ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
XV Всероссийская научная конференция
Ярославль,
14-17 октября 2013 года

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

Д.В. ЛАНДЭ,
Институт проблем регистрации информации НАН Украины, Киев

А.А. СНАРСКИЙ
НТУУ «Киевский политехнический институт», Украина

Е.В. ЯГУНОВА,
Санкт-Петербургский гос. университет, Россия

Ярославль, 17 октября 2013 г.

Аннотация

Предлагается алгоритм компактифицированного графа горизонтальной видимости. На его основе строятся сети 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.

Актуальность

Наряду с последовательным, «линейным» анализом текстов, построение сетей, узлами которых являются их элементы – 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 теоретически определенных компонент текста, таких как коллокации, сверхфразовые единства, например, при поиске подобных фрагментов в различных текстах.

Сети слов

Первым шагом при применении теории сложных сетей к анализу текста является представление этого текста в виде совокупности узлов и связей, построение сети языка (Language Network).

Существуют различные способы интерпретации узлов и связей, что приводит, соответственно, к различным представлениям сети языка.

Узлы могут быть соединены меду собой, если соответствующие им слова стоят рядом в тексте, принадлежат одному предложению, соединены синтаксически или семантически.

Простейшие сети слов

- **L-пространство**. Связываются соседние слова, которые принадлежат одному предложению. Количество соседей для каждого слова (окно слова) определяется радиусом взаимодействия R, чаще всего рассматривается случай R = 1.

B-пространство. Рассматриваются узлы двух видов, соответствующие предложениям и словам, которые им принадлежат.

P-пространство. Все слова, которые принадлежат одному предложению, связываются между собой.

C-пространство. Предложения связываются между собой, если в них употреблены одинаковые слова.

Из рядов - в графы

На стыке теорий цифровой обработки сигналов (Digital Signal Processing) и сложных сетей (Complex Network) существует несколько методов построения сетей на основе временных рядов, среди которых можно назвать несколько методов построения графов видимости, в частности, так называемый граф горизонтальной видимости (Horizontal Visibility Graph – HVG). Эти подходы позволяют строить сетевые структуры также и на основании текстов, в которых отдельным словам или словосочетаниям некоторым специальным образом поставлены в соответствие некоторые весовые значения.

Весовые оценки слов

В качестве функции, ставящей в соответствие слову из текста число, можно рассматривать, например, порядковый номер уникального слова в тексте, длину слова, «вес» слов в текстах, общепринятую оценку TFIDF или ее варианты, а также другие весовые оценки, в частности, статистические дисперсионные.

TFIDF

В качестве весовой оценки из полного текста, состоящего из слов, текст разбивается на фрагменты, содержащие заданное количество N 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 рассчитывается среднее значение весовой оценки каждого слова в тексте, близкое по идеологии к классическому TFIDF:

