2006a, 2008), agrammatism (Ferrer i Cancho 2005) and the relation among Zipf's law, syntax, and communication needs (Ferrer i Cancho, Riordan, Bollobás 2005; Solé 2005; Ferrer i Cancho 2006b). All these analyses are based on quantitative characteristics of language.³

The aim of this article is to compare properties of two syntactic dependency networks based on the same language data. The first network is created by using raw word forms, the second network by using canonical word forms – lemmas. The study reveals that both networks display some properties which are typical of complex networks: small world effect, which is given by high clustering coefficient and low average path length between nodes, and

1 Department of Czech Language, University of Ostrava, Reální 5, 70103 Ostrava, Czech Republic; e-mail: radek.cech@osu.cz

2 Institut für Slawistik, Karl-Franzens Universität, Merangasse 70, 8010 Graz, Austria; e-mail: jan.macutek@uni-graz.at or jmacutek@yahoo.com

3 To our knowledge, there is one attempt to combine traditional non-statistical approach to grammar with the network analysis (Hudson 2007), but this approach is not followed in this article.

high heterogeneity; both networks are scale free. The comparison allows to investigate (1) which network properties depend specially on the fact that one uses word forms or lemmas and (2) which factors influence prospective differences between the network based on word forms (hereinafter WFN) and the network based on lemmas (hereinafter LN).

So far, most of syntactic network analyses have worked with networks in which each node represents a word form (e.g., Ferrer i Cancho, Solé, Köhler 2004; Ferrer i Cancho 2005). For instance five singular forms for seven cases of the Czech noun *kluk* (a boy) (nominative *kluk*; genitive & accusative *kluka*; dative & local *klukovi*; vocative *kluku*; instrumental *klukem*) are represented by five different nodes. On the other hand, if one uses lemmas, all word forms are represented by only one node – for example words such as *do*, *does*, *did*, *done* and *doing* are word forms of the lemma *DO* while words *kluk*, *kluka*, *klukovi*, *kluku*, *klukem* and all plural forms are word forms of the lemma *KLUK* (*BOY*).

To our knowledge, only Caldeira et al. (2006) analysed a syntactic network based on lemmas, but they used *co-occurrence* syntactic network i.e., any two words were connected if they were concomitantly in one (or more) sentence. So, the present study is the first attempt to use lemmas for syntactic *dependency* network analysis.

At first sight, the possible discrepancies between WFN and LN are caused by inflection; if there are no inflected words in a language, both networks would be equal. It opens the question about the use of the network analysis for other typological characteristics of languages. But there is no straightforward influence of inflection on network properties in the syntactic dependency networks (in the sense the more inflected words in WFN the more discrepancy between WFN and LN) because syntactic relationships could also have an important impact on the properties of LN, and consequently on the differences between WFN and LN. Hence, this article is primarily focused on the exploration of all factors that have influence on LN properties in comparison to WFN. The results show that it is possible to partly hypothesize a relationship between the typological characteristics of language and network properties (average degree, clustering coefficient).

The use of WFN is the best way for analysis of global properties of syntactic networks, primarily because it reflects syntactic dependencies of words in actual language use. The concept of lemma is artificial; nothing as lemmas actually exists in a language. However, it is obvious that using LN simplifies linguistic analysis, especially in high inflectional languages as Czech (for example the lemma of the verb *JÍST* (*EAT*) could be realized by 30 different word forms), and it is not hard to imagine why an analysis of a role of certain group of words (for instance transitive verbs) is easier and purposeful in LN than in WFN. So, if one agrees that “networks are means, not the goal” (Liu 2008) for linguists, the use of LN seems to be a reasonable way for language inquiries. But for adequate linguistic analysis of LN it is first necessary to explore the relationship between WFN and LN based on the same language data. And also prospective comparative analyses of LNs based on different languages require knowledge about factors influencing properties of LN.

It has been shown (Ferrer i Cancho, Solé, Köhler 2004; Ferrer i Cancho 2005) that syntactic dependency networks based on word forms display some properties that are typical of complex networks (i.e., a small world effect, high heterogeneity). So first, it is necessary to explore whether LN has the properties of complex networks as well as WFN, and then possible discrepancies between both networks have to be analysed.

This article is organized as follows. Properties of syntactic networks are presented in Section 2; the comparison of syntactic networks based on word forms and lemmas is given in Section 3; and the article is closed by Discussion.

2. Data and methodology

The data used in this study come from the Prague Dependency Treebank 2.0 (hereinafter PDT) (Hajič et. al. 2006). The PDT is a Czech corpus which contains a large amount of texts with interlinked morphological, syntactic, and semantic annotation. For the present purposes we used data annotated on the analytical layer, which contains 87,913 sentences and 1,503,739 word tokens. Thanks to the PDT lemmatization it has been possible to create WFN as well as LN. The PDT consists of articles from newspapers and journals.

