Использование графов горизонтальной видимости для выявления слов, определяющих информационную структуру текста

© Д. В. Ландэ
Институт проблем регистрации информации НАН Украины, НТУУ «Киевский политехнический институт», Украина, Киев
dwlande@gmail.com

© А. А. Снарский
asnarskii@gmail.com

© Е. В. Ягунова
С.-Петербургский гос. унив., С.-Петербург, Россия
iagounova.elena@gmail.com

Аннотация

Предлагается методика компактифицированного графа горизонтальной видимости для создания сети слов и выявления тех слов в тексте, которые определяют его информационную структуру. Исследованы свойства таких сетей слов, показано, что они являются безмасштабными, а также, что среди узлов с наибольшими степенями имеются слова, определяющие не только структуру связности текста, но и его информационную структуру.

Наряду с последовательным, «линейным» анализом текстов, построение сетей, узлами которых являются их элементы – слова или словосочетания, фрагменты естественного языка, позволяет выявлять структурные элементы текста, без которых он теряет свою связность. При этом актуальной является задача определения того, какие из важных структурных элементов оказываются также информационно-значимыми, определяющими информационную структуру текста. Такие элементы могут использоваться также для идентификации еще не достаточно четко определенных компонент текста, таких как коллокации, сверхфразовые единства [1], например, при поиске подобных фрагментов в различных текстах [2].

Известно несколько подходов к построению сетей из текстов, так называемых сетей слов (Language Network), и различные способы интерпретации узлов и связей, что приводит, соответственно, к различным видам представления таких сетей. Узлы могут быть соединены между собой, если соответствующие им слова стоят рядом в тексте [3, 4], принадлежат одному предложению или абзацу [5], соединены синтаксически [6, 7] или семантически [8, 9].

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

В рамках теорий цифровой обработки сигналов (Digital Signal Processing) и сложных сетей (Complex Network) [10, 11] предложено несколько методов построения сетей на основе временных рядов, среди которых можно назвать несколько методов построения графов видимости (см. обзор [12]), в частности, так называемый граф горизонтальной видимости (Horizontal Visibility Graph – HVG) [13,14]. Эти подходы также позволяют строить сетевые структуры на основании текстов, в которых отдельным словам или словосочетаниям некоторым специальным образом поставлены в соответствие числовые весовые значения. В качестве функции, ставящей в соответствие слову число, можно рассматривать, например, порядковый номер уникального слова в тексте, длину слова, «вес» слов в текстах, общепринятую оценку TFIDF (в каноническом виде, равную произведению частоты слова в фрагменте текста – term frequency – на двоичный логарифм от величины, обратной количеству фрагментов текста, в которых это слово встретилось – inverse document frequency) или ее варианты [15, 16], а также другие весовые оценки.

В качестве весовой оценки TFIDF из полного текста, состоящего из N слов, текст разбивается на фрагменты, содержащие заданное количество слов M (например, $M = 500$). Затем для каждого слова i, входящего в текст, подсчитывается количество фрагментов $df(i)$, в которые это слово входит, а также общее количество вхождений данного слова i в текст – $n(i)$. После этого по формуле

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

рассчитывается среднее значение TFIDF весовой оценки каждого слова.

При построении сетей слов в данной работе также будет использована дисперсионная оценка важности слов [17], которая реализуется следующим образом: пусть текст состоит из N слов ($n = 1,...N$, n – порядковый номер слова в тексте, позиция слова). Некоторое слово, например A, обозначается как A_k^n, где индекс $k = 1,2,...,K$ – номер появления данного слова в тесте, а n – позиция данного слова в тексте. Например, A_3^{50}

означает, что на 50-й позиции текста находится слово A, которое встретилось третий раз.

Интервал между последовательными появлениями слова при таких обозначениях будет величина $\Delta A_k = A_{k+1}^m - A_k^n = m - n$, где на m-м и n-м позициях в тесте находится слово A, которое встретилось $k+1$-й и k-й разы.

Предложенная в [27] дисперсионная оценка рассчитывается как

$$\sigma_A = \frac{\sqrt{\langle \Delta A^2 \rangle - \langle \Delta A \rangle^2}}{\langle \Delta A \rangle},$$

где: $\langle \Delta A \rangle$ – среднее значение последовательности $\Delta A_1, \Delta A_2, ..., \Delta A_K$, $\langle \Delta A^2 \rangle$ – последовательности $\Delta A_1^2, \Delta A_2^2, ..., \Delta A_K^2$, K –количество появления слова A в тексте.

По сути, дисперсионная оценка позволяет отделить слова, встречающиеся в тексте относительно равномерно (для равномерно распределенных слов эта оценка равна нулю), от слов, распределенных неравномерно. Т.е. это оценка различительной, дискриминантной силы слов, в частности, для информационного поиска. Идея дисперсионной оценки очень близка к TFIDF, при этом менее распространена, однако более корректно применима к полным единичным текстам, а не к массивам текстов, как TFIDF.

