2013 12th Mexican International Conference on Artificial Intelligence

The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text

D. V. Lande, A. A. Snarskii
Institute for Information Recording NAS of Ukraine,
NTUU “Kiev Polytechnic Institute”
Kiev, Ukraine
dwlande@gmail.com, *asnarskii@gmail.com*

E. V. Yagunova, E. V. Pronoza
Saint-Petersburg State University
St.-Petersburg, Russian Federation
iagounova.elena@gmail.com, *katpronoza@gmail.com*

***Abstract*—A compactified horizontal visibility graph for the language network and identification of the words that define the informational structure of a text is proposed. It was found that the networks constructed in such a way are scale free, and have a property that among the nodes with largest degrees there are words that determine not only communicative text structure, but also its informational structure.**

Keywords-horizontal visibility graph; language network; dispersion estimated value; TFIDF; text coherence; informational structure

I. Introduction

Along with successive, or “linear” text analysis, construction of a net with text elements such as words and word combinations as its nodes can help reveal the structural elements of a text which make it coherent. Finding those structural elements which also have informational significance and form informational structure of a text is an important problem. These elements can be used for identification of text components which are not yet clearly defined, e.g., collocations, supra-phrasal units [1, 2, 3], when finding such components in various texts [2, 4].

II. Language Network Construction

A. Language Network

There are several methods used for constructing a network out of text. Such network is known as language network, and there are different ways of nodes and edges interpretation which results in different net representations. Nodes can be linked if the corresponding words are adjacent [5, 6], belong to the same sentence or paragraph [7], are connected syntactically [8, 9] or semantically [10, 11].

There are several methods of network construction on the basis of temporal series within digital signal processing and complex network theories [12, 13]. Some of them are based on visibility graph construction (see overview in [14]), e.g., so called horizontal visibility graph (HVG) [15, 16]. Such methods allow building network structures on the basis of texts where single words or word combinations map to some special numeric weights. Such mapping function may relate a word to its sequence number among the unique words in the text, its length, “weight” in the text, a common score like TFIDF (which, in its canonical form, equals to term frequency multiplied by the logarithm of inverse document frequency with respect to base 2) or its variants [17, 18], and other weighted scores.

The originality of the research is contained within the application of horizontal visibility graph used in digital signal processing to a linguistic object, namely a piece of literature. Language network of a text used to be constructed only by the traditional algorithms which are described later in the paper. The proposed algorithm enables to extract the words which not only have informational significance but also are important for text coherence. According to the algorithm, the words of the text are attributed with the intensionally important numeric values (TFIDF or dispersion estimated value which is given a closer look at later in the paper). Then horizontal visibility graph is built based on these values, and it is compactified according to the procedure given later in the paper. Weights of the nodes corresponding to the words of the text are then calculated, and it turns out that the maximum-weight words are the most important ones for both text coherence and the informational structure of the text. The examples given in the paper demonstrate the effectiveness of the proposed method in comparison with traditional simple language network analysis used in computational linguistics.

When calculating TFIDF score, a text consisting of N words is divided into equal fragments of M words (e.g., $M = 500$). Then for each i word in the text $df(i)$ – the number of fragments which include this word – is counted, as well as $n(i)$ – the total number of times it is seen in the text. After that the average TFIDF score is calculated for each word according to the formula

$$tfidf(i) = \frac{n(i)}{N} \log\left(\frac{N}{M \times df(i)}\right) \qquad (1)$$

A word significance dispersion estimated value [19] was also used in our research when constructing language network. This score is calculated as follows: let us consider the text consisting of N words ($n = 1,...,N$, where n is the position of a word in the text read from left to write). A word A is denoted as A_k^n, where $k = 1,2,...,K$ is the number of a word occurrence in the text and n is the position of the word in the text. For example, A_3^{50} refers to the word A which is on the 50th position and is seen for the third time in the text.

978-1-4799-2605-3/13 $31.00 © 2013 IEEE
DOI 10.1109/MICAI.2013.33

An interval between sequential occurences of a word is denoted by $\Delta A_k = A_{k+1}^m - A_k^n = m - n$, where the word A is at the m-th place when seen for the $(k+1)$-th time and at the n-th place and when seen for the k-th time.

Dispersoin estimated value proposed in [19] is calculated as

$$\sigma_A = \frac{\sqrt{\left\langle \Delta A^2 \right\rangle - \left\langle \Delta A \right\rangle^2}}{\left\langle \Delta A \right\rangle} \qquad (2)$$

where $\left\langle \Delta A \right\rangle$ is the average value of the $\Delta A_1, \Delta A_2, ..., \Delta A_k$ sequence, $\left\langle \Delta A^2 \right\rangle$ denotes the sequences $\Delta A_1^2, \Delta A_2^2, ..., \Delta A_k^2$ and K is the number of times the word A is seen in the text.

In fact, dispersion estimated value helps to separate words into those which are uniformly distributed and those which are not (it equals 0 for the uniform distribution case). Thus, this value estimates the discriminative power of the words, which can be used in Information Retrieval. The idea of dispersion estimated value is close to that of TFIDF. However, this value is less widespread than TFIDF and is appropriate for full single texts rather than for composite text corpora as in the case of TFIDF.