In constructing the networks, we follow the method developed by Ferrer i Cancho et al. (2004). This approach is based on dependency grammar formalism which defines the structure of a sentence as a set of linked lexical nodes. The links represent binary relations of dependency between nodes. The direction of the links, going from the head to its modifier, (1) reflects types of syntactic relations which determine the morphological form of the subordinate word (agreement and regimen) and (2) reflects the valency properties (see Figure 1)⁴.

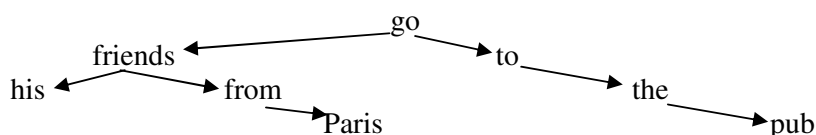


Figure 1. Direction of dependencies

The syntactic dependency network contains all words in the corpus. Two words are linked if there is a dependency relationship between them in the corpus.⁵ Thus, a global syntactic dependency network is constructed by cumulative sentence structures, and the network is an emergent property of sentence structures (Ferrer i Cancho, Solé, Köhler 2004; Ferrer i Cancho 2005). Figure 2 shows an example of a small network containing 51 vertices.

The free software Pajek 1.22 was used for the network creation and computing (<http://vlado.fmf.uni-lj.si/pub/networks/pajek/>).

3. Comparison of WFN and LN

There are several statistical measures typically used for studying complex networks. This article makes observations about the following: an average degree, degree distribution, clustering coefficient, and average path length. The results of both WFN and LN are summarized in Table 1. The large sample sizes do not allow the use of the usual statistical tests (which were designed for much smaller amount of data; with sample sizes of tens of thousands almost all null hypotheses are rejected).

⁴ Of course, there is no definite agreement about direction among linguists. We are aware that subject–predicate relationship is reverse (subject governs predicate morphology), but on the other hand, we attach importance to valency. So, our approach is a compromise.

⁵ We used a simple graph that does not reflect a frequency of connections between particular nodes (as in a multigraph). This approach was used in all previous syntactic networks analysis and we follow it because of the possibility to compare the results.

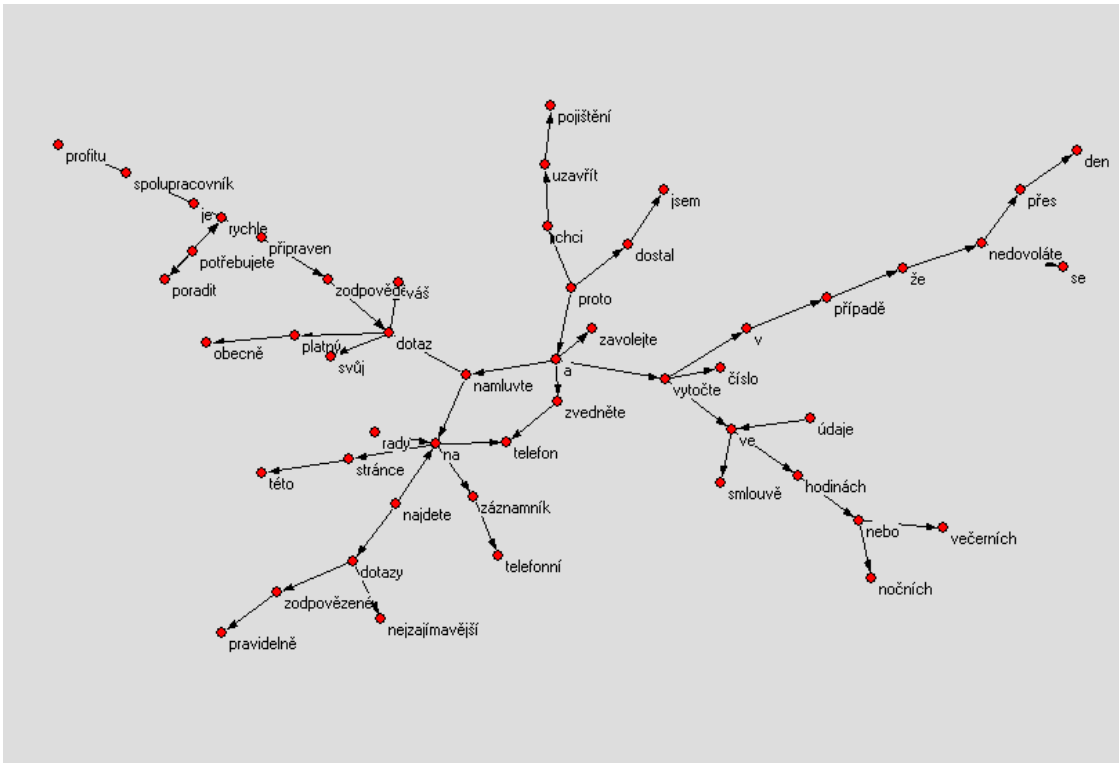


Figure 2. An example of a syntactic dependency network containing 51 vertices

Table 1
A summary of WFN and LN characteristics.