В отличие от остальных рядов, изучаемых в рамках цифровой обработки сигналов, ряды из цифровых значений, соответствующих словам, преобразуются в графы горизонтальной видимости, в которых узлам соответствуют не только цифровые значения, но сами слова, выражающие определенное смысловое значение.

Сеть слов с использованием алгоритма горизонтальной видимости строится в три этапа. На первом на горизонтальной оси отмечается ряд узлов, каждый из которых соответствует словам в порядке появления в тексте, а по вертикальной оси откладываются весовые численные оценки (визуально – набор вертикальных линий, см. рис.1).

На втором этапе строится традиционный граф горизонтальной видимости [21]. Для этого между узлами существует связь, если они находятся в «прямой видимости», т.е. если их можно соединить горизонтальной линией, не пересекающей никакую другую вертикальную линию. Этот (геометрический) критерий можно записать, согласно [15,16] следующим образом: два узла (слова) слова, например, B_3^n и C_7^m $(m = n+5)$ соединены связью, если (см. рис. 1) $\sigma_n, \sigma_m > \sigma_p$ для всех $n < p < m$.

Алгоритм построения можно представить удобным для вычисления способом. Так например, на рис. 1 для узла-слова A_1^{n+2} смежными в сети считаются слова B_3^n и C_1^{n+5} (и устанавливаются ребра-связи), такие что B_3^n – ближайшее слева от A_1^{n+2} слово, с дисперсионной оценкой $\sigma_n = \sigma_B$, превышающей дисперсионную оценку слова A $\sigma_{n+2} = \sigma_A$, а $C_7\,(m = n+5)$ – ближайшее справа от A_1^{n+2} слово, для которого $\sigma_{105} > \sigma_{102}$.

На третьем, заключительном этапе, полученная на предыдущем этапе сеть компактифицируется. Все узлы с данным словом, например словом A, объединяются в один узел (естественно, индекс и номер положения слова при этом исчезают). Все связи таких узлов также объединяются. Важно отметить, что между любыми двумя узлами при этом остается не более одной связи – кратные связи изымаются. В частности это означает, что степень (число связей) узла A не превышает суммы степеней $\sum_k A_k^n$. В результате получается новая сеть слов – *компактифицированный граф горизонтальной видимости* (КГГВ) – рис.2.

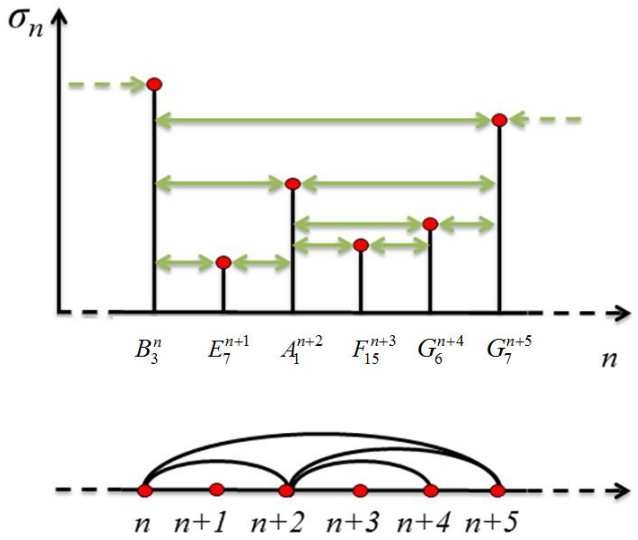

Рис. 1. Пример построения графа горизонтальной видимости

В качестве текстов при построении сетей слов в рамках данной статьи рассматриваются рассказы В. Астафьева «Ловля пескарей в Грузии», Ю. Бондарева «Река», И. Грековой «Без улыбок», Л. Петрушевской «Свой круг» и В. Пелевина «Проблема верволка в средней полосе». Следует, отметить, что авторами проводились подобные исследования на базе десятков других произведений, в том числе, значительно более объемных. Анализировались также законодательные акты Украины и России. Концептуальные результаты анализа при этом совпадали с приведенными ниже, поэтому остановимся на предложенных произведениях, как примерах.