Unlike other types of series within digital signal processing theory, those with numeric values corresponding to the words can be transformed to horizontal visibility graphs which allow for non-numeric node values that may be represented by words conveying a particular meaning.

B. Horizontal Visibility Algorithm

According to the horizontal visibility algorithm, language network is built in three stages. First, a series of nodes are plotted on the horizontal axis with uniform spacing, each node corresponding to a word, in the order the words appear in the text. At the same time numeric weights ω_n (we consider ω_n to denote either TFIDF or dispersion estimated value) corresponding to the words are plotted on the vertical axis (they are depicted as a set of vertical lines – see Fig. 1) so that there is a set of columns with height varying from 0 to ω_n.

At the second stage a traditional horizontal visibility graph is built [11, 12]. Visibility is considered for the highest points of the nodes' columns. An edge is put between the nodes, if there is a visible connection between them, e.g., if they can be connected by a horizontal line which does not cross any column.

According to [15, 16], this geometric criterion can be written as follows: two nodes $B_i^n, G_j^m, i < j$ are connected if $\omega_n, \omega_m > \omega_p$ for all $n < p < m$ (e.g., B_3^n and G_7^{n+5} are connected – see Fig. 1).

The construction algorithm can be represented as follows. As shown in Fig. 1, for the A_1^{n+2} node the adjacent nodes are B_3^n and G_7^{n+5} (and the edges are drawn between them), so that B_3^n is the nearest neighbor of A_1^{n+2} to the left with $\omega_B = \omega_n$ weight value which is higher than that of A: $\omega_B > \omega_A = \omega_{n+2}$, and G_7^{n+5} is the nearest neighbor of A_1^{n+2} to the right, with $\omega_G = \omega_{n+5} > \omega_{n+2}$.

At the third stage language network is compactified. All the nodes with the given word, e.g., A, are merged into one node (and the index and word number disappear). All the edges of such nodes are also merged. It is important to note that multiple edges are removed, and there is no more than one edge left between every two nodes. In particular, this means that the degree of A does not exceed the sum $\sum_k A_k^n$ of degrees. As a result, we have a new language network – a compactified horizontal visibility graph (CHVG) – see Fig. 2.

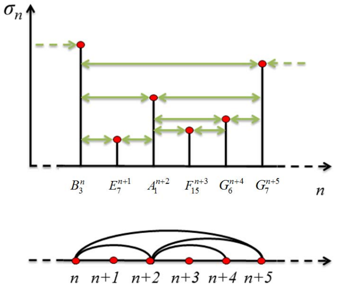


Figure 1. An Example of Horizontal Visibility Graph Construction

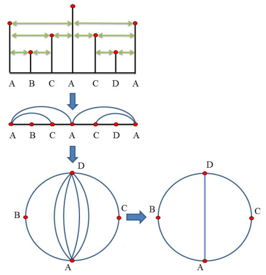


Figure 2. Stages of Compactified Horizontal Visibility Graph Construction

III. Experiment

We considered language network construction by the example of "Lovlya peskarey v Gruzii" ("The Catching of Gudgeons in Georgia") by Viktor Astafiev, "Reka" ("The River") by Yuri Bondarev, "Bez ulybok" ("Without Smiles") by Irina Grekova, "Svoy krug" ("Our Crowd") by Lyudmila Petrushevskaya and "Problema vervolka v sredney polose" ("A Werewolf Problem in Central Russia") by Viktor Pelevin. It should be noted that the authors have also conducted similar research on the dozens of other pieces of literature, including the novels "The Master and Margarita" by Mikhaill Bulgakov and "Moby-Dick, or The Whale" by Herman Melville, "The Hobbit, or There and Back Again" by J. R. R. Tolkien, "Dandelion Wine" by Ray Bradbury, e.t.c. News corpora and the Ukrainian and Russian acts of law texts were also analyzed. When considering news corpora, we had to filter out the words important for text coherence but bearing no informational significance and therefore stop words dictionaries were constructed based on various web resources[1]. As we have seen in our research, this way of choosing the words shows more conformance to

[1] *http://code.google.com/p/stop-words/source/browse/trunk/stop-words/stop-words/stop-words-russian.txt?spec=svn3&r=3*
https://github.com/punbb/langs/blob/master/Russian/stopwords.txt
http://www.ranks.nl/stopwords/russian.html
https://trac.mysvn.ru/punbb/punbb/browser/trunk/Russian/stopwords.txt

assessors' views than traditional methods. As the key results have shown coincidence with the ones given below, in this paper we only consider a few examples of the pieces of literature mentioned above.

For all CHVG language networks constructed the distribution of node degrees appeared to be close to power-series distribution ($p(k) = Ck^{\alpha}$), which means that these networks are scale free. Network parameters were calculated for the given pieces of literature. It turned out that for all of them α coefficient varies from –1 to –0.97 at relatively small approximation accuracy R^2 of power-series distribution, which improves as text length increases. This R^2 value totals 0.5–0.7 for short stories, and 0.95 – for larger pieces of literature, e.g., "Master i Margarita" ("The Master and Margarita") by Mikhail Bulgakov.

Along with personal pronouns and other functional parts of speech (particles, prepositions, conjunctions, etc.), the words which determine the informational structure of a text are among the maximum degree nodes in CHVG networks [21, 22].