	WFN	LN
n	73989	36037
k	8.19	13.34
C	0.12	0.18
D	3.84	3.58

n = the number of vertices;

k = the average degree, it expresses the average connectivity (number of links) of all words;

C = the clustering coefficient, it is defined as the probability that two words that are neighbours of a given vertex are neighbours of each other;

D = the average minimum distance between words.

3.1 Average degree

The average degree expresses an average connectivity (number of links) of all words. It is given by

$$k = \sum_i k_i / N$$

where k is the number of undirected connections of each word i and N is the total number of words. The average degrees were measured on the undirected versions of both networks for simplicity reasons.

The comparison of WFN and LN (see Table 1) shows an average degree 1.6 times higher in LN. Furthermore, the network density, defined as

$$\delta = \frac{k}{n-1}$$

is about three times larger in LN (LN: $\delta = 0.000370185$; WFN: $\delta = 0.000110692$).

A discrepancy of k (and consequently the network density) between WFN and LN seems to be caused by inflection at first sight – if there are no inflected words in the language, the average degree of WFN and LN would be equal. But the influence of inflection upon average degree is not straightforward; syntactic relationships also have an important impact on it. Now, we will consider three possible types of syntactic connections between words with regard to inflection and the consequences for discrepancy size between WFN and LN (see Figure 3 and Table 2).

Type 1

There is only one word form (from all possible word forms) of a lemma connected to only one word form (from all possible word forms) of another lemma in WFN. This type has zero influence on discrepancy size, and it can appear in all three possible kinds of connections (with regard to inflection):

- (i) between two indeclinable words (the only possible case);
- (ii) between an indeclinable word and a declinable word which is connected to particular indeclinable word only by one word form;
- (iii) between two declinable words, each of which is connected to the other only by one word form.

Type 2

There is only one word form (from all possible word forms) of a lemma connected to more than one word form of another lemma in WFN. This type causes a *higher average degree of WFN* and it can appear:

- (i) between an indeclinable word and a declinable word which is connected to a particular indeclinable word by more than one word form;
- (ii) between two declinable words, one of which is realized only by one word form (from all possible word forms) of a lemma and the other is connected to it by more than one word form.

Type 3

There is more than one word form of a lemma; each is connected to only one word form of different lemmas in WFN. This type causes a *higher average degree of LN* and it appears:

- (i) between a declinable word which is realized by more than one word form and indeclinable words which are connected to only one word form of a declinable word (no indeclinable word can be connected to more than one word form of declinable word);
- (ii) between a declinable word which is realized by more than one word form and

other declinable words; each word form of former declinable word is connected to a different word which is realized only by one word form (from all possible word forms) of a lemma.

