Properties of the network of semantic relations in the Russian language based on the RuWordNet data

Vladimir V. Bochkarev, Valery D. Solovyev
Kazan Federal University, Kazan, Russia

vbochkarev@mail.ru, maki.solovyev@mail.ru

Abstract. It was previously shown that the WordNet thesaurus has a small-world structure. We obtained a similar result for RuWordNet, a recently created thesaurus of the Russian language, and determined the main characteristics of the network of semantic relations in the Russian language. They are the average length of the path between the vertices, the maximum path length between the vertices, the cluster coefficient and the assortativity coefficient. The distribution of the vertices degrees was also considered. We compared the values of the structural parameters of the networks for WordNet and RuWordNet. A significant difference between RuWordNet and WordNet is that RuWordNet has many semantic relations of the 'POS-synonymy' type (synonymy between words of different parts of speech) that are hardly found in WordNet. The values of structural constants were determined for RuWordNet both regarding this type of relations and without taking into account these relations, which allows one to estimate the influence of relations between parts of speech on the structure of the words network.

1. Introduction
One of the most important digital linguistic resources of the English language, which has numerous applications, is a well-known thesaurus (lexical ontology) WordNet (https://wordnet.princeton.edu/) [1]. The RuWordNet thesaurus (https://ruwordnet.ru/ru/) is its analogue for the Russian language. RuWordNet was created semi-automatically using corpora of news and legislative texts. The authors of this thesaurus [2] consider that news corpora provide a lot of important information concerning current social events and life. Like Wordnet, RuWordNet contains words grouped into synsets (synonimic sets). These synsets are connected to other synsets by means of semantic relations. RuWordNet contains 110.5 thousand of words and word combinations belonging to 49798 synsets. The words in the synsets refer to three main parts of speech: nouns (29297 synsets), verbs (7636 synsets) and adjectives (12865).

The following semantic relations are established between the synsets in RuWordNet: hyponym-hypernym, instance-class, part-whole, antonymy (for adjectives), cause, entailment, and domain. The first four types of semantic relations are typical of the thesauri and are also presented in WordNet. Synsets that refer to different parts of speech but express one and the same sense are connected by means of POS synonymy. Such relations can also be found in WordNet. However, their amount is small. The word combination is connected to the words that it contains by means of the relation "consists of". Cognate words are connected by means of the relation "cognate words". More information about the principles of the thesaurus constructing can be found in [2, 3]. The article studies the structural properties of RuWordNet compared to WordNet.
Many networks have been studied from a structural point of view. It is showed in [4] that semantic networks, such as WordNet, the Roget’s Thesaurus and associative networks have many common features such as a small-world structure and scale-free organization. The latter features were studied in [5, 6, 7], where it was noted that many networks (including WWW, networks of scientific cooperation, etc.) have these properties. Probably, this reflects the fundamental properties of the structures that ensure efficient storage and search for large amounts of information [4].

Two promising areas for further research are mentioned in [4]: comparative studies for different languages and studies of the structures of narrow groups of words, for example, certain parts of speech. The authors of this article make the first attempt to answer these questions for the Russian language.

2. Data analysis

As it was mentioned above, there are 49798 synsets in the RuWordNet thesaurus. It contains more than 110 thousand of Russian words and word combinations. There are 57865 words in it. This paper considers networks of semantic relations for words and synsets. Therefore, word combinations were excluded from the analysis. The network of semantic relations is regarded as undirected graph.

First, let us consider the distribution of the degrees of the vertices of the network of semantic relations. Figure 1, A shows distribution of the degrees as a function of rank $k$ for the networks of synsets (solid line) and words (dash-dotted line). It can be seen that only the curve for the network of words can be described by a power law with some degree of conditionality.

Let us consider in more detail the deviations of empirical curves from the expected power dependence. According to [8, 9], let us represent the distribution of the degrees of the vertices $P(k)$ in the locally power form:

$$P(k) = A(k) k^{-\gamma(k)}$$

(1)

Here, it is assumed that the amplitude $A(k)$ and the local power exponent $\gamma(k)$ are relatively slowly varying functions. To estimate the local power exponent near the target rank value $k_0$, we select all vertices whose rank differs from $k_0$ by no more than a given factor:

$$k_0(1 + \varepsilon) \leq k \leq k_0(1 + \varepsilon)$$

(2)
Here, \(\varepsilon\) is a small number (in this work, the values of 0.25-0.4 were used). The dependence is approximated by a power function for the set of vertices chosen in accordance with (2). The method is detailedly described in [9].