For the sake of comparison we analyzed the simplest types of language networks where on the first stage of the algorithm neighbouring words in the text are connected, and on the second one network compactification takes place. It is obvious that in such networks node weights correspond to word frequencies, and their distribution – to Zipf's law [20]. At the same time we have maximum degree nodes for the maximum frequency words including conjunctions, prepositions, etc. – i.e., the words which are of great importance for text coherence and of little interest for the informational structure of a text [20, 21].

If we denote a set of N different words corresponding to the maximum degree nodes of a simple language network as Ψ – (we considered N = 100), and a set of words corresponding to the maximum degree nodes of CHVG as Λ , then $\Omega = \Lambda \setminus \Psi$ corresponds to the most informative words which are also important for text coherence. 100 maximum degree nodes for the three considered types of language network for "Lovlya peskarey v Gruzii", "Svoy krug" and "Problema vervolka v sredney polose" stories can be found in the appendix.

In particular, in CHVG network for the "Lovlya peskarey v Gruzii" story, with TFIDF as weight score, Ω includes words like "Дядя" ("Uncle"), "Вася" ("Vasya"), "Собора" (the genitive case of "Cathedral"), "Хозяин" ("Master"), "Грузии" (the genitive case of "Georgia"). Ω for the same CHVG network with dispersion estimated values as weight scores also includes words "Пескаря" (the genitive case for "Gudgeon"), "Рыбы" (the genitive case for "Fish"), "Храм" ("Temple"), "Горы" ("Mountains"), "Витязь" ("Knight"), etc.

As for "Svoy krug" by L. Petrushevskaya, such words as "Алешка" ("Alyoshka"), "Отец" ("Father"), "Время" ("Time"), "Жизни" (the genitive or locative case for "Life"), "Улице" (the dative or locative case for "Street") are in Ω . CHVG network Ω for this story with dispersion estimated values as weight scores includes "Любви" (the genitive or dative case for "Love"), "Ребенка" (the genitive or accusative case for "Child"), "Глаз" ("Eye") and "Андрея" ("Andrey").

For the "Problema vervolka v sredney polose" story by V. Pelevin Ω includes the words "Поляны" ("Meadows"), "Лапы" ("Paws" or "Boughs"), "Декан" ("Dena"), "Дороги" ("Roads"), "Машины" ("Cars") and "Девочка" ("Girl"). For CHVG network with dispersion estimated values as weight scores (for the same short story) Ω includes the words mentioned and also "Волки" ("Wolves") and "Волков" (the genitive or accusative case for "Wolves") – the words playing a special informational role in this piece of literature.

Top 100 nodes with largest degrees across the short stories mentioned are presented in tables 1–6[2] and in Fig. 3. Keywords based on the experiments with assessors across the short stories are presented in table 7.

Fig. 3 represents CHVG subgraphs for the short stories mentioned in the paper[3]. The subgraphs are based on dispersion estimated values. They only include large degree nodes except for the stop words nodes, and the edges are drawn only between the largest-degree nodes. The CHVG subgraphs demonstrate the presence of a rich club phenomenon in the language network of each of the short stories in question as the largest-degree nodes are most densely linked with each other while being sparsely linked with smaller-degree nodes.

The subgraph nodes correspond to words and not to lexemes because, although lemmatized graphs may appear more illustrative, we consider it important to maintain the tie with word forms from the original text[4].

The idea of informational significance of the given sets of words and their importance for the comprehension of the meaning of a piece of literature is proved in the experiments with assessors. Thus, for all the texts mentioned in the paper experiments were held with instructions as follows: "Read the text. Think about its content. Write down 10-15 words which are most important for the content of the text" and more than 20 assessors [23]. Keywords based on the experiments for "Lovlya peskarey v Gruzii", "Svoy krug" and "Problema vervolka v sredney polose" stories can be found in the appendix, and they can be compared with 100 maximum-degree nodes for the three considered types of language network for these short stories. The experiments have shown that assessors' views conform to the results obtained by the method of identifying the words that define the informational structure of the text proposed in this paper and to the principles of information density visualization. In

[2] The words from the CHVG node list which are absent from the simple network nodes list are in bold font. The words which have informational significance and are also in TOP-100 of the simple network are in italics.

[3] The graphs are drawn using software developed by D. Lande for power-law network illustration.

[4] It is well known that the Russian language has rich morphology, and for some NLP problems it is considered a drawback rather than an advantage.

this paper we only give the examples of fiction texts but we have also applied the proposed method to the other text genres including news, law and scientific texts. As part of the research, additional series of experiments were held concerning readability and missing-word tests (cloze tests).

IV. Conclusion

As a result of the research, the algorithm for compactified horizontal visibility graph (CHVG) construction is proposed.

Language networks for different texts are built on the basis of dispersion estimated values and CHVG.

For literary texts the CHVG nodes with maximum degree correspond to the words which not only provide for text coherence but also determine its informational structure and the semantics of the pieces of literature.

The algorithm of word weight calculation based on dispersion estimated values has proved to be more effective for identifying the words, which have informational significance and play an important role for text coherence, than the one based on TFIDF score.