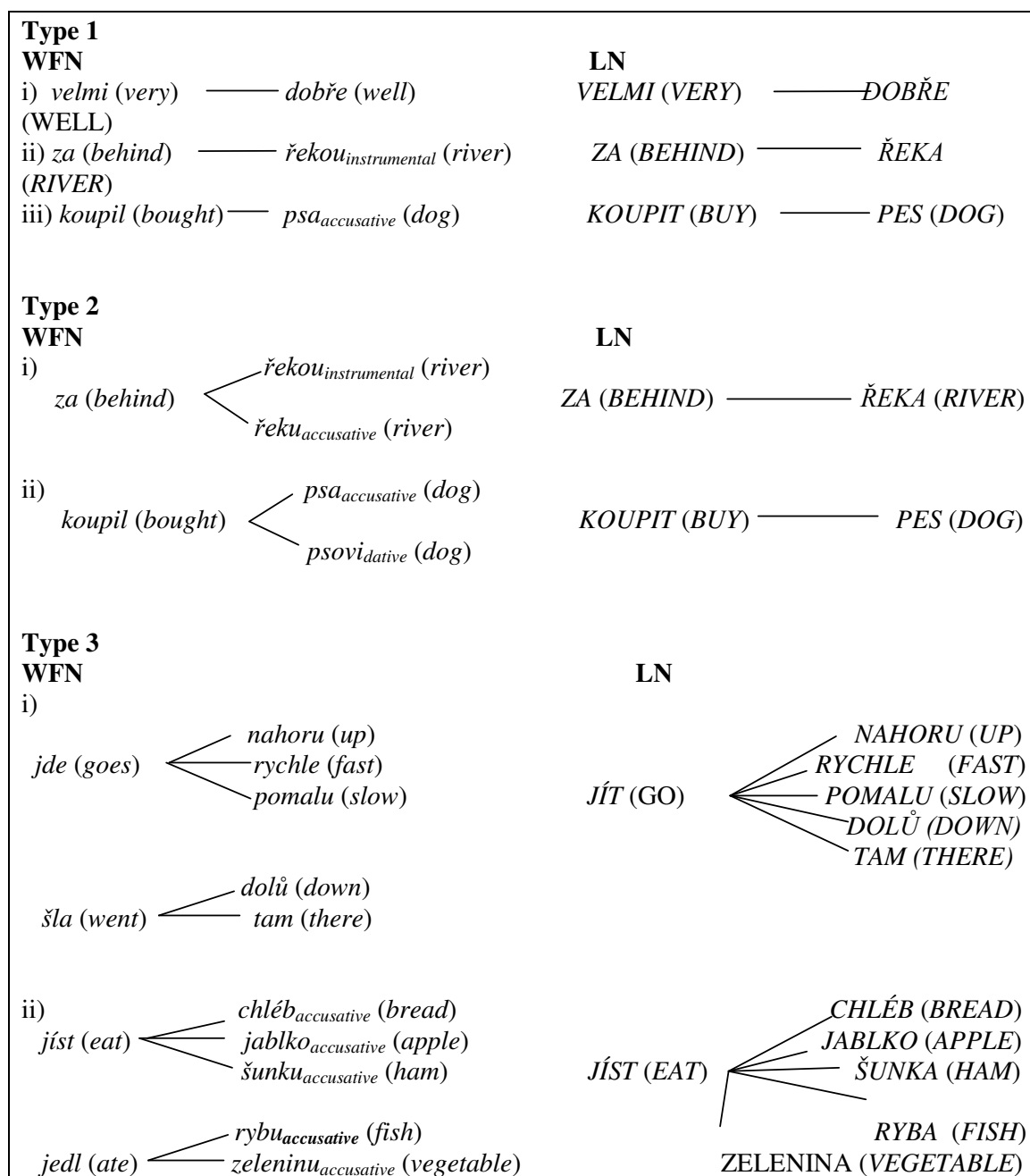


Figure 3. Three types of syntactic connections between words with regard to inflection. Each type is an illustrative example of a small language network. It shows how inflection and syntactic relationships determine a discrepancy size between WFN and LN. For values of k for each type see Table 2.

Table 2
Average degree values for each type of syntactic connections from Figure 3.

	WFN	LN
k_{type1}	1	1
k_{type2}	1.33	1
k_{type3}	1.42	1.67

The comparison of Types 1–3 shows that a discrepancy size is given by both inflectional richness (a number of particular word forms) and syntactic relationships. Moreover, an actual value of k is caused simultaneously by the grammar (if there is only one possible word form which could be connected to another word form e.g., connections between indeclinable words or a connection of an indefinite article with just a singular noun in English) and by the language usage (it determines how many word forms of a declinable word are realized in Types 2 and 3, and how many different indeclinable words are connected to a particular word form of a declinable lemma in Type 3). Consequently, a mere presence of inflection in the language does not necessarily entail higher values of average degree in WFN or LN and, theoretically, it is possible that the language with rich inflection will have the same average degree as the language with absolutely no inflection. So, the comparison of the average degree of WFN and LN cannot be used for typological characterisation of languages.

As for an observed discrepancy between WFN and LN (Table 1), the higher value of LN means that rich inflection is coupled with high lexical variability in actual language use. In other words, there is a strong tendency of a particular word form of a lemma to connect with many other lemmas through one word form of each of these lemmas (see Type 3). In Czech it is typical for instance of transitive verbs, as well as other types of verb, to have many raw word forms and which are connected to a host of different objects (accusative nouns). So again, the discrepancy is simultaneously a product of grammar and language usage.

3.2 Degree distribution

The degree distribution, $P(k)$, describes the number of nodes (e.g., word forms or lemmas) with a connectivity k . For most of large complex networks highly heterogeneous distribution is typical: the probability $P(k)$ that a randomly chosen node in the network interacts with k other nodes decays as a power law, following $P(k) \sim k^{-\gamma}$ (Barabási, Albert 1999). Moreover, as Ferrer i Cancho (2005) shows, the distribution of word degrees could be a consequence of Zipf's law for word frequencies and the distribution, following the power law, should be a universal property of language.

However, in general our frequencies are not decreasing; in two cases we have $f(0) < f(1) > f(2) > f(3) > \dots$. Hence we suggest the (shifted) gamma function as a model, namely,

$$f(x) = a(x+1)^b e^{-cx}.$$

We replaced the parameter a with the frequency $f(0)$. The obtained fit is very good (see Table 3) and, moreover, we do not reject the previous model; we generalize it (obviously, for $c = 0$ one obtains the power law).

Table 3
Fitting the gamma function to the degree distributions

	a	b	c	R^2
IN-distribution (WFN)	4665	11.936	6.100	0.9828
OUT-distribution (WFN)	28205	-0.440	0.322	0.9993
IN-distribution (LN)	1531	12.099	5.982	0.9653
OUT-distribution (LN)	13406	-0.211	0.444	0.9982

As can be seen, there are obvious differences between parameter values for nodes representing subordinate words and governing words.