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

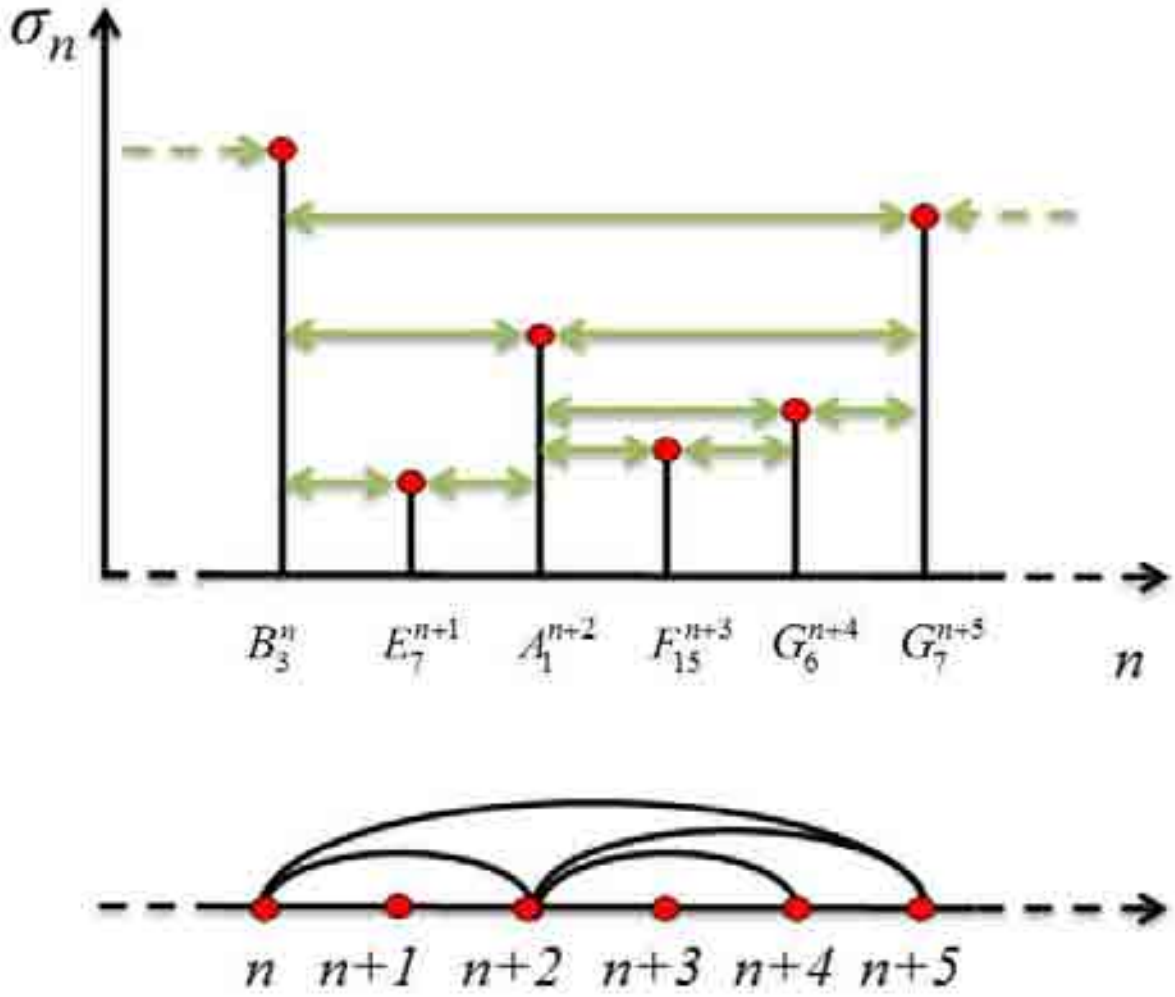
Дисперсионная оценка

Дисперсионная σ_A оценка для некоторого слова *A* из текста рассчитывается как

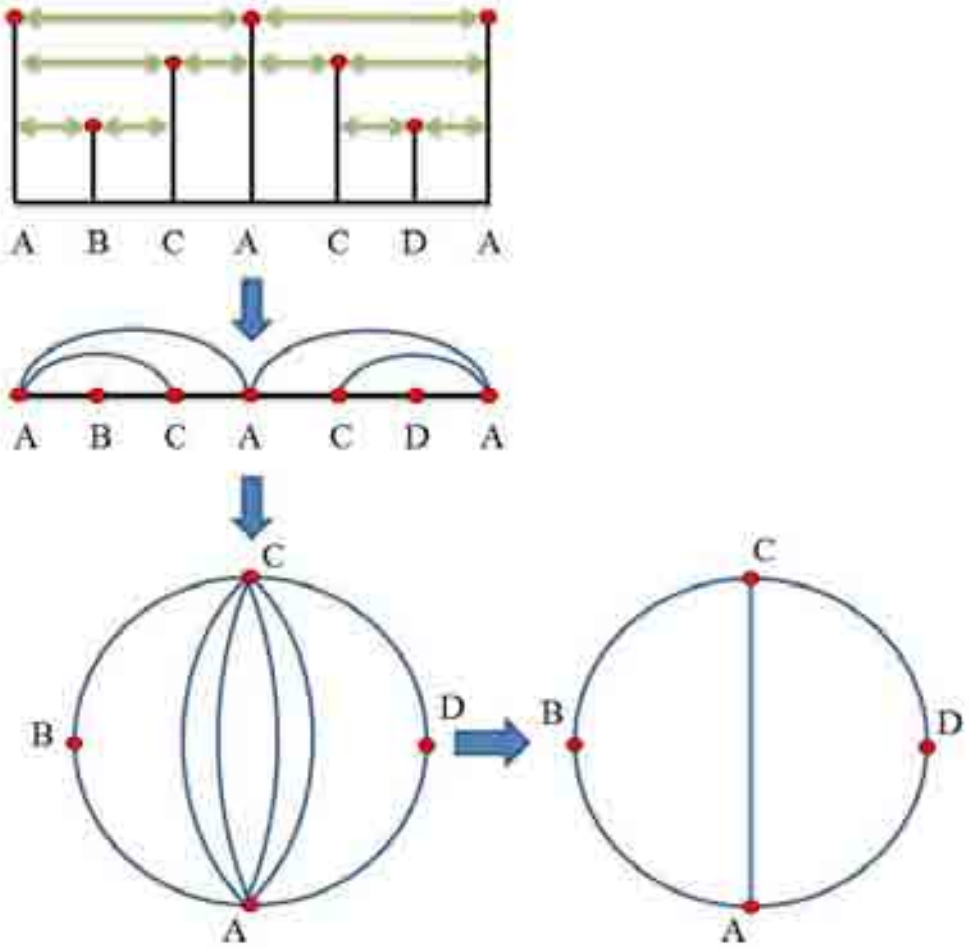
$$\sigma_A = \frac{\sqrt{\langle \Delta A^2 \rangle - \langle \Delta A \rangle^2}}{\langle \Delta A \rangle},$$