References

[1] G. Y. Solganik, Sintaksicheskaya stilistika. Slozhnoe sintaksicheskoe tzeloe. Moscow: Vishaya shkola, 1991.

[2] E. V. Yagunova, “Variativnost’ strukturi narrativa i raznoobrazie strategiy ponimaniya,” Chelovek i yazik. K yubileyu Tamari Ivanovni Erofeevoy. Sb. nauchn. st. Perm, 2012, pp. 65–83.

[3] E. V. Yagunova, “Struktura spontannogo narrativa: multimediynost’ ishodnogo diskursa i ego otrazhenie v tekstah detey i vzroslih,” Sankt-Peterburgskaya shkola ontolingvistiki: sbornik statey k yubileyu S. N. Tzeytlin. SPb, 2013.

[4] A. Broder, “Identifying and filtering near-duplicate documents,” Proc. 11th Annual Symposium in Combinational Pattern Matching (CPM 2000), Montreal, Canada, June 2000, pp. 1–10, doi: 10.1007/3-540-45123-4_1.

[5] R. Ferrer i Cancho and R. V. Sole, “The small world of human language,” Proc. R. Soc. Lond., vol. B 268, Nov. 2011, pp. 2261–2265, doi: 10.1098/rspb.2001.1800.

[6] S. N. Dorogovtsev and J. F. F. Mendes, “Language as an evolving word web,” Proc. R. Soc. Lond., vol. B 268, Dec. 2011, pp. 2603–2606, doi: 10.1098/rspb.2001.1824.

[7] S. M. G. Caldeira, T. C. Petit Lobao, R. F. S. Andrade, A. Neme and J. G. V. Miranda, “The network of concepts in written texts,” Eur. Phys. J. B, vol. 49, pp. 523–529, 2006, doi: 10.1140/epjb/e2006-00091-3.

[8] R. Ferrer i Cancho, R. V. Sole and R. Kohler, “Petterns in syntactic dependency networks,” Physical Review E, vol. 69, May 2004, doi: 10.1103/PhysRevE.69.051915.

[9] R. Ferrer i Cancho, “The variation of Zipf’s law in human language,” Eur. Phys. J. B, vol. 44, pp. 249–257, 2005, doi: 10.1140/epjb/e2005-00121-8.

[10] A. E. Motter, A. P. S. de Moura, Y.-C. Lai and P. Dasgupta, “Topology of the conceptual network of language,” Physical Review E, vol. 65, June 2002, doi: 10.1103/PhysRevE.65.065102.

[11] M. Sigman and G.A. Cecchi, “Global properties of the Wordnet lexicon,” Proc. Natl. Acad. Sci. USA, 99(3), Feb. 2002, pp. 1742–1747, doi:10.1073/pnas.022341799.

[12] S. H. Strogatz, “Exploring complex networks,” Nature, vol. 410, Mar. 2001, pp. 268–276, doi:10.1038/35065725.

[13] R. Albert and A.-L. Barabasi, “Statistical mechanics of complex networks,” Reviews of Modern Physics, vol. 74, Mar. 2002, pp. 47–97, doi:10.1103/RevModPhys.74.47.

[14] A. M. Nunez, L. Lacasa, J. P. Gomez and B. Luque, “Visibility algorithms: A short review,” in New Frontiers in Graph Theory, ch. 6, Y. G. Zhang, Eds. Intech Press, 2002, pp. 119–152.

[15] B. Luque, L. Lacasa, F. Ballesteros, and J. Luque, “Horizontal visibility graphs: Exact results for random time series,” Physical Review E, vol. 80, Oct. 2009, doi: 10.1103/PhysRevE.80.046103.

[16] G. Gutin, T. Mansour and S. Severini, “A characterization of horizontal visibility graphs and combinatoris on words,” Physica A: Statistical Mechanics and its Applications, vol. 390(12), June 2011, pp. 2421–2428, doi: 10.1016/j.physa.2011.02.031.

[17] K. S. Jones, “Statistical Interpretation of Term Specificity and Its Application in Retrieval,” Journal of Documentation, vol. 28(1), 1972, pp. 11–21.

[18] G. Salton and M. J. McGill, Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

[19] M. Ortuño, P. Carpena, P. Bernaola, E. Muñoz and A. M. Somoza, “Keyword detection in natural languages and DNA,” Europhys. Lett., vol. 57(5), Mar. 2002, pp. 759–764.

[20] G. K. Zipf, Human Behavior and the Principle of Least Effort. Cambridge, MA: Addison-Wesley Press, 1949.

[21] R. Giora, “Segmentation and segment cohesion: On the thematic organization of the text,” Text – Interdisciplinary Journal for the Study of Discourse, vol. 2, 1983, pp. 155–181.

[22] L. A. Chernyahovkaya, “Smislovaya Struktura Teksta i ee Edinitzi,” Voprosi yazikoznaniya, Nov. –Dec. 1983, vol. 6, pp. 118–126.

[23] E. V. Yagunova, “Experiment i Vichisleniya v Analize Klyuchevih Slov Hudozhestvennogo Teksta,” in Sbornik Nauchnih Trudov Kafedri Inostrannih Yazikov i Filosofii PNTZ UrO RAN. Filosofiya yazika. Lingvistika. Lingvodidaktika, vol. 1, V. T. Yungblyud, Eds. Perm, 2010, pp. 83-89.