We do not present particular frequencies in the degree distributions – that would mean inserting four tables with approximately 250 lines.⁶ The data can be sent upon request. A discrepancy between the previous model (i.e., power law distribution) and the result we obtained requires an explanation. As Newman (2006) shows, only “[f]ew real-world distributions follow a power law over their entire range, and in particular not for smaller values of the variable being measured. (...) [F]or any positive value of the exponent α the function $p(x) = Cx^{-\alpha}$ diverges as $x \rightarrow 0$. In reality therefore, the distribution must deviate from the power-law form below some minimum value x_{\min} .” In many analyses a distribution altogether below x_{\min} is cut off and “one often hears it said that the distribution of such-and-such a quantity “has a power-law tail.” This precisely happened in our case. If we cut off nodes with frequency $f(0)$, the all distributions follow the power law, so it seems reasonable to consider the power law as a universal language property for distributions for $x_{\min} = 1$.

However, we consider the cutting off of distribution below some x_{\min} improper in our analysis because the nodes with the frequency $f(0)$ are linguistically important. They represent (1) word forms/lemmas which occur purely as terminal nodes of sentence structure, in the case of nodes which represent modifiers or (2) word forms/lemmas which occur purely as the highest elements in sentence structure (e.g., predicative verbs or various types of words in elliptical expressions), in the case of nodes which represent heads. That is why we propose the new model for degree distribution (see above).

Figure 4 presents cumulative in-degree and out-degree distributions in WFN and LN – the proportions of words whose input or output degree is k or more are shown. All graphs display a highly heterogeneous distribution; we can see that 90% of the words have less than 10 connections, whereas only 0.1% of all words have more than 100 connections. The distribution of words in LN follows the power law as well ($P(k) = (k+1)^{-b}$, see Table 4; we note that because of zero degrees the function had to be shifted to the left). Consequently, if the lemma network follows the power law (for $x_{\min} = 1$), it means that the concept of lemma, despite of its artificiality, “obeys” the same language universal rule (Ferrer i Cancho 2005) and LN can be properly used for linguistic analysis.

⁶ The data are available at the web page www.cechradek.ic.cz/@files/WFN_LN_distributions.xls

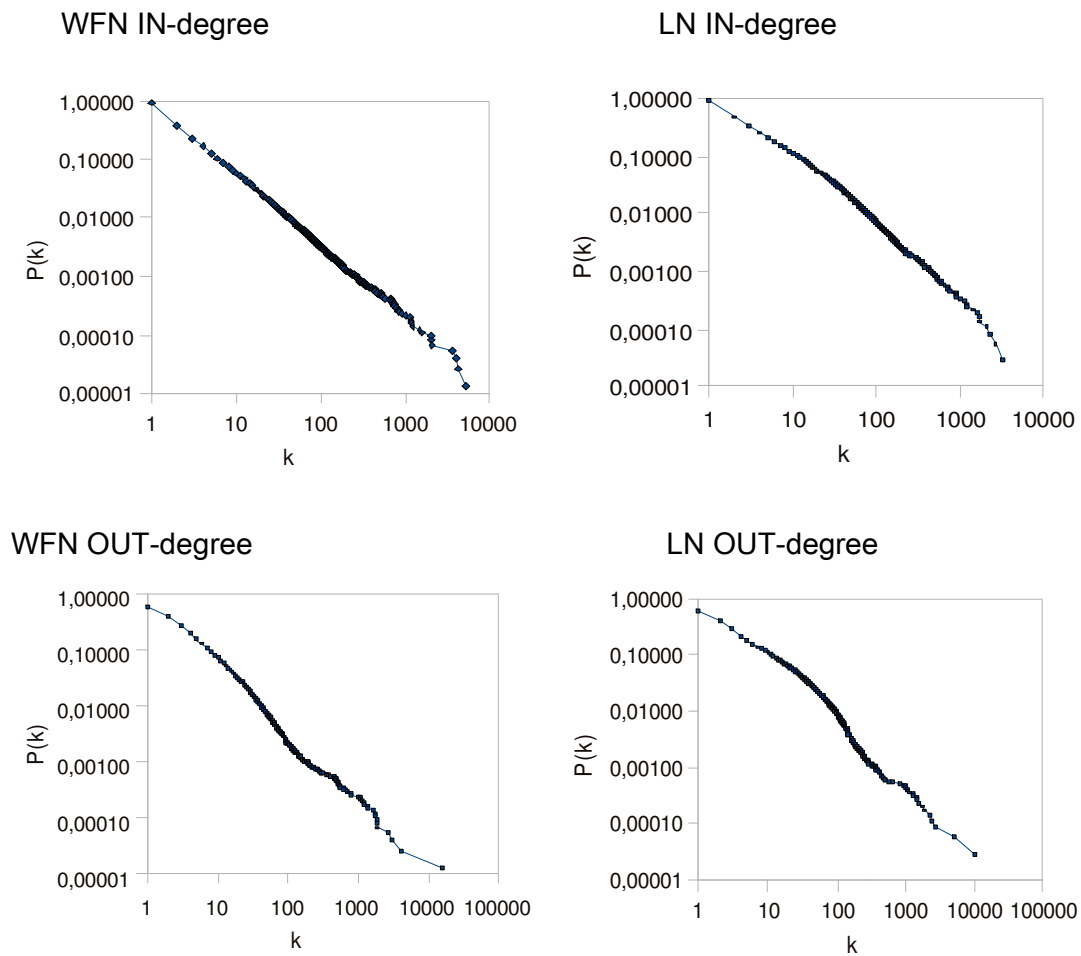


Figure 4. Cumulative degree distributions of network

Table 4

Fitting the power law function to the cumulative degree distributions.

