УДК 004.7

Д.В. Ландэ, А.А. Снарский, В.Г. Путятин

Институт проблем регистрации информации НАН Украины

(ул. Шпака, 2, 03113 Киев, Украина, тел.(044) 4542163)

Построение терминологической сети предметной области

Описывается методика построения сетей иерархий терминов на основе анализа массива текстов по выбранной проблематике (живучести технических и информационных систем). Методика базируется на применении компактифицированных графов горизонтальной видимости для терминов – отдельных слов, биграмм и триграмм, а также установлении связей между терминами. Построена и исследована сеть языка, сформированная на основе полных текстов научных статей по проблематике живучести

Ключевые слова: *языковая сеть, граф горизонтальной видимости, сеть иерархии терминов, живучесть, текстовый корпус*

Введение

Для решения актуальных задач построения онтологий (детальной формализации выбранных областей знаний, отрасли) требуется проведение комплексных исследований, определенным этапом которых является построение так называемых словарных номенклатур, тезаурусов, предметных словарей. Эффективный автоматический отбор отдельных терминов для таких конструкций – не решенная окончательно задача, а проблема установления связей, автоматического построения сетей из таких терминов до сих пор остается открытой.

Как терминологическую основу для формирования соответствующей терминологических онтологии предлагается использовать сеть естественной иерархии терминов, которая базируется на информационно-значимых элементах текста, опорных словах и словосочетаниях [1], методология выявления которых приведена в [2]. Опорные слова и словосочетания, как правило, выбираются с учетом такого их свойства, как дескриминантная сила. Вместе с тем, одного этого свойства часто оказывается недостаточно для отражения содержания предметной области. Иногда слова с низкой дескриминантной силой, в частности, наиболее частотные слова из выбранной предметной области (например, слова «Живучесть», «Система», «Надежность» в корпусе текстов по проблемам живучести) оказываются важнейшими для рассматриваемой задачи.

Постановка задачи

В данной работе решается задача формирования терминологической сетевой основы для построения онтологий предметной области, для чего рассматриваются принципы формирования сети естественных иерархий терминов (СЕИТ), базирующейся на контенте научных статей выбранной направленности. "Естественность" иерархий терминов в этом случае понимается как отказ при формировании терминологической сети от специальных методов смыслового анализа, ограничение фактически статистическим анализом текстов. Связи в такой сети определяются естественным взаимным положением слов и словосочетаний, которые экстрагируются из текстов. Такая сеть, создаваемая полностью автоматически, может рассматриваться как основа для дальнейшего автоматизированного формирования терминологической онтологии с участием экспертов.

Методика формирования СЕИТ

Методика формирования сети естественных иерархий терминов, рассматриваемая в данной работе, предусматривает реализацию последовательности шагов, охватывающей предварительную обработку исходного текста, определение и сортировку терминов, выбор из них необходимого количества наиболее весомых, непосредственное построение СЕИТ и ее отображение [3]. Рассмотрим эти шаги более подробно.

1. На первом этапе формируется исходный текстовый корпус. Как пример такого корпуса рассматриваются полные тексты научных статей, посвященных проблематике живучести в информационных и технических системах, представленных на русском языке. В состав текстового корпуса было включено около 50 научных статей общим объемом около 1 млн. символов. Предварительная обработка такого текстового корпуса предусматривала выделение фрагментов текстов (отдельных статей, абзацев, предложений, слов), исключение нетекстовых символов, отсечение флективных окончаний (стемминг).
2. На втором этапе каждому отдельному термину из текста (слову, биграмме или триграмме) ставится в соответствие оценка их "дескриминантная сила", а именно TFIDF, которая в каноническом виде равна произведению частоты соответствующего термина (Term Frequency) в фрагменте текста на двоичный логарифм от величины, обратной к количеству фрагментов текста, в которых этот термин встретился (Inverse Document Frequency) [4].