где: $\langle \Delta A \rangle$ – среднее расстояние (в словах) между появлениями слова *A* в тесте; : $\langle \Delta A^2 \rangle$ – среднее квадрата расстояния между появлениями слова *A* в тексте.

Формирование графа горизонтальной видимости



Этапы построения КГГВ



Компактифицированный граф горизонтальной видимости

Основы оценки методов

Если обозначить Ψ – множество из N различных слов (рассматривался случай N = 100), 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.

Сопоставление 100 наиболее весомых узлов сетей по рассказу

В. Астафьев, «Ловля пескарей в Грузии»

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	НАШЕИ	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	НАШЕИ	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

Сопоставление 100 наиболее весомых узлов сетей по рассказу

В. Пелевин, «Проблема верволка в средней полосе»

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	*НИКОЛАИ*	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	*НИКОЛАИ*	39	МНЕ	59	ТОЖЕ	79	**МАШИНА**	99	**ВОЛКИ**
20	КАК	40	ВДРУГ	60	ТАК	80	НАОБОРОТ	100	**САШЕ**

Сопоставление 100 наиболее весомых узлов сетей по рассказу Л. Петрушевской, «Свой круг»

KITB–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

KITB–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

Сопоставление 100 наиболее весомых узлов (топ-100) сетей слов по роману М. Булгакова «Мастер и Маргарита»*

Простейшая сеть		КГГВ	
Вес	Слово	Вес	Слово
5724	И	14724	И
3591	В	12880	В
2235	НА	8069	НЕ
1893	НЕ	7550	НА
1616	С	6511	ЧТО
1396	ЧТО	6050	ОН
1204	ОН	5225	ТО
1081	А	5224	Я
979	ЕГО	5105	С
936	ТО	4518	*МАРГАРИТА*
936	КАК	3642	ЕГО
899	НО	3396	А
809	К	3009	К
760	Я	2996	КАК
709	ИЗ	2848	**ИВАН**
680	ПО	2847	ОНА
634	ЗА	2562	ИЗ
555	ОТ	2509	ВЫ
553	У	2441	*ПРОКУРАТОР*
534	ЭТО	2317	ЗА
521	ВСЕ	2313	ПО
520	ЖЕ	2206	БЫЛО
514	ОНА	2076	ЭТО
484	МАРГАРИТА	2057	НО
460	ЕЕ	2000	У
409	БЫЛО	1989	О
403	ПОД	1940	ЕЕ
403	БЫЛ	1914	ВСЕ
400	ТАК	1904	*КОРОВЬЕВ*
382	ВЫ	1859	*ВОЛАНД*
379	УЖЕ	1815	БЫ
375	ЕМУ	1761	БЫЛ
333	БЫ	1721	*КОТ*
328	О	1696	ТАК
321	ТУТ	1693	*АЗАЗЕЛЛО*