	b	R^2
IN-distribution (WFN)	1.0170	0.8944
OUT-distribution (WFN)	1.0286	0.9766
IN-distribution (LN)	0.8721	0.9161
OUT-distribution (LN)	0.9211	0.9898

In this case one does not see any obvious differences in parameter values. We are aware of the vagueness of this statement; however, as we already mentioned, the huge amount of data is the reason why the usual statistical approaches cannot be applied.

3.3 Clustering coefficient and average path length

The clustering coefficient, C , expresses the probability that two nodes that are both neighbors of the same third node will be neighbors of one another (Newman 2001). The value of the clustering coefficient for a whole network is given by the average over all nodes and it indicates interconnectivity of a complex network. Because of problems with a clustering coefficient measurement in directed networks (Caldarelli 2007), values of C were measured on the undirected versions of both networks, WFN and LN.

Table 1 shows that there is a higher clustering coefficient in LN than in WFN. As in average degree (Section 3.1), a discrepancy is caused by the presence of declinable words in the language and by syntactic relationships. Next we will consider three types of syntactic relationship with regard to inflection and consequences for differences of C between WFN and LN.

Type 1

There is an indeclinable word connected to only indeclinable words. Obviously, interconnectivity among indeclinable words has zero influence on differences between clustering coefficients of WFN and LN, because all syntactic relationships are the same in WFN and LN thanks to the absence of inflection.

Type 2

There is an indeclinable word connected to a declinable word or words. Theoretically, all three possible cases with regard to discrepancies of C between WFN and LN should appear:

- (i) higher C of LN,
- (ii) equal C of WFN and LN,
- (iii) higher C of WFN.

But if we take account of grammar properties of the observed language, we would be able to hypothesize, at least partly, what kind of a clustering coefficient discrepancy (i.e., equal C of WFN and LN, higher C of WFN or higher C of LN) between WFN and LN should be more often detected than the others.

For (i) let us start with the clustering coefficient higher in LN than in WFN. For illustration (see Figure 5), assume the connections between the preposition, e.g., $o_{(locative)}$ (*about*)⁷, and two nouns, e.g., *stůl* (*table*) and *noha* (*foot*). The preposition $o_{(locative)}$ (*about*) governs the grammatical case of dependent nouns, but there is no possible syntactic relationship between the two nouns in the locative case in Czech. Consequently, the clustering coefficient of preposition $o_{(locative)}$ (*about*) equals zero in WFN. Now, assume the same example in LN: all word forms of declinable words fall to one lemma, so if there is a possible syntactic relationship between *stůl* (*table*) and *noha* (*foot*) in other grammatical cases i.e., in an attributive connection *noha_{nominative} stolu_{genitive}* (*table foot*), there is a potentiality that lemmas *O* (*ABOUT*), *STŮL* (*TABLE*), and *NOHA* (*FOOT*) can be interconnected in LN. So, if there is an actual connection (it depends on the usage) between these nouns, the clustering coefficient is higher in LN than in WFN (Figure 5).

⁷ Preposition o can also be used with accusative noun in Czech; in presented example we consider only locative coligation.

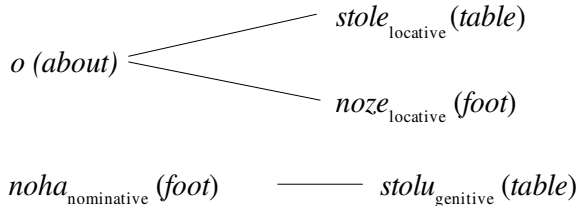
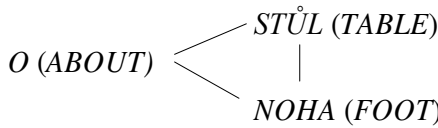
WFN	LN
	
$C_{o (about)} = 0$	$C_{O (ABOUT)} = 1$

Figure 5. An example of small networks, WFN and LN, based on five raw word forms. The connection between declinable words *noha_{nominative} stolu_{genitive} (foot table)* in WFN causes higher C of lemma *O (ABOUT)* in LN

For (ii), now, consider an instance when clustering coefficients in WFN and LN are equal. Again for illustration (see Figure 6), assume the connections between genitive preposition *bez (without)* and two nouns, *stůl (table)* and *noha (foot)*. Contrary to nouns in locative, genitive nouns could connect each other e.g., *nohy_{genitive} stolu_{genitive} (table foot)*. So if these nouns are actually connected in WFN, the clustering coefficient is equal in WFN and LN.