The calculation results are shown in Figure 1, B. As for the network of synsets, the exponent \(\gamma(k)\) varies over the entire range of ranks. As for the network of words, one can see a section in the range of \(k\) values (approximately, from 75 to 4000) in which the indicator \(\gamma(k)\) varies slightly. Thus, the dependence power approximation \(P(k)\) can be considered satisfactory for this range. At that, the power exponent is close to 0.5. If the approximation \(P(k)\) is performed by a power dependence in the specified range of 75–4000 using the criterion of the relative MSE, the obtained value of the exponent is 0.4956.

This value is much lower than that obtained in most analysed works on real networks [6, 7]. In [4], the values of 3.11-3.19 are obtained for the networks of semantic relations of the English language using the WordNet and Roget’s Thesaurus data.

Besides the degree distribution, other parameters of the semantic relation networks were calculated. They are the maximum length of the shortest path between the vertices (the diameter of the network), the average length of the shortest path, the cluster coefficient and the assortativity coefficient. The calculation results are shown in Table 1. The table shows data on networks of synsets and words (both considering and excluding POS-synonymy), as well as data for networks of words referring only to one part of speech (adjective, noun and verb).

**Table 1.** Parameters of the semantic relation networks for the Russian language and the estimation obtained in [4] for the English language using the WordNet and Roget’s thesaurus.

| Parameters                      | Synsets | Words | Words excluding POS-synonymy | Adjectives | Nouns | Verbs | English words (WordNet) | English words (Roget’s Thesaurus) |
|---------------------------------|---------|-------|-----------------------------|------------|-------|-------|------------------------|----------------------------------|
| Diameter of the network         | 15      | 12    | 24                          | 14         | 17    | 24    | 27                     | 10                               |
| Average length of the shortest path | 5.617  | 4.360 | 5.956                       | 5.300      | 5.476 | 8.686 | 10.56                  | 5.60                             |
| Cluster coefficient             | 0.305   | 0.564 | 0.755                       | 0.595      | 0.769 | 0.887 | 0.027                  | 0.875                            |
| Assortativity coefficient       | -0.113  | -0.072| 0.094                       | -0.125     | -0.059| 0.632 | -                      | -                                |

According to the WordNet thesaurus data, the network diameter and the average length of the shortest path for the network of words of the Russian language is more than 2 times smaller than those of the English language. As for the value obtained using the Roget’s Thesaurus, it is quite close to ours (the network diameter is 12 in Russian and 10 in English, the average length of the shortest path is 4.36 in Russian and 5.6 in English).

The cluster coefficient behaves in a similar way. Its value is relatively high for Russian words. According to the WordNet thesaurus data, the cluster coefficient value is low for English. The value obtained using Roget’s are high and equals 0.875. The value obtained in this work is 0.564.

The behaviour of the local cluster coefficient is also interesting. Figure 2 shows the dependence of the local cluster coefficient on the vertex degree. It is seen that the local cluster coefficient does not depend on the vertex degree for small and moderate values of the degree of the vertex. However, if the degree values are high, the cluster coefficient decreases along with degree increase. Having divided the values of the degree of the vertices into ranges in accordance with expression (2), we calculate the conditional average values of the local cluster coefficient depending on the degree of the vertices.
(solid line). It is seen that the decrease nature is close to a power law, as, for example, the hierarchical network model predicted [10, 11]. It is a difficult task to determine the power exponent, since, firstly, the result will depend on the choice of the range for approximation, and secondly, there are very few points in the area of large values. Approximating the dependence for the vertices with a degree of more than 200 (there are 248 of such vertices in the network), the value of the power exponent (which equals -1.576) is obtained (the result of the approximation is shown by the dash-dotted line). If the dependence is approximated for the vertices with a degree of more than 400 (there are 70 of such vertices), the value which slightly differs from -1 is obtained (exactly this value is predicted in [12] for deterministic hierarchical networks).