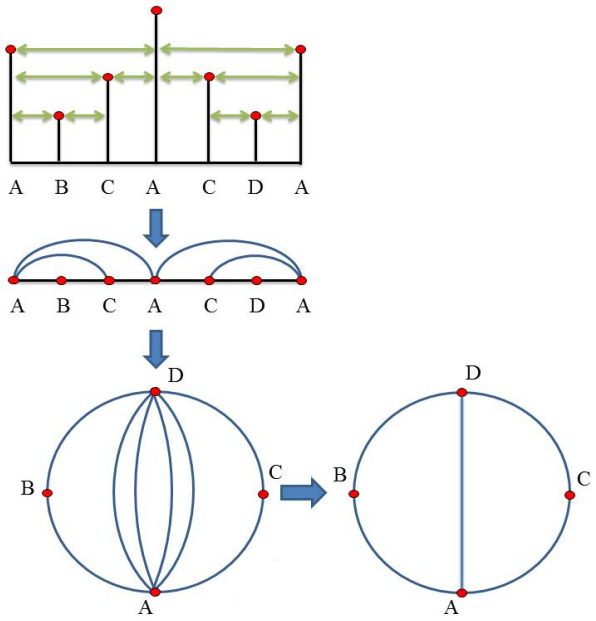

Рис. 2. Этапы построения компактификационного графа горизонтальной видимости

Для всех построенных КГГВ-сетей слов было определено распределение степеней узлов, которое оказалось близким к степенному ($p(k) = Ck^{\alpha}$), т.е. эти сети являются безмасштабными. Были проведены расчеты параметров сетей для всех рассмотренных литературных произведений. В результате оказалось, что для всех из них коэффициент α изменялся в диапазоне от –1 до –0,97 при относительно небольшой точности аппроксимации R^2 степенного распределения, которая повышается при увеличении длины текста. Для рассказов эта точность составила 0,5-0,7, а для сравнительно больших произведений, исследованных авторами, например, для романа М. Булгакова «Мастер и Маргарита» – 0,95.

В состав узлов с наибольшими степенями в для КГГВ-сетей, наряду с личными местоимениями и другими служебными словами (частицы, предлоги, союзы и т.д.), попали слова, определяющие информационную структуру текста [18, 19].

Для сравнения исследовано поведение простейших сетей языка, когда не первом этапе построения сети связываются соседние слова, входящие в текст, а на втором происходит компактификация сети. Очевидно, вес узлов в этой сети соответствует частоте встречаемости слов, а их распределение – закону Ципфа [20]. При этом самые большие степени имеют узлы, соответствующие словам с наибольшей частотой – союзам, предлогами и т.п., имеющим большое значение для связности текста, но малоинтересным с точки зрения информационной структуры.

Если обозначить Ψ – множество из N различных слов, соответствующих наиболее весомым узлам приведенной простейшей сети языка, а Λ – множество из слов, соответствующих наиболее весомым узлам КГГВ, то множество $\Omega = \Lambda \setminus \Psi$ соответствует информативным словам, имеющим, кроме того, важное значение и для связности текста. В Приложении приведены сопоставления 100 наиболее весомых узлов для трех рассматриваемых типов сетей слов по рассказам В. Астафьева «Ловля пескарей в Грузии», Л. Петрушевской «Свой круг» и В. Пелевина «Проблема верволка в средней полосе». Рассматривался случай N = 100, что было выбрано достаточно произвольно, с учетом того, что для рассматриваемых небольших по объему произведений важнейшие с точки зрения смысла слова попали в данный диапазон.

В частности, в КГГВ-сети по весовым значениям TFIDF, по рассказу В. Астафьева ««Ловля пескарей в Грузии» в состав множества Ω попали такие слова, как «Дядя», «Вася», «Собора», «Хозяин», «Грузии». В КГГВ-сети для этого же рассказа по весовым значениям, соответствующим дисперсионным оценкам, в состав множества Ω попали дополнительно такие слова, как «Пескаря», «Рыбы», «Храм», «Горы», «Витязь» и др.

При анализе рассказа Л. Петрушевской «Свой круг» в множество Ω попали такие слова, как «Алешка», «Отец», «Время», «Жизни», «Улице». В КГГВ-сети для этого рассказа по весовым значениям, соответствующим дисперсионным оценкам, в состав множества Ω попали дополнительно такие слова, как «Любви», «Ребенка», «Глаз», «Андрея».

В случае рассказа В. Пелевина «Проблема верволка в средней полосе» в состав множества Ω попали такие слова, как «Поляны», «Лапы», «Декан», «Дороги», «Машины», «Девочка». В КГГВ-сети для этого же рассказа по весовым значениям, соответствующим дисперсионным оценкам, в состав множества Ω попали те же слова, что и в предыдущем случае, и, кроме того, слова «Волки» и «Волков», играющие особую информационную роль в данном произведении.

Представления об информационной значимости рассматриваемых наборов слов, степени их важности для понимания смысла литературного произведения были подтверждены в ходе экспериментов с информантами. Так, для всех текстов были проведены эксперименты со стандартной инструкцией «Прочитайте текст. Подумайте над его содержанием. Выпишите 10-15 слов, наиболее важных для его содержания» (более 20 информантов для каждого текста) [21].

В результате проведенных исследований сетей:

1. Предложен алгоритм построения компактифицированного графа горизонтальной видимости (КГГВ).
2. На основе последовательности дисперсионных оценок слов текста и TFIDF, с помощью метода КГГВ, построены сети слов различных текстов.

3. Для литературных текстов среди узлов соответствующих КГГВ с наибольшими степенями присутствуют слова, не только обеспечивающие связность структуры текста, но и определяющие его информационную структуру, отражают семантику литературных произведений.
4. Алгоритм определения веса слов, базирующийся на дисперсионной оценке оказался более эффективным для определения информационно-значимых слов, играющих важное значение для структурной связности в литературных текстах, чем алгоритм TFIDF.

Литература

[1] Солганик Г. Я. Синтаксическая стилистика. Сложное синтаксическое целое. – 2-е изд., испр. и доп. – М.: Высш. шк. – 182 с. (1991).