Простейшая сеть		КГГВ	
Вес	Слово	Вес	Слово
237	ЭТОТ	1020	*ЧЕЛОВЕК*
222	КОТ	1007	ВАС
219	ПРОКУРАТОР	978	СКАЗАЛ
219	ГЛАЗА	961	ЭТОГО
215	СО	944	**ГОСТЬ**
213	ВАС	919	ГДЕ
212	ИЛИ	905	**ВАРЕНУХА**
210	ВОТ	886	**МАСТЕР**
209	СОВЕРШЕННО	871	**НИКАНОР**
207	ЧЕЛОВЕК	866	**БУФЕТЧИК**
206	ЛИ	861	УЖЕ
206	КОРОВЬЕВ	825	ТЕПЕРЬ
204	ТЕПЕРЬ	815	ЕЩЕ
199	АЗАЗЕЛЛО	807	ЧТОБЫ
197	ИХ	805	**ИВАНОВИЧ**
193	СКАЗАЛ	803	НУ
187	НАД	798	**СТЕПА**
184	ВАМ	790	НАД
183	СЕБЯ	766	ВАМ
183	ОНИ	761	ВО
183	КТО	740	**РИМСКИЙ**
182	БЫЛА	738	ОЧЕНЬ
177	ПЕРЕД	724	ОТВЕТИЛ
175	ТОТ	722	СО
172	ЧЕРЕЗ	720	КОГДА
171	БЫЛИ	719	НИЧЕГО
166	ВО	671	**МАРГАРИТЕ**
165	ВОЛАНД	663	*ЛИЦО*
165	НЕГО	657	**ПРОФЕССОР**
162	ТОГДА	656	ЛИ
157	ОТВЕТИЛ	652	**ИВАНА**
157	ЛИЦО	651	ЧЕРЕЗ
156	ДАЖЕ	649	МЫ
153	ВРЕМЯ	644	ВРЕМЯ
150	СЕЙЧАС	641	ДО

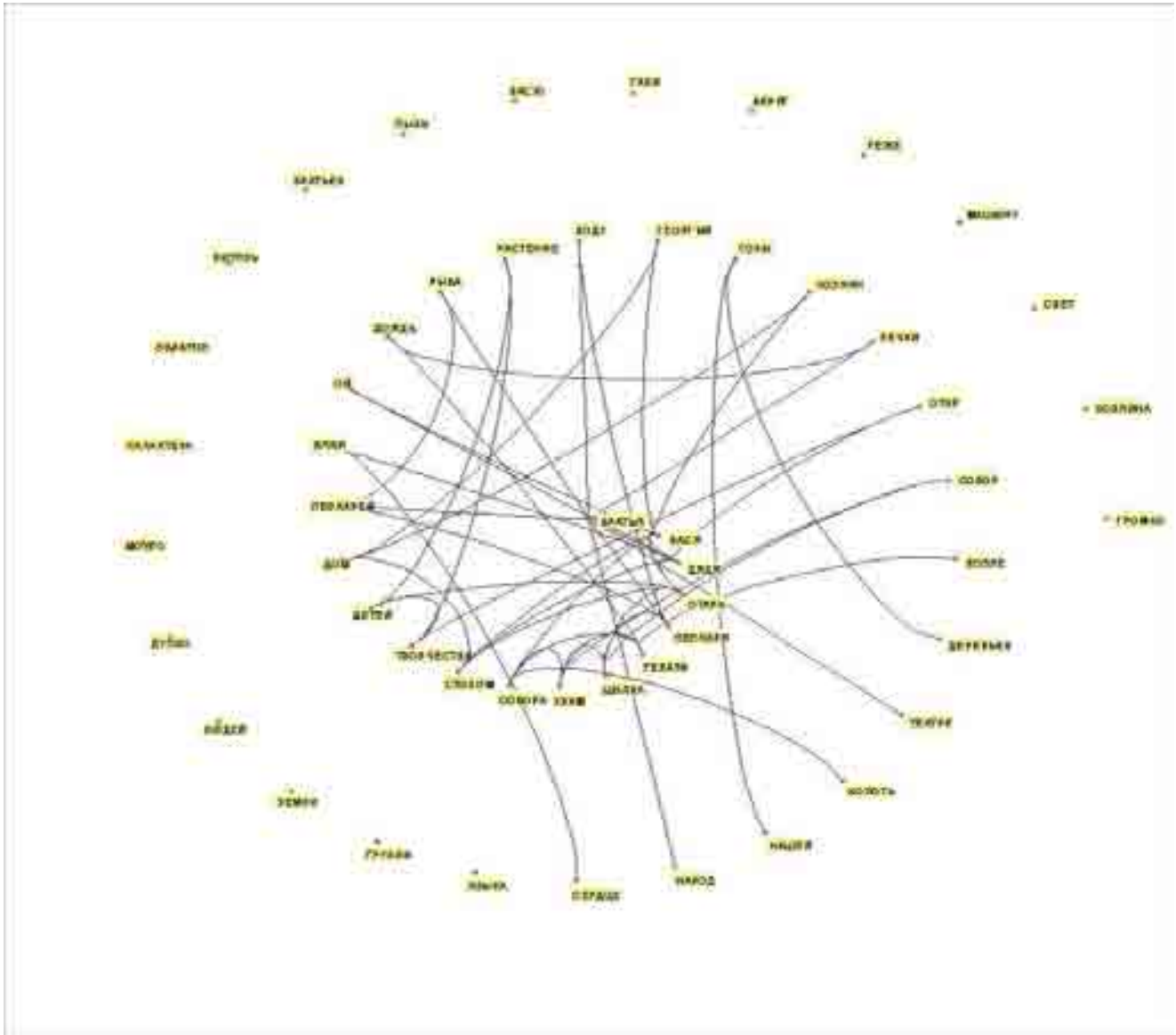
Сопоставление 100 наиболее весомых узлов (топ-100) сетей слов по роману Г. Мелвилла «Моби Дик, или Белый кит» (Moby-Dick, or The Whale)