   Для последовательностей терминов и их весовых значений по TFIDF строятся компактифицированные графы горизонтальной видимости (CHVG) и выполняется переопределение весовых значений слов уже по этому алгоритму. Данная процедура позволяет учитывать в дальнейшем кроме терминов с большой дескриминантной силой также высокочастотные термины, которые имеют большое значение для общей тематики. В соответствии с [3], сеть слов с использованием алгоритма горизонтальной видимости строится также 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 (TFIDF). На втором этапе строится традиционный граф горизонтальной видимости [5]. 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. На третьем, заключительном этапе, сеть компактифицируется. Все узлы с одним и тем же словом объединяются в один узел, связи таких узлов также объединяются. В качестве весовых оценок отдельных слов в дальнейшем используются степени соответствующих им узлов в CHVG. После этого все термины текста сортируются по убыванию рассчитанных весовых значений соответствующих узлов CHVG. Дальнейшему анализу не подлежат термины из так называемого стоп-словаря, являющиеся важными для связности текста, но не несущие информационной нагрузки. Это, как правило, фиксированный набор служебных слов. Используемый в рамках данной работы стоп-словарь был построен на основе различных стоп-словарей, представленных в доступном виде на веб-ресурсах:

   http://code.google.com/p/stop-words/source/browse/trunk/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;
   http://trac.mysvn.ru/punbb/punbb/browser/trunk/Russian/stopwords.txt.

   Экспертным методом определяется необходимый размер СЕИТ (число N), после чего выбирается соответствующее количество единичных слов, биграмм и триграмм (всего N+N+N элементов) с наибольшими весовыми значениями по CHVG.
3. Из отобранных терминов строятся сети естественных иерархий терминов, в которых как узлы рассматриваются сами термины, а связи соответствуют вхождением одних терминов в другие. На рис. 1 проиллюстрирован принцип построения связей СЕИТ.

Различные геометрические фигуры на этой иллюстрации соответствуют различным словам. Первой строке соответствует выбранное множество единичных слов, второму – множество биграмм, а третьему – множество триграмм. Если единичное слово входит в биграмму или триграмму, или биграмма входит в триграмму, образуется связь, которая обозначается стрелкой. Множество узлов, которым соответствуют термины, и связи образуют трехуровневую сеть естественной иерархии терминов.

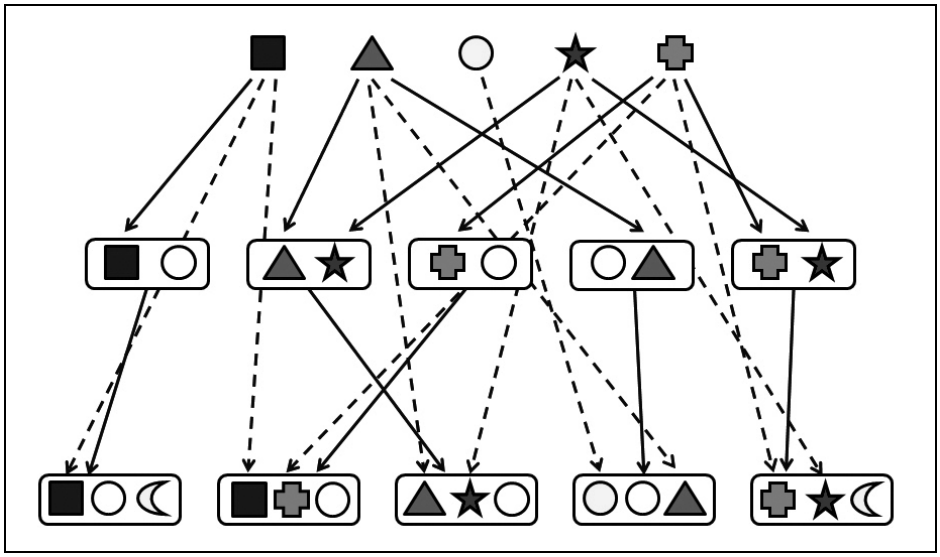

Рис. 1 – Трехуровневая сеть естественной иерархии терминов

На последнем этапе формирования СЕИТ осуществляется ее отображение программными средствами анализа и визуализации графов. Для загрузки сетей естественных иерархий терминов в базы данным формируется матрица инцидентности общепринятого формата csv.