[2] Broder A. Identifying and Filtering Near-Duplicate Documents, COM'00 // Proceedings of the 11th Annual Symposium on Combinatorial Pattern Matching. – P. 1-10 (2000).

[3] Ferrer-i-Cancho R., Sole R. V. The small world of human language // Proc. R. Soc. Lond. – B 268, 2261 (2001).

[4] Dorogovtsev S.N., Mendes J. F. F. Language as an evolving word web // Proc. R. Soc. Lond. – B 268, 2603 (2001).

[5] Caldeira S. M. G., Petit Lobao T. C., Andrade R. F. S., Neme A., Miranda J. G. V. The network of concepts in written texts // Preprint physics/0508066 (2005).

[6] Ferrer-i-Cancho R., Sole R.V., Kohler R. Patterns in syntactic dependency networks // Phys. Rev. E 69, 051915 (2004).

[7] Ferrer-i-Cancho R. The variation of Zipf's law in human language. // Phys. Rev. E 70, 056135 (2005).

[8] Motter A. E., de Moura A. P. S., Lai Y.-C., Dasgupta P. Topology of the conceptual network of language // Phys. Rev. E 65, 065102(R) (2002).

[9] Sigman M., Cecchi G. A. Global Properties of the Wordnet Lexicon // Proc. Natl. Acad. Sci. USA, 99, 1742 (2002).

[10] Strogatz S. H. Exploring Complex Networks // Nature. – 410. – P. 268-276 (2001).

[11] Albert R., Barabasi A.-L. Statistical mechanics of complex networks // Reviews of Modern Physics. – 74. – P. 47 (2002).

[12] Nunez A. M., Lacasa L., Gomez J. P., Luque B. Visibility algorithms: A short review // New Frontiers in Graph Theory, Y. G. Zhang, Ed. Intech Press, ch. 6. – P. 119 – 152 (2012).

[13] Luque B., Lacasa L., Ballesteros F., Luque J. Horizontal visibility graphs: Exact results for random time series // Physical Review E,– P. 046103-1–046103-11 (2009).

[14] Gutin G., Mansour T., Severini S. A characterization of horizontal visibility graphs and combinatoris on words // Physica A, – 390 – P. 2421-2428 (2011).

[15] Jones K.S. A statistical interpretation of term specificity and its application in retrieval // Journal of Documentation. – 28 (1). – P. 11–21 (1972).

[16] 16. Salton G., McGill M. J. Introduction to Modern Information Retrieval. – New York: McGraw-Hill. – 448 p. (1983).

[17] Ortuño M., Carpena P., Bernaola P., Muñoz E., Somoza A.M. Keyword detection in natural languages and DNA // Europhys. Lett, – 57(5). – P. 759-764 (2002).

[18] Черняховская Л.А. Смысловая структура текста и ее единицы // Вопросы языкознания. – № 6. – С. 118–126. (1983).

[19] Giora R. Segmentation and Segment Cohesion: On the Thematic Organization of the Text // Text. An Interdisciplinary Journal for the Study of Discourse Amsterdam. – 3. – № 2. – P. 155-181 (1983).

[20] Zipf G.K. Human Behavior and the Principle of Least Effort. – Cambridge, MA: Addison-Wesley Press. – 573 p. (1949).

[21] Ягунова Е.В. Эксперимент и вычисления в анализе ключевых слов художественного текста // Сборник научных трудов кафедры иностранных языков и философии ПНЦ УрО РАН. Философия языка. Лингвистика. Лингводидактика / Отв. ред. В.Т. Юнгблюд.– Пермь, 2010. – Вып. 1. – С. 85- 91.

The Use Of Horizontal Visibility Graphs To Identify The Words That Define The Information Structure Of The Text

D.V. Lande, A.A. Snarskii, E.V. Yagunova

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

Приложение. Сопоставление 100 наиболее весомых узлов сетей по рассказам*

1. В. Астафьев, «Ловля пескарей в Грузии»