Простейшая сеть		КГГВ	
Вес узла	Слово	Вес узла	Слово
6612	THE	41291	THE
5589	AND	23567	OF
4257	OF	17704	I
3083	A	16585	A
2862	TO	16577	AND
2730	IN	14853	HIS
2050	THAT	11976	IS
1915	HIS	11961	TO
1568	BUT	11582	HE
1524	IT	11431	WAS
1400	HE	10956	IN
1341	WITH	9883	WHALE
1301	FOR	9516	THAT
1281	I	9244	IT
1248	AS	7483	AS
1166	IS	7224	YOU
1152	WAS	6640	AHAB
1148	THIS	6457	HIM
1086	ALL	5727	BE
1008	BY	4867	BY
977	SO	4753	THIS
924	OR	4747	ALL
887	AT	4647	WITH
847	FROM	4578	ME
832	ON	4511	BUT
796	NOW	4403	HAD
784	NOT	4182	YE
733	WERE	4147	THEIR
721	THERE	4143	FROM
713	ONE	4038	FOR
703	HIM	3921	MY
697	THEIR	3645	WERE
694	YOU	3618	NOT

Простейшая сеть		КГГВ	
Вес узла	Слово	Вес узла	Слово
467	MORE	2591	OUT
458	OUT	2590	SPERM
451	WE	2575	HAVE
445	UP	2538	OLD
441	INTO	2482	THOU
433	THESE	2351	THEM
431	OLD	2317	WHALES
429	AHAB	2291	ONE
425	THEM	2259	ITS
425	ITS	2252	MAN
414	YE	2214	WHAT
397	YET	2187	STARBUCK
381	HER	2159	LIKE
380	WHO	2085	WHITE
369	OVER	2053	INTO
361	STILL	2010	MORE
360	THOUGH	1981	NO
360	ONLY	1944	THEN
353	MAN	1934	SOME
352	HERE	1903	UP
351	WILL	1891	AN
348	SEA	1872	UPON
343	SUCH	1846	THESE
343	LONG	1836	SUCH
339	VERY	1788	WHEN
338	WOULD	1694	BEEN
336	ABOUT	1665	PEQUOD
331	THOSE	1634	ABOUT
326	BEEN	1592	THOUGH
321	OTHER	1589	SEEMED
320	YOUR	1574	YOUR
318	THOU	1549	OVER
317	IF	1544	OUR