TABLE I. CHVG-TFIDF for "LOVLYA peskarey v Gruzii" by V. Astafiev

#	Word	#	Word	#	Word	#	Word	#	Word
1	и (and)	21	под (under)	41	над (above)	61	этот (this)	81	даже (even)
2	в (in)	22	бы (a particle)	42	ты (you)	62	без (without)	82	потом (afterwards)
3	я (i)	23	что (that)	43	время (time)	63	всех (everybody)	83	*речки (river)*
4	за (behind)	24	во (at)	44	уже (already)	64	тут (here)	84	**дом (home)**
5	на (on)	25	когда (when)	45	мне (me)	65	ли (whether)	85	чтоб (so that)
6	у (at)	26	только (only)	46	это (this)	66	**хозяин (master)**	86	про (about)
7	по (along)	27	о (about)	47	но (but)	67	вот (here)	87	среди (among)
8	не (not)	28	ни (not)	48	то (then)	68	нашей (our)	88	такой (such)
9	так (so)	29	*Отара (Otara)*	49	**дома (at home)**	69	себя (oneself)	89	совсем (quite)
10	к (to)	30	мы (we)	50	**дядя (uncle)**	70	где (where)	90	раз (if)
11	с (with)	31	было (was)	51	**Вася (Vasya)**	71	тогда (then)	91	нет (no)
12	еще (yet)	32	*братья (brothers)*	52	**собора (cathedral)**	72	куда (where)	92	**дождь (rain)**

13	от (from)	33	до (before)	53	со (with)	73	меня (me)	93	нам (us)
14	может (maybe)	34	их (them)	54	как (how)	74	которые (which)	94	**Грузии (Georgia)**
15	*Отар (Otar)*	35	они (they)	55	для (for)	75	**земли (ground)**	95	моего (my)
16	из (from)	36	*Гелати (Gelati)*	56	был (was)	76	здесь (here)	96	**сердце (heart)**
17	его (him)	37	возле (near)	57	да (yes)	77	тоже (too)	97	**гор (mountains)**
18	все (all)	38	*Шалва (Shalva)*	58	нас (us)	78	чтобы (so that)	98	если (if)
19	или (or)	39	же (but)	59	почти (almost)	79	лишь (only)	99	была (was)
20	а (and)	40	*столом (table)*	60	он (he)	80	чем (than)	100	**человек (human)**

TABLE II. CHVG-DISPERSION VALUE FOR „LOVLYA PESKAREY V GRUZII“ BY V. ASTAFIEV

#	Word	#	Word	#	Word	#	Word	#	Word
1	и (and)	21	*Отар (Otar)*	41	*Гелати (Gelati)*	61	без (without)	81	**друг (friend)**
2	в (in)	22	под (under)	42	о (about)	62	возле (near)	82	**детей (children)**
3	на (on)	23	ни (not)	43	меня (me)	63	ли (whether)	83	нам (us)
4	с (with)	24	еще (yet)	44	они (they)	64	со (with)	84	тут (here)
5	не (he)	25	когда (when)	45	над (above)	65	да (yes)	85	тоже (too)
6	я (i)	26	как (how)	46	этот (this)	66	совсем (quite)	86	чем (than)
7	за (after)	27	или (or)	47	же (but)	67	**дом (home)**	87	**пескаря (gudgeon)**
8	что (that)	28	ты (you)	48	**собора (cathedral)**	68	**дождь (rain)**	88	**горы (mountains)**
9	по (along)	29	время (time)	49	себя (oneself)	69	был (was)	89	раз (if)
10	от (from)	30	было (was)	50	до (before)	70	про (about)	90	потом (afterwards)
11	все (all)	31	**дома (at home)**	51	*столом (table)*	71	такой (such)	91	даже (even)
12	его (him)	32	это (this)	52	**хозяин master)**	72	**пескарей (gudgeons)**	92	где (where)
13	он (he)	33	во (in)	53	*Шалва (Shalva)*	73	нет (no)	93	среди (among)
14	у (at)	34	**дядя (uncle)**	54	только (only)	74	нашей (our)	94	против (against)
15	то (then)	35	бы (a particle)	55	для (for)	75	здесь (here)	95	чтобы (so that)
16	из (from)	36	мы (we)	56	почти (almost)	76	*речки (river)*	96	всего (altogether)
17	к (to)	37	может (maybe)	57	мне (me)	77	**храм (temple)**	97	**Витязь (Knight)**
18	а (and)	38	**Вася (Vasya)**	58	их (them)	78	уже (already)	98	всех (everybody)
19	так (so)	39	*Отара (Otara)*	59	**рыба (fish)**	79	**творчества (creation)**	99	вот (here)
20	но (but)	40	*братья (brothers)*	60	нас (us)	80	которые (which)	100	куда (where)