Простейшая сеть

1	И	21	ОТАР	41	ТОЖЕ	61	СЕБЯ	81	ЛЮДЕЙ
2	В	22	ПОД	42	ТОЛЬКО	62	ЛИ	82	КОТОРЫЕ
3	НА	23	БЫЛО	43	ОТАРА	63	КУДА	83	ЕСТЬ
4	С	24	ОНИ	44	МНЕ	64	ЕМУ	84	ЕСЛИ
5	НЕ	25	ЕЩЕ	45	ВОЗЛЕ	65	БРАТЬЯ	85	ДОМА
6	ЧТО	26	ТАК	46	ВО	66	СЕРДЦЕ	86	ЧЕЛОВЕК
7	Я	27	МЫ	47	ТЫ	67	НАД	87	ВСЕГДА
8	ПО	28	ИЛИ	48	ДЛЯ	68	МЕНЯ	88	СОВСЕМ
9	ЗА	29	ЖЕ	49	ТУТ	69	ГЕЛАТИ	89	ПОТОМ
10	ИЗ	30	ДО	50	ГДЕ	70	БЕЗ	90	НАМ
11	ТО	31	КОГДА	51	РАЗ	71	ЧЕМ	91	ИМ
12	НО	32	ЭТО	52	ВРЕМЯ	72	НАС	92	ДАЖЕ
13	ОТ	33	БЫ	53	О	73	БЫЛА	93	БЫЛИ
14	ОН	34	НИ	54	ЧТОБЫ	74	ЗДЕСЬ	94	ЗЕМЛИ
15	ВСЕ	35	ДА	55	ВСЕХ	75	ВСЕГО	95	ВСЕМ
16	КАК	36	БЫЛ	56	ПОЧТИ	76	УЖ	96	ТОГО
17	ЕГО	37	МОЖЕТ	57	ЧТОБ	77	СРЕДИ	97	СТОЛОМ
18	А	38	ИХ	58	ЭТОТ	78	РЕЧКИ	98	ПРО
19	У	39	ВОТ	59	СО	79	НЕТ	99	ОДНАКО
20	К	40	УЖЕ	60	ШАЛВА	80	НАШЕЙ	100	ОДИН

КГГВ–TFIDF

1	И	21	ПОД	41	НАД	61	ЭТОТ	81	ДАЖЕ
2	В	22	БЫ	42	ТЫ	62	БЕЗ	82	ПОТОМ
3	Я	23	ЧТО	43	ВРЕМЯ	63	ВСЕХ	83	*РЕЧКИ*
4	ЗА	24	ВО	44	УЖЕ	64	ТУТ	84	**ДОМ**
5	НА	25	КОГДА	45	МНЕ	65	ЛИ	85	ЧТОБ
6	У	26	ТОЛЬКО	46	ЭТО	66	**ХОЗЯИН**	86	ПРО
7	ПО	27	О	47	НО	67	ВОТ	87	СРЕДИ
8	НЕ	28	НИ	48	ТО	68	НАШЕЙ	88	ТАКОЙ
9	ТАК	29	*ОТАРА*	49	**ДОМА**	69	СЕБЯ	89	СОВСЕМ
10	К	30	МЫ	50	**ДЯДЯ**	70	ГДЕ	90	РАЗ
11	С	31	БЫЛО	51	**ВАСЯ**	71	ТОГДА	91	НЕТ
12	ЕЩЕ	32	*БРАТЬЯ*	52	**СОБОРА**	72	КУДА	92	**ДОЖДЬ**
13	ОТ	33	ДО	53	СО	73	МЕНЯ	93	НАМ
14	МОЖЕТ	34	ИХ	54	КАК	74	КОТОРЫЕ	94	**ГРУЗИИ**
15	*ОТАР*	35	ОНИ	55	ДЛЯ	75	**ЗЕМЛИ**	95	МОЕГО
16	ИЗ	36	*ГЕЛАТИ*	56	БЫЛ	76	ЗДЕСЬ	96	**СЕРДЦЕ**
17	ЕГО	37	ВОЗЛЕ	57	ДА	77	ТОЖЕ	97	**ГОР**
18	ВСЕ	38	*ШАЛВА*	58	НАС	78	ЧТОБЫ	98	ЕСЛИ
19	ИЛИ	39	ЖЕ	59	ПОЧТИ	79	ЛИШЬ	99	БЫЛА
20	А	40	*СТОЛОМ*	60	ОН	80	ЧЕМ	100	**ЧЕЛОВЕК**

КГГВ–дисперсионная оценка

1	И	21	*ОТАР*	41	*ГЕЛАТИ*	61	БЕЗ	81	**ДРУГ**
2	В	22	ПОД	42	О	62	ВОЗЛЕ	82	**ДЕТЕЙ**
3	НА	23	НИ	43	МЕНЯ	63	ЛИ	83	НАМ
4	С	24	ЕЩЕ	44	ОНИ	64	СО	84	ТУТ
5	НЕ	25	КОГДА	45	НАД	65	ДА	85	ТОЖЕ
6	Я	26	КАК	46	ЭТОТ	66	СОВСЕМ	86	ЧЕМ
7	ЗА	27	ИЛИ	47	ЖЕ	67	**ДОМ**	87	**ПЕСКАРЯ**
8	ЧТО	28	ТЫ	48	**СОБОРА**	68	**ДОЖДЬ**	88	**ГОРЫ**
9	ПО	29	ВРЕМЯ	49	СЕБЯ	69	БЫЛ	89	РАЗ
10	ОТ	30	БЫЛО	50	ДО	70	ПРО	90	ПОТОМ
11	ВСЕ	31	**ДОМА**	51	*СТОЛОМ*	71	ТАКОЙ	91	ДАЖЕ
12	ЕГО	32	ЭТО	52	**ХОЗЯИН**	72	**ПЕСКАРЕЙ**	92	ГДЕ
13	ОН	33	ВО	53	*ШАЛВА*	73	НЕТ	93	СРЕДИ
14	У	34	**ДЯДЯ**	54	ТОЛЬКО	74	НАШЕЙ	94	ПРОТИВ
15	ТО	35	БЫ	55	ДЛЯ	75	ЗДЕСЬ	95	ЧТОБЫ
16	ИЗ	36	МЫ	56	ПОЧТИ	76	*РЕЧКИ*	96	ВСЕГО
17	К	37	МОЖЕТ	57	МНЕ	77	**ХРАМ**	97	**ВИТЯЗЬ**
18	А	38	**ВАСЯ**	58	ИХ	78	УЖЕ	98	ВСЕХ
19	ТАК	39	*ОТАРА*	59	**РЫБА**	79	**ТВОРЧЕСТВА**	99	ВОТ
20	НО	40	*БРАТЬЯ*	60	НАС	80	КОТОРЫЕ	100	КУДА