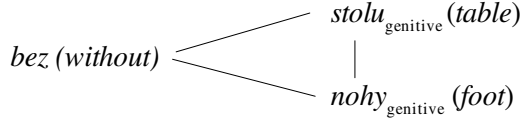
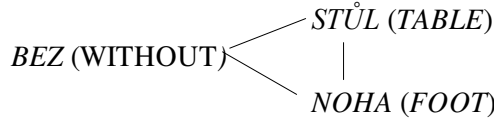
NW	NL
	
$C_{bez (without)} = 1$	$C_{BEZ (WITHOUT)} = 1$

Figure 6. An example of small networks based on three raw word forms. The connection between dependent genitive nouns, *nohy_{genitive} stolu_{genitive} (table foot)*, causes equality of C of preposition *bez (without)* in WFN and LN.

Of course, an equality of clustering coefficients in WFN and LN would be also given, if there is no actual connection between declinable words in WFN and LN – a value of the clustering coefficient of an indeclinable word equals zero in both networks.

Finally, for (iii), the higher value of the clustering coefficient in WFN would be possible, if there is a connection between indeclinable word and at least two word forms of the same lemma of declinable word, and moreover between these same word forms in WFN: e.g., all instances as *bez písň_{genitive sg.} (without song)*, *bez písní_{genitive pl.} (without songs)*, and *písň_{genitive sg.} písni_{genitive pl.} (song of songs)* have to be present in WFN for the higher value of the clustering coefficient in WFN (see Figure 7).

NW	NL
<p><i>bez (without)</i> is connected to <i>písň_{genitive sg.} (table)</i> and <i>písni_{genitive pl.} (foot)</i>.</p>	<p><i>BEZ (WITHOUT)</i> — <i>PÍSEŇ (SONG)</i></p>
$C_{bez (without)} = 1$	$C_{BEZ (WITHOUT)} = 0$

Figure 7. An example of small networks based on three raw word forms. The connection between two dependent word forms of the same lemma, *PÍSEŇ (SONG)*, causes higher C of preposition *bez (without)* in WFN.

Because case (i) is more typical than case (ii), and case (iii) is very rare in Czech we can expect that words with no inflection should have more often a higher clustering coefficient in LN than in WFN. So, with regard to syntactic connections between indeclinable and declinable words the higher C in LN (Table 1) is not surprising.

Type 3

Type 3 represents syntactic relationships between declinable words. All the three possible cases with regard to discrepancies of C between WFN and LN should appear: equal C of WFN and LN, higher C of WFN or higher C of LN. Contrary to Type 2, it is primarily the language usage that determines how many word forms of the lemma are realized and which are connected to each other in WFN, so the influence of grammar is much weaker than in Type 2. Consequently, it is practically impossible to hypothesize which kind of clustering coefficient discrepancy (i.e., equal C of WFN and LN, higher C of WFN or higher C of LN) between WFN and LN should be more often detected than the others.

Moreover, the clustering coefficient is crucial in the characterization of complex networks: a high C in real networks, in comparison of C in Erdős-Rényi random networks, indicates the small world structure of real networks (Watts, Strogatz, 1998). Table 3 shows that $C_{WFN} \gg C_{random}$ and $C_{LN} \gg C_{random}$. The small world phenomenon means that despite the large number of network elements (e.g., word forms or lemmas) the distance between them is strikingly small. The value of this distance is called average path length, D , and it expresses the shortest distance (it is defined by a number of links) between any randomly chosen pair of nodes of a network. Values of D in both WFN and LN are close to D_{random} (see Table 5) which expresses the value of D for Erdős-Rényi random network. It exhibits, with considerably high clustering formation that both networks are really small worlds (Watts, Strogatz, 1998).

Table 5

Clustering coefficients and average path lengths of real and random networks.

	C	D
WFN	0.12	3.84
WFN _{random}	0.0009	5.57
LN	0.18	3.58
LN _{random}	0.004	4.35

4. Discussion and future work

We presented a study which compares two syntactic dependency networks based on the same language data; in one network each node represents a raw word form (WFN), in the other network each node represents a basic word form, lemma (LN). The analysis shows discrepancies between WFN and LN in observed values. These discrepancies are caused by the inflectional characteristics of the Czech grammar, syntactic relationships, and by language usage.

As regards *average degree* we can partly hypothesize a relationship between the typological character of language and the average degree discrepancy between WFN and LN:

- networks based on languages with no inflection (as a highly isolating language) will have zero discrepancy,
- networks based on languages with low inflection (as English) will have zero discrepancy or higher average degree of WFN,
- for networks based on highly inflectional languages it is not possible to make theoretical hypotheses; all the three potential kinds of discrepancy could appear because the discrepancy value is significantly influenced by language usage. The present study shows why one could expect a higher average degree in LN (Section 3.1), but only further observations of networks based on different highly inflectional languages could reveal typical characteristics of WFN and LN with regard to average degree.