TABLE III. CHVG-TFIDF FOR „SVOY KRUG“ BY L. PETRUSHEVSKAYA

#	Word	#	Word	#	Word	#	Word	#	Word
1	и (and)	21	я (I)	41	из (from)	61	всех (everybody)	81	**ночь (night)**
2	в (in)	22	так (so)	42	была (was)	62	своей (his/her)	82	**дверь (door)**
3	он (he)	23	ни (not)	43	будет (will be)	63	*Маришу (Marisha)*	83	этот (this)
4	*Андрей (Andrey)*	24	ему (him)	44	*Алеша (Alyosha)*	64	*Алешу (Alyosha)*	84	чтобы (so that)
5	*Валера (Valera)*	25	к (to)	45	то (then)	65	при (by)	85	стал (began)
6	*Коля (Kolya)*	26	были (were)	46	тут (here)	66	вообще (at all)	86	спросила (asked)
7	не (not)	27	же (but)	47	ли (whether)	67	мой (my)	87	лет (years)
8	на (on)	28	был (was)	48	всегда (always)	68	тот (that)	88	им (them)
9	это (this)	29	мы (we)	49	от (from)	69	жить (live)	89	без (without)
10	с (with)	30	*Жора (Zhora)*	50	*Надя (Nadya)*	70	того (that)	90	**Алешка (Alyoshka)**
11	*Серж (Serge)*	31	как (how)	51	до (before)	71	где (where)	91	**улице (street)**
12	по (along)	32	бы (a particle)	52	потом (afterwards)	72	там (there)	92	под (under)
13	она (she)	33	у (at)	53	один (one)	73	себя (oneself)	93	**отец (father)**
14	а (and)	34	еще (yet)	54	нас (us)	74	мне (me)	94	**время (time)**
15	они (they)	35	было (was)	55	о (about)	75	со (with)	95	который (which)
16	ее (her)	36	сказала (said)	56	меня (me)	76	за (after)	96	только (only)
17	*Ленка (Lenka)*	37	*Мариша (Marisha)*	57	*Мариши (Marisha)*	77	нее (her)	97	**жизни (life)**
18	что (that)	38	вот (here)	58	но (but)	78	для (for)	98	сказал (said)
19	когда (when)	39	*Сержа (Serge)*	59	его (him)	79	или (or)	99	тоже (too)
20	все (all)	40	уже (already)	60	очень (very)	80	*Таня (Tanya)*	100	их (them)

TABLE IV. CHVG-DISPERSION VALUE FOR „SVOY KRUG“ BY L. PETRUSHEVSKAYA

#	Word	#	Word	#	Word	#	Word	#	Word
1	и (and)	21	это (this)	41	будет (will be)	61	со (with)	81	один (one)
2	в (in)	22	же (but)	42	те (those)	62	*Маришу (Marisha)*	82	стал (began)
3	не(not)	23	так (so)	43	до (before)	63	была (was)	83	мой (my)
4	а (and)	24	к (to)	44	ему (him)	64	своей (his/her)	84	**любви (love)**
5	я (i)	25	его (him)	45	всегда (always)	65	только (only)	85	им (them)
6	с (with)	26	за (after)	46	*Алешу (Alyosha)*	66	вот (here)	86	где (where)
7	на (on)	27	*Мариша (Marisha)*	47	ли (whether)	67	мне (me)	87	для (for)
8	все (all)	28	*ленка (lenka)*	48	очень (very)	68	быть (to be)	88	давно (long ago)
9	то (then)	29	ее (her)	49	**отец (father)**	69	который (which)	89	чем (than)
10	он (he)	30	из (from)	50	уже (already)	70	перед (before)	90	**время (time)**
11	*андрей (andrey)*	31	но (but)	51	тут (here)	71	ничего (nothing)	91	спросила (asked)
12	у (at)	32	*Жора (Zhora)*	52	был (was)	72	**глаз (eye)**	92	сказал (said)
13	что (that)	33	ни (not)	53	от (from)	73	ты (you)	93	**ребенка (child)**

14	*Серж (Serge)*	34	мы (we)	54	бы (a particle)	74	всех (everybody)	94	**Алешка (Alyoshka)**
15	как (how)	35	она (she)	55	*Надя (Nadya)*	75	потом (afterwards)	95	под (under)
16	*Валера (Valera)*	36	*Алеша (Alyosha)*	56	были (were)	76	над (above)	96	того (that)
17	*Коля (Kolya)*	37	*Сержа (Serge)*	57	*Мариши (Marisha)*	77	кто (who)	97	*Таня (Tanya)*
18	они (they)	38	еще (yet)	58	о (about)	78	себя (oneself)	98	**Андрея (Andrey)**
19	по (along)	39	когда (when)	59	раз (if)	79	при (by)	99	**улице (street)**
20	было (was)	40	сказала (said)	60	даже (even)	80	жить (live)	100	почему (why)