Figure 2. Dependence of the local cluster coefficient on the vertex degree for the network of words of the Russian language

Comparing the results for the synsets and words, it is seen that in the second case the diameter of the network and the average length of the shortest path are smaller, and the cluster coefficient is higher for words than for synsets. Both effects are expected. Indeed, a word can be included in several synsets. This increases the connectivity of the network and reduces its diameter and the average shortest path between the vertices. At the same time, if the word synonymy is considered the cluster coefficient increases.

Let us consider the effect of POS-synonymy on the network parameters. If the POS-synonymy relation is excluded, the network splits into 3 connectivity components, corresponding to the parts of speech presented in the thesaurus. The average length of the shortest path between the connected vertices of the network and the cluster coefficient in this case can be obtained as weighted averages of the values of the corresponding quantities for the three subnets. Interestingly, the average length of the shortest path for the network of words of each part of speech separately is significantly higher than that for the network of all words. In other words, the shortest path in the network can be constructed for the large part of words using words belonging to other part of speech. Accordingly, the average length of the shortest path without considering the POS-synonymy relation is significantly higher than it is considered (5.96 vs. 4.36). The cluster coefficient shows the opposite behaviour. It is higher for each of the subnets separately than for the network of all words. Considering the POS-synonymy relation, we get a significantly higher value than without taking it into account (0.755 vs. 0.564).

The values of the assortativity coefficient are close to zero for all networks, except for the network of verbs. In the latter case, a sufficiently high positive value of the coefficient is observed. It should be noted that the network of verbs has some other distinguished features.

3. Conclusion
The obtained results show that the network of semantic relations (built according to the RuWordNet thesaurus) has the property of a small-world structure and a high level of clustering. If the obtained data are compared with the WordNet thesaurus data, one can find significant differences in their
network structure. However, most of the parameters show a good agreement with the data on the English language obtained using the Roget’s Thesaurus. The most significant difference is observed for the power exponent distribution of the vertex degrees (0.5 vs. 3.11-3.19 for the English language). Probably, it can be said that RuWordNet provides a denser network with many connections at a qualitative level. It is unlikely that this is a feature of the Russian language. It is rather a result of the chosen strategy for constructing a thesaurus. This can possibly be due to three additional types of relations between the synsets in the RuWordNet which were mentioned above. Another interesting result is that the asymptotic dependence of the local cluster coefficient on the degree of the network vertex is detected. This dependence is close to a power law, and the power exponent is apparently close to the value of -1, which is predicted by the hierarchical network model.

Acknowledgements
This research was financially supported by the Russian Government Program of Competitive Growth of Kazan Federal University, state assignment of Ministry of Education and Science, grant agreement № 2.8303.2017/8.9 and by RFBR, grants № 17-29-09163, 18-00-01238.

References
[1] Miller G A 1995 Communications of the ACM 38 (11) 39
[2] Loukachevitch N V, Lashevich G, Gerasimova A A, Ivanov V V and Dobrov B V 2016 Proc. of Conf. on Computational linguistics and Intellectual technologies Dialog-2016 (Moskow: Russian State University for the Humanities) pp 405-415
[3] Loukachevitch N and Lashevich G 2017 Proc. of the AINL FRUCT 2016 Conf. pp 66-71
[4] Steyvers M and Tenenbaum J B 2005 Cogn. Sci. 29(1) 41
[5] Watts D J and Strogatz S 1998 Nature 393 (6684) 440
[6] Barabási A L and Albert R 1999 Science 286 509
[7] Dorogovtsev S N and Mendes J F F 2003 Evolution of Networks: From Biological Networks to the Internet and WWW (Oxford, USA: Oxford University Press)
[8] Bochkarev V V, Lerner E Yu and Shevlyakova A V 2014 J. Phys.: Conf. Ser. 490 012009
[9] Belashova I A, Bochkarev V V and Tyurin V A 2017 J. Phys.: Conf. Ser. 936 012074
[10] Dorogovtsev S, Goltsev A and Mendes J 2002 Phys. Rev. E 65 066122
[11] Ravasz E B and Barabási A L S 2003 Phys. Rev. E 67 026112