* Слова, присутсвующие в списке узлов КГГВ, но отсутствующие в списке узлов простейшей сети выделены жирным шрифтом. Наиболее информационно-значимые слова, также присутствующие и в топ-100 простейшей сети, выделены курсивом.

2. Л. Петрушевская, «Свой круг»

Простейшая сеть

1	И	21	СЕРЖ	41	ЖОРА	61	ЛИ	81	МАРИШУ
2	В	22	ЖЕ	42	БЫ	62	ВООБЩЕ	82	МАРИШЕ
3	НЕ	23	ТАК	43	ТУТ	63	ТЕПЕРЬ	83	ИХ
4	НА	24	МАРИША	44	СЕРЖА	64	СВОЕЙ	84	РАЗ
5	А	25	ИЗ	45	ДО	65	ОДИН	85	ПОСЛЕ
6	С	26	АНДРЕЙ	46	ВОТ	66	НАДЯ	86	МАРИШИ
7	ВСЕ	27	КОЛЯ	47	БЫЛИ	67	ЛЕТ	87	ИМ
8	КАК	28	КОГДА	48	ОТ	68	ГДЕ	88	ЭТОТ
9	Я	29	ВАЛЕРА	49	МЫ	69	ВСЕГДА	89	ТЫ
10	ЧТО	30	БЫЛ	50	ЛЕНКА	70	ТАМ	90	ТАНЯ
11	ОН	31	ТОЛЬКО	51	ПОТОМ	71	ОЧЕНЬ	91	ПОСКОЛЬКУ
12	У	32	УЖЕ	52	ВСЕХ	72	О	92	НИЧЕГО
13	ТО	33	ЕЩЕ	53	СО	73	АЛЕША	93	НАДО
14	ЭТО	34	ЕЕ	54	МНЕ	74	ВРЕМЯ	94	ДАЖЕ
15	НО	35	ОНИ	55	ДЛЯ	75	ТОЖЕ	95	БЕЗ
16	ЕГО	36	НИ	56	БУДЕТ	76	ТОГО	96	ТОТ
17	ПО	37	ЕМУ	57	СКАЗАЛА	77	СВОЮ	97	СКАЗАЛ
18	К	38	КОТОРЫЙ	58	МЕНЯ	78	СТАЛ	98	НАД
19	ЗА	39	ОНА	59	ЧТОБЫ	79	ПОД	99	ДОМА
20	БЫЛО	40	БЫЛА	60	СЕБЯ	80	МОЙ	100	АЛЕШУ

КГГВ-TFIDF

1	И	21	Я	41	ИЗ	61	ВСЕХ	81	**НОЧЬ**
2	В	22	ТАК	42	БЫЛА	62	СВОЕЙ	82	**ДВЕРЬ**
3	ОН	23	НИ	43	БУДЕТ	63	*МАРИШУ*	83	ЭТОТ
4	*АНДРЕЙ*	24	ЕМУ	44	*АЛЕША*	64	*АЛЕШУ*	84	ЧТОБЫ
5	*ВАЛЕРА*	25	К	45	ТО	65	ПРИ	85	СТАЛ
6	*КОЛЯ*	26	БЫЛИ	46	ТУТ	66	ВООБЩЕ	86	СПРОСИЛА
7	НЕ	27	ЖЕ	47	ЛИ	67	МОЙ	87	ЛЕТ
8	НА	28	БЫЛ	48	ВСЕГДА	68	ТОТ	88	ИМ
9	ЭТО	29	МЫ	49	ОТ	69	ЖИТЬ	89	БЕЗ
10	С	30	*ЖОРА*	50	*НАДЯ*	70	ТОГО	90	**АЛЕШКА**
11	*СЕРЖ*	31	КАК	51	ДО	71	ГДЕ	91	**УЛИЦЕ**
12	ПО	32	БЫ	52	ПОТОМ	72	ТАМ	92	ПОД
13	ОНА	33	У	53	ОДИН	73	СЕБЯ	93	**ОТЕЦ**
14	А	34	ЕЩЕ	54	НАС	74	МНЕ	94	**ВРЕМЯ**
15	ОНИ	35	БЫЛО	55	О	75	СО	95	КОТОРЫЙ
16	ЕЕ	36	СКАЗАЛА	56	МЕНЯ	76	ЗА	96	ТОЛЬКО
17	*ЛЕНКА*	37	*МАРИША*	57	*МАРИШИ*	77	НЕЕ	97	**ЖИЗНИ**
18	ЧТО	38	ВОТ	58	НО	78	ДЛЯ	98	СКАЗАЛ
19	КОГДА	39	*СЕРЖА*	59	ЕГО	79	ИЛИ	99	ТОЖЕ
20	ВСЕ	40	УЖЕ	60	ОЧЕНЬ	80	*ТАНЯ*	100	ИХ