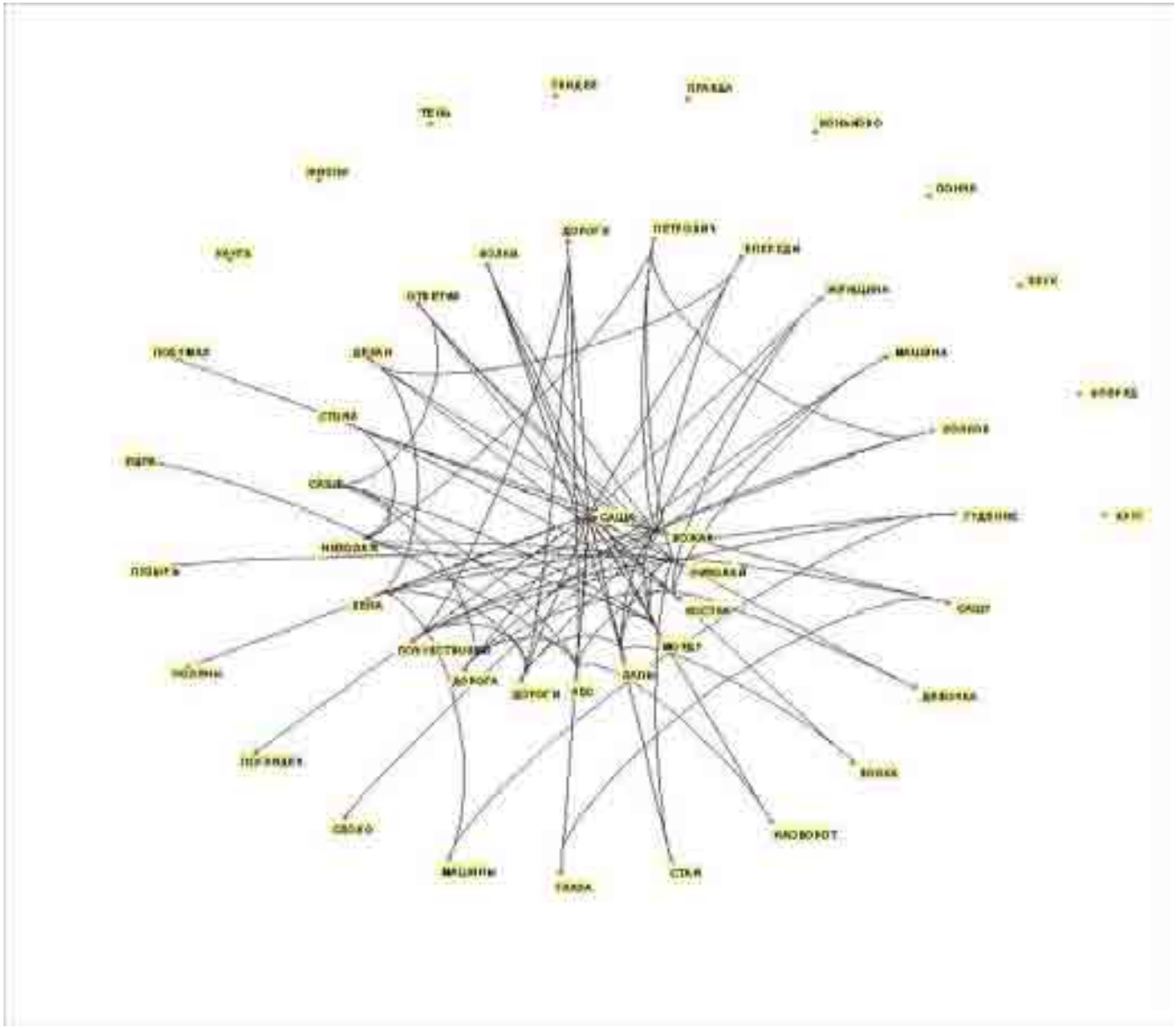
Визуализация КГГВ

В. Астафьев, «Ловля пескарей в Грузии»