TABLE V. CHVG-TFIDF for „Problema vervolka v sredney polose“ by V. Pelevin

#	Word	#	Word	#	Word	#	Word	#	Word
1	и (and)	21	только (only)	41	ему (him)	61	кто (who)	81	время (time)
2	я (I)	22	вдруг (suddenly)	42	был (was)	62	*костра (fire)*	82	рядом (near)
3	в (in)	23	к (to)	43	чтобы (so that)	63	мне (me)	83	**машины (cars)**
4	было (was)	24	*Лена (Lena)*	44	же (but)	64	если (if)	84	поглядел (looked)
5	это (this)	25	бы (a particle)	45	где (where)	65	для (for)	85	сразу (at once)
6	с (with)	26	потом (afterwards)	46	из (from)	66	**лапы (paws)**	86	увидел (saw)
7	*вожак (leader)*	27	теперь (now)	47	раз (if)	67	или (or)	87	*лес (forest)*
8	на (on)	28	все (all)	48	тоже (too)	68	жизни (life)	88	почувствовал (felt)
9	ты (you)	29	как (how)	49	сейчас (now)	69	*морду (muzzle)*	89	чуть (scarcely)
10	его (him)	30	какой (what)	50	но (but)	70	почему (why)	90	стал (began)
11	что (that)	31	по (along)	51	под (under)	71	подумал (thought0	91	**девочка (girl)**
12	*Николай (Nikolay)*	32	когда (when)	52	понял (understood)	72	за (after)	92	перед (before)
13	она (she)	33	*глаза (eyes)*	53	него (him)	73	вот (here)	93	будет (will be)
14	он (he)	34	у (at)	54	еще (yet)	74	они (they)	94	идти (go)
15	сказал (said)	35	до (before)	55	так (so)	75	**декан (dean)**	95	вверх (up)
16	*Саша (Sasha)*	36	уже (already)	56	*дороге (road)*	76	**дороги (roads)**	96	назад (back)
17	не (not)	37	от (from)	57	себя (oneself)	77	во (in)	97	ее (her)
18	то (the)	38	о (about)	58	**поляны (meadows)**	78	одна (one)	98	заметил noticed)
19	вы (you)	39	были (were)	59	несколько (some)	79	через (through)	99	тебя (you)
20	а (and)	40	была (was)	60	ответил (answered)	80	чем (than)	100	здесь (here)

TABLE VI. CHVG-dispersion value for „Problema vervolka v sredney polose“ by V. Pelevin

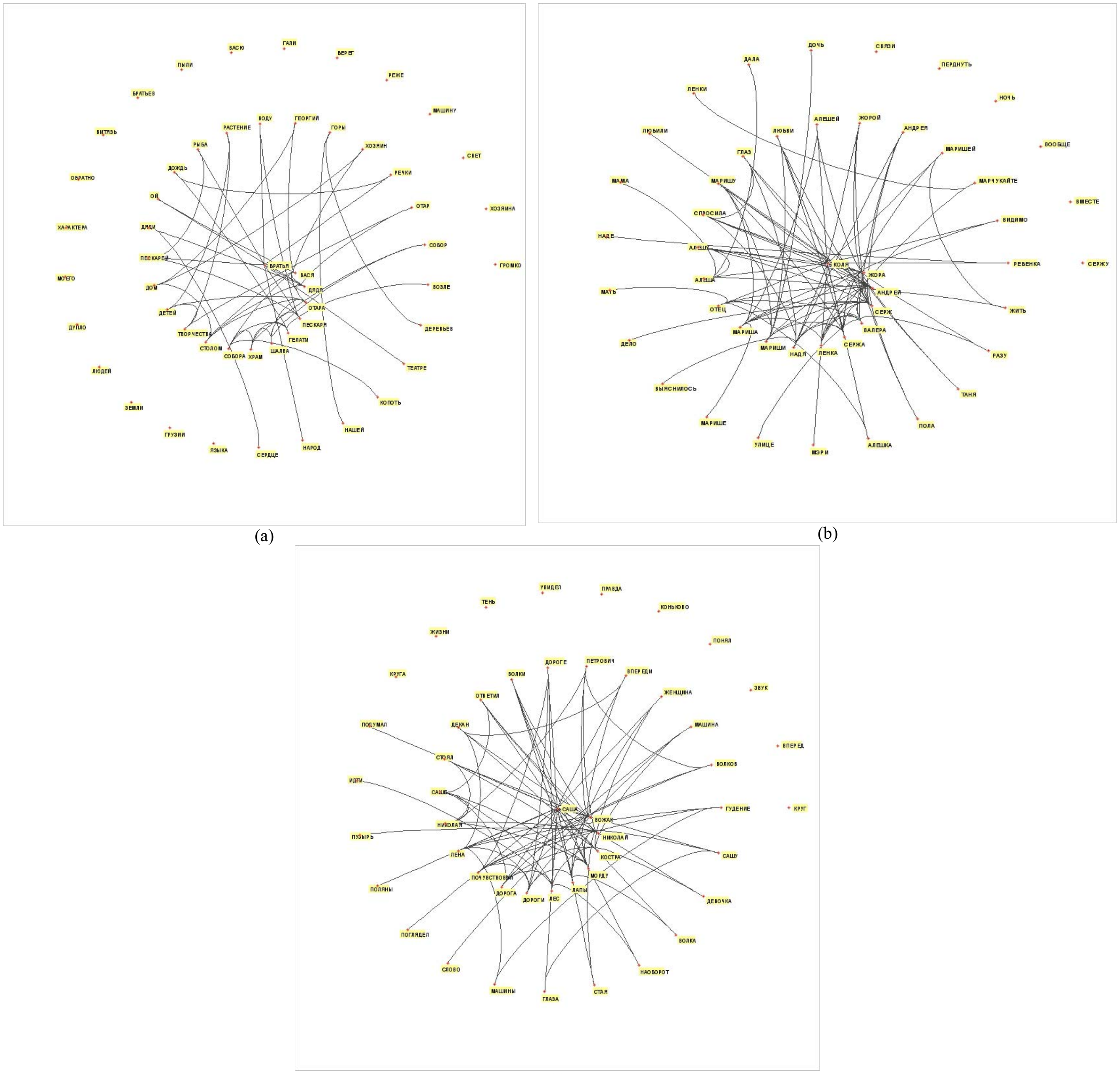
#	Word	#	Word	#	Word	#	Word	#	Word
1	и (and)	21	из (from)	41	него (him)	61	*морду (muzzle)*	81	**Волков (wolves)**
2	в (in)	22	она (she)	42	если (if)	62	ответил (answered)	82	*стая (troop)*
3	он (he)	23	уже (already)	43	*дорога (road)*	63	ему (him)	83	раз (if)
4	*Саша (Sasha)*	24	*костра (fire)*	44	через (through)	64	для (for)	84	мы (we)
5	на (on)	25	сказал (said)	45	был (was)	65	**лапы (paws)**	85	вверх (up)
6	то (then)	26	*Лена (Lena)*	46	они (they)	66	*глаза (eyes)*	86	при (at)
7	не (not)	27	за (after)	47	*лес (forest)*	67	*дороге (road)*	87	под (under)
8	это (this)	28	до (before)	48	же (but)	68	**девочка (girl)**	88	почувствовал (felt)
9	что (that)	29	но (but)	49	у (at)	69	почему (why)	89	назад (back)
10	с (with)	30	только (only)	50	была (was)	70	или (or)	90	их (them)
11	было (was)	31	вы (you)	51	во (in)	71	**декан (dean)**	91	вам (you)
12	я (i)	32	все (all)	52	о (about)	72	где (where)	92	слово (word)
13	его (him)	33	еще (yet)	53	будет (will be)	73	теперь (now)	93	сейчас (now)
14	к (to)	34	когда (when)	54	одна (one)	74	**поляны (meadow)**	94	кто (who)
15	по (along)	35	потом (afterwards)	55	чтобы (so that)	75	мимо (past)	95	друг (friend)
16	а (and)	36	бы (a particle)	56	были (were)	76	вокруг (around)	96	время (time)
17	*вожак (leader)*	37	какой (what)	57	**дороги (roads)**	77	такое (such)	97	будто (that)
18	ты (you)	38	от (from)	58	вот (here)	78	несколько (some)	98	этот (this)
19	*Николай (Nikolay)*	39	мне (me)	59	тоже (too)	79	**машина (car)**	99	**волки (wolves)**
20	как (how)	40	вдруг (suddenly)	60	так (so)	80	наоборот (vice versa)	100	**Саше (Sasha)**