КГГВ-дисперсионная оценка

1	И	21	ЭТО	41	БУДЕТ	61	СО	81	ОДИН
2	В	22	ЖЕ	42	ТЕ	62	*МАРИШУ*	82	СТАЛ
3	НЕ	23	ТАК	43	ДО	63	БЫЛА	83	МОЙ
4	А	24	К	44	ЕМУ	64	СВОЕЙ	84	**ЛЮБВИ**
5	Я	25	ЕГО	45	ВСЕГДА	65	ТОЛЬКО	85	ИМ
6	С	26	ЗА	46	*АЛЕШУ*	66	ВОТ	86	ГДЕ
7	НА	27	*МАРИША*	47	ЛИ	67	МНЕ	87	ДЛЯ
8	ВСЕ	28	*ЛЕНКА*	48	ОЧЕНЬ	68	БЫТЬ	88	ДАВНО
9	ТО	29	ЕЕ	49	**ОТЕЦ**	69	КОТОРЫЙ	89	ЧЕМ
10	ОН	30	ИЗ	50	УЖЕ	70	ПЕРЕД	90	**ВРЕМЯ**
11	*АНДРЕЙ*	31	НО	51	ТУТ	71	НИЧЕГО	91	СПРОСИЛА
12	У	32	*ЖОРА*	52	БЫЛ	72	**ГЛАЗ**	92	СКАЗАЛ
13	ЧТО	33	НИ	53	ОТ	73	ТЫ	93	**РЕБЕНКА**
14	*СЕРЖ*	34	МЫ	54	БЫ	74	ВСЕХ	94	**АЛЕШКА**
15	КАК	35	ОНА	55	*НАДЯ*	75	ПОТОМ	95	ПОД
16	*ВАЛЕРА*	36	*АЛЕША*	56	БЫЛИ	76	НАД	96	ТОГО
17	*КОЛЯ*	37	*СЕРЖА*	57	*МАРИШИ*	77	КТО	97	*ТАНЯ*
18	ОНИ	38	ЕЩЕ	58	О	78	СЕБЯ	98	**АНДРЕЯ**
19	ПО	39	КОГДА	59	РАЗ	79	ПРИ	99	**УЛИЦЕ**
20	БЫЛО	40	СКАЗАЛА	60	ДАЖЕ	80	ЖИТЬ	100	ПОЧЕМУ

3. В. Пелевин, «Проблема верволка в средней полосе»

Простейшая сеть

1	И	21	ЗА	41	ЕСЛИ	61	ПОД	81	ПЕРЕД
2	В	22	ТЫ	42	ГЛАЗА	62	НЕСКОЛЬКО	82	МОРДУ
3	НА	23	ОНА	43	ДО	63	ЕЕ	83	ЧУТЬ
4	ОН	24	ОТ	44	ЧТОБЫ	64	МНЕ	84	ЭТОТ
5	САША	25	ТАК	45	СЕЙЧАС	65	КАКОЙ	85	ЗДЕСЬ
6	НЕ	26	ЕЩЕ	46	О	66	СТАЛ	86	ПЕРЕД
7	ЧТО	27	КОГДА	47	ГДЕ	67	НЕГО	87	В
8	А	28	ТОЛЬКО	48	БЫЛ	68	ПОЧЕМУ	88	ТЕБЯ
9	ТО	29	ВОЖАК	49	ВОКРУГ	69	КТО	89	РЯДОМ
10	С	30	НИКОЛАЙ	50	ИЛИ	70	ЭТОГО	90	ДОРОГЕ
11	ПО	31	ПОТОМ	51	ВОТ	71	ОТВЕТИЛ	91	ЧЕГО
12	КАК	32	ОНИ	52	БЫЛИ	72	МОЖЕТ	92	ВВЕРХ
13	ЭТО	33	У	53	БЫЛА	73	ДА	93	ВРЕМЯ
14	ЕГО	34	СКАЗАЛ	54	ДЛЯ	74	БУДЕТ	94	ВО
15	К	35	ЖЕ	55	ТОЖЕ	75	ЧЕМ	95	СТАЯ
16	НО	36	ЕМУ	56	СЕБЯ	76	ПОДУМАЛ	96	ОЧЕНЬ
17	ИЗ	37	ВДРУГ	57	ЛЕНА	77	ЛЕС	97	ОПЯТЬ
18	БЫЛО	38	УЖЕ	58	ЧЕРЕЗ	78	ИХ	98	ОДНА
19	Я	39	БЫ	59	ТЕПЕРЬ	79	УВИДЕЛ	99	НИБУДЬ
20	ВСЕ	40	ВЫ	60	РАЗ	80	ПОСЛЕ	100	НЕТ