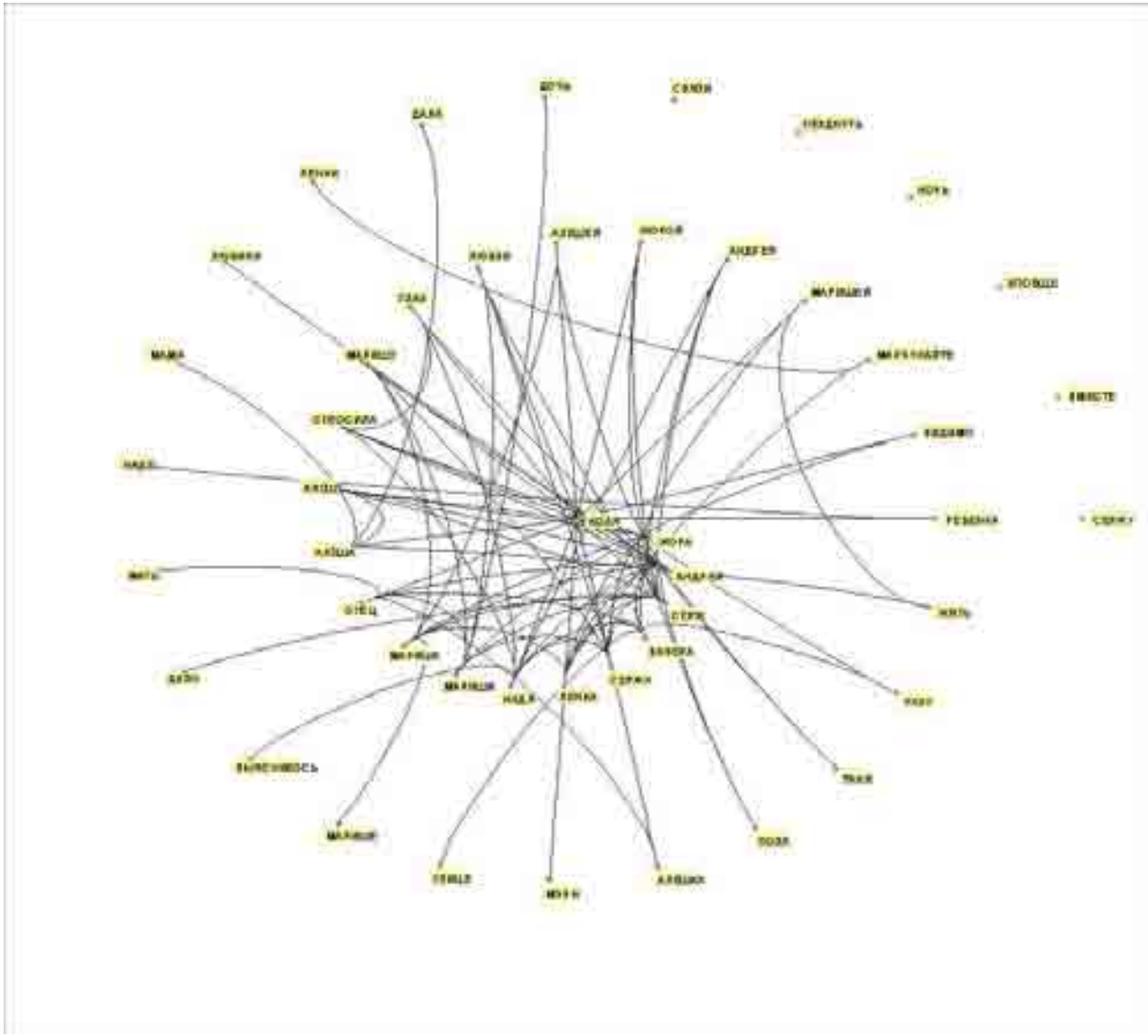
Визуализация КГГВ

В. Пелевин, «Проблема верволка в средней полосе»



Визуализация КГГВ

Л. Петрушевская, «Свой круг»



Некоторые результаты

- Предложен алгоритм компактифицированного графа горизонтальной видимости (КГГВ).
- На основе дисперсионных оценок слов, метода TFIDF и КГГВ, построены сети слов различных текстов.
- 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.
- 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.

Перспективы

Изучение свойства выявленных одновременно информационно и структурно важных лексических единиц как опорных слов для различных текстовых жанров, документов, представленных на разных языках. Это позволит:

- Формировать «более осмысленные» информационные портреты текстов;
- Выполнять автоматическое реферирование текстов;
- Формировать цепочки подобных документов, объединять тематические сюжеты, используя выявленные слова в качестве дескрипторов;
- Выявлять возможное содержательное дублирование документов, представленных на различных языках (необходимо дальнейшее исследование инвариантности опорных слов, для исходных текстов и их переводов);
- Составлять словари опорных слов, формировать тезаурусы и онтологии предметных областей.

Спасибо за внимание!

Д.В. ЛАНДЭ

dwlande@gmail.com