The kind of *clustering coefficient* differences between WFN and LN could be partly hypothesized (for words with no inflection) by closer grammar observation, as we presented in Section 3.3. Not only grammar or typological characteristics play a crucial role for clustering coefficient discrepancy, language usage also notably influences the clustering coefficient as well the average degree.

The findings presented in the article are important for potential matching of individual lemma networks based on different languages; for appropriate comparison of these networks one has to take into account at least typological characteristics of languages. Because of the importance of language usage one should also consider the possible influence of text types, but only further observations of networks based on different registers could reveal a real impact of text types on both the average degree and clustering coefficient.

Acknowledgements

R. Čech and J. Mačutek were supported by GAČR (Czech Science Foundation) No. 405/08/P157 – Components of transitivity analysis of Czech sentences (emergent grammar approach) and by the Lise Meitner Stipendium (FWF, Austria), respectively.

References

- Barabási, A.L., Albert, R. (1999). Emergence of Scaling Random Networks. *Science* 286, 509-512.
- Caldarelli, G. (2007). *Scale-Free Networks: Complex Webs in Nature and Technology*. Oxford: Oxford University Press.
- Caldeira, S.M.G., Lobão, T.P., Andrade, R.F.S., Neme, A., Miranda, J.G.V. (2006). The network of concepts in written texts. *European Physical Journal, B* 49, 523-529.
- Dorogovtsev, S.N., Mendes, J.F. (2001). Language as an evolving word web. *Proceedings of the Royal Society of London B* 268, 2603–2606.
- Ferrer i Cancho, R. (2005). The structure of syntactic dependency networks: insight from recent advances in network theory. In: Altmann, G., Levickij, V.V., Perebyinis, V.

- (eds.), *Problems of Quantitative Linguistics: 60-75*. Chernivtsi: Ruta.
- Ferrer i Cancho, R.** (2006a). Why do syntactic links not cross? *Europhysics Letters* 76, 1228-1235.
- Ferrer i Cancho, R.** (2006b). When language breaks into pieces. A conflict between communication through isolated signals and language. *Biosystems* 84, 242–253.
- Ferrer i Cancho, R.** (2008). Some word order biases from limited brain resources. A mathematical approach. *Advances in Complex Systems* 11 (3), 394–414.
- Ferrer i Cancho, R., Riordan, O., Bollobás, B.** (2005). The consequences of Zipf's law for syntax and symbolic reference. *Proceedings of the Royal Society of London Series B* 272, 561–565.
- Ferrer i Cancho, R., Solé, R.V.** (2001). The small-world of human language. *Proceedings of the Royal Society of London B* 268, 2261–2265.
- Ferrer i Cancho, R., Solé, R.V., Köhler, R.** (2004). Patterns in syntactic dependency networks. *Physical Review E* 69, 051915.
- Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M.** (2006). *Prague Dependency Treebank 2.0*. Philadelphia: Linguistic Data Consortium.
- Holanda, A.J., Pisa, I.T., Kinouchi, O., Martinez, A.S., Ruiz, E.E.S.** (2004). Thesaurus as a complex network. *Physica A* 344, 530–536.
- Hudson, R.** (2007). *Language Networks: The New Word Grammar*. Oxford: Oxford University Press.
- Ke, J., Yao, Y.** (2008). Analysing Language Development from a Network Approach. *Journal of Quantitative Linguistics* 15(1), 70–99.
- Liu, H.** (2008). The complexity of Chinese syntactic dependency networks. *Physica A* 387, 3048–3058.
- Mehler, A.** (2007). Large Text Networks as an Object of Corpus Linguistic Studies. In: Lüdeling, A., Kytö, M. (eds.), *Corpus Linguistic. An International Handbook: 328-382*. Berlin: Mouton de Gruyter.
- Motter, A.E., de Moura, A.P.S., Lai, Y.-Ch., Dasgupta, P.** (2002). Topology of the conceptual network of language. *Physical Review E* 65, 065102(R).
- Newman, M.E.J.** (2001). Clustering and preferential attachment in growing networks. *Physical Review E* 64, 025102(R).
- Newman, M.E.J.** (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46, 323–351.
- Ninio, A.** (2006). *Language and the learning curve. A new theory of syntactic development*. Oxford: Oxford University Press.
- Sigman M., Cecchi, G.A.** (2002). Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences of the United States of America* 99(3), 1742-1747.
- Soares, M., Corso, G., Lucena, L.S.** (2005). The network of syllables in Portuguese. *Physica A*, 355 (2-4), 678–684.
- Solé, R.V.** (2005). Syntax for free? *Nature* 434, 289.
- Steyvers, M., Tenenbaum, J.B.** (2005). The Large Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science* 29(1), 41–78.
- Watts, D.J., Strogatz, S.H.** (1998). Collective dynamics of 'small-world' networks. *Nature* 393, 440–442.