КГГВ-TFIDF

1	И	21	ТОЛЬКО	41	ЕМУ	61	КТО	81	ВРЕМЯ
2	Я	22	ВДРУГ	42	БЫЛ	62	*КОСТРА*	82	РЯДОМ
3	В	23	К	43	ЧТОБЫ	63	МНЕ	83	**МАШИНЫ**
4	БЫЛО	24	*ЛЕНА*	44	ЖЕ	64	ЕСЛИ	84	ПОГЛЯДЕЛ
5	ЭТО	25	БЫ	45	ГДЕ	65	ДЛЯ	85	СРАЗУ
6	С	26	ПОТОМ	46	ИЗ	66	**ЛАПЫ**	86	УВИДЕЛ
7	*ВОЖАК*	27	ТЕПЕРЬ	47	РАЗ	67	ИЛИ	87	*ЛЕС*
8	НА	28	ВСЕ	48	ТОЖЕ	68	ЖИЗНИ	88	ПОЧУВСТВОВАЛ
9	ТЫ	29	КАК	49	СЕЙЧАС	69	*МОРДУ*	89	ЧУТЬ
10	ЕГО	30	КАКОЙ	50	НО	70	ПОЧЕМУ	90	СТАЛ
11	ЧТО	31	ПО	51	ПОД	71	ПОДУМАЛ	91	**ДЕВОЧКА**
12	*НИКОЛАЙ*	32	КОГДА	52	ПОНЯЛ	72	ЗА	92	ПЕРЕД
13	ОНА	33	*ГЛАЗА*	53	НЕГО	73	ВОТ	93	БУДЕТ
14	ОН	34	У	54	ЕЩЕ	74	ОНИ	94	ИДТИ
15	СКАЗАЛ	35	ДО	55	ТАК	75	**ДЕКАН**	95	ВВЕРХ
16	*САША*	36	УЖЕ	56	*ДОРОГЕ*	76	**ДОРОГИ**	96	НАЗАД
17	НЕ	37	ОТ	57	СЕБЯ	77	ВО	97	ЕЕ
18	ТО	38	О	58	**ПОЛЯНЫ**	78	ОДНА	98	ЗАМЕТИЛ
19	ВЫ	39	БЫЛИ	59	НЕСКОЛЬКО	79	ЧЕРЕЗ	99	ТЕБЯ
20	А	40	БЫЛА	60	ОТВЕТИЛ	80	ЧЕМ	100	ЗДЕСЬ

КГГВ-дисперсионная оценка

1	И	21	ИЗ	41	НЕГО	61	*МОРДУ*	81	**ВОЛКОВ**
2	В	22	ОНА	42	ЕСЛИ	62	ОТВЕТИЛ	82	*СТАЯ*
3	ОН	23	УЖЕ	43	*ДОРОГА*	63	ЕМУ	83	РАЗ
4	*САША*	24	*КОСТРА*	44	ЧЕРЕЗ	64	ДЛЯ	84	МЫ
5	НА	25	СКАЗАЛ	45	БЫЛ	65	**ЛАПЫ**	85	ВВЕРХ
6	ТО	26	*ЛЕНА*	46	ОНИ	66	*ГЛАЗА*	86	ПРИ
7	НЕ	27	ЗА	47	*ЛЕС*	67	*ДОРОГЕ*	87	ПОД
8	ЭТО	28	ДО	48	ЖЕ	68	**ДЕВОЧКА**	88	ПОЧУВСТВОВАЛ
9	ЧТО	29	НО	49	У	69	ПОЧЕМУ	89	НАЗАД
10	С	30	ТОЛЬКО	50	БЫЛА	70	ИЛИ	90	ИХ
11	БЫЛО	31	ВЫ	51	ВО	71	**ДЕКАН**	91	ВАМ
12	Я	32	ВСЕ	52	О	72	ГДЕ	92	СЛОВО
13	ЕГО	33	ЕЩЕ	53	БУДЕТ	73	ТЕПЕРЬ	93	СЕЙЧАС
14	К	34	КОГДА	54	ОДНА	74	**ПОЛЯНЫ**	94	КТО
15	ПО	35	ПОТОМ	55	ЧТОБЫ	75	МИМО	95	ДРУГ
16	А	36	БЫ	56	БЫЛИ	76	ВОКРУГ	96	ВРЕМЯ
17	*ВОЖАК*	37	КАКОЙ	57	**ДОРОГИ**	77	ТАКОЕ	97	БУДТО
18	ТЫ	38	ОТ	58	ВОТ	78	НЕСКОЛЬКО	98	ЭТОТ
19	*НИКОЛАЙ*	39	МНЕ	59	ТОЖЕ	79	**МАШИНА**	99	**ВОЛКИ**
20	КАК	40	ВДРУГ	60	ТАК	80	НАОБОРОТ	100	**САШЕ**