TABLE VII. Keywords based on the experiments (lemma) with informants for „Problema vervolka v sredney polose“ by V. Pelevin, for „Lovlya peskarey v Gruzii“ by V. Astafiev and „Svoy krug“ by L. Petrushevskaya

Problema vervolka...	**Lemma**	**Lolvlya perskarey**	**Lemma**	**Svoy krug**	**Lemma**
Саша (Sasha)	0,67	Грузия (Georgia)	0,71	Мариша (Marisha)	0,62
дорога (road)	0,48	Витязь (Knight)	0,57	Алеша (ка) (Alyosha(ka))	0,52
драка (fight)	0,43	гость(и) (guest)	0,43	любовь (love)	0,43
стая (troop)	0,43	собор (cathedral)	0,33	умная(ый) (clever)	0,43
зов (call)	0,43	Гелати (Gelati)	0,33	кровь (blood)	0,38
Коньково (Kon'kovo)	0,43	храм (temple)	0,33	пятница (Friday)	0,38
машина (car)	0,38	рыбалка (fishing)	0,29	ребенок (дети) (child(ren))	0,38
костер (у костра) (fire)	0,38	брат (тья) (brother(s))	0,29	Пасха (Easter)	0,33
деревня (village)	0,38	земля (ground)	0,29	человек (human)	0,33
Вервольф(ки) (Werewolf(ves))	0,38	русский (Russian)	0,29	Серж (Serge)	0,33
волк(и) (wolf(ves))	0,33	Дом (Home)	0,29	отец (father)	0,24
поездка (trip)	0,33	гостеприимство	0,24	садиться на колени	0,24

лес (forest)	0,33	(hospitability)		(to sit on one's knees)	
девочка (girl)	0,33	грузины (Georgian(s))	0,24	друзья (friends)	0,24
поляна (meadow)	0,33	подарочек (a present)	0,24	Коля (Kolya)	0,24
Лена (Lena)	0,33	праздник (holiday)	0,24	болезнь (disease)	0,24

луна (moon)	0,33	рыба (fish)	0,24	я (I)	0,24
страх (fear)	0,33				



(a)

(b)

(c)

Figure 3. CHVG subgraphs for „Lovlya peskarey v Gruzii“ by V. Astafiev (a), „Svoy krug“ by L. Petrushevskaya (b) and „Problema vervolka v sredney polose“ by V. Pelevin (c)