На рис. 2 представлена небольшая сеть естественной иерархии терминов размером 20+20+20, которая визуализирована средствами системы Gephi (https://gephi.org/).

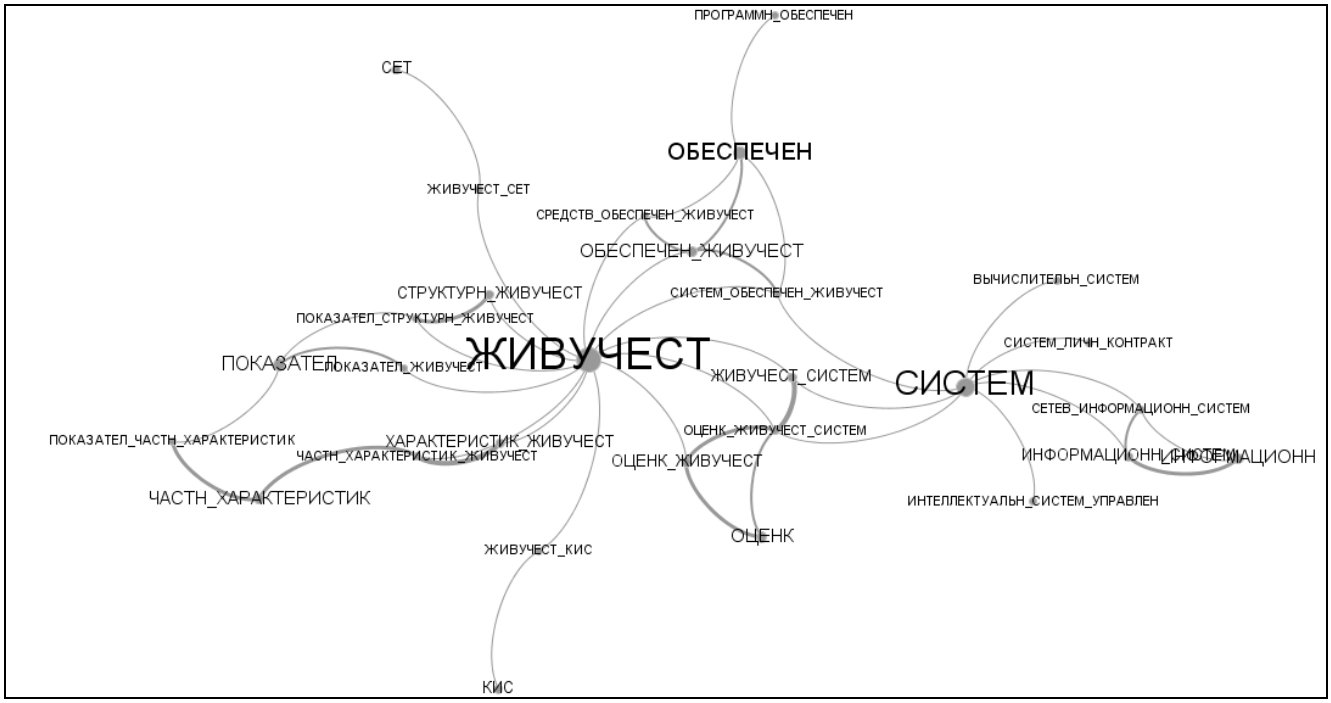

Рис. 2 – Визуализация связного фрагмента СЕИТ размером 20+20+20

На рис. 3 приведены отдельные фрагменты более крупной сети естественной иерархии терминов размером 200+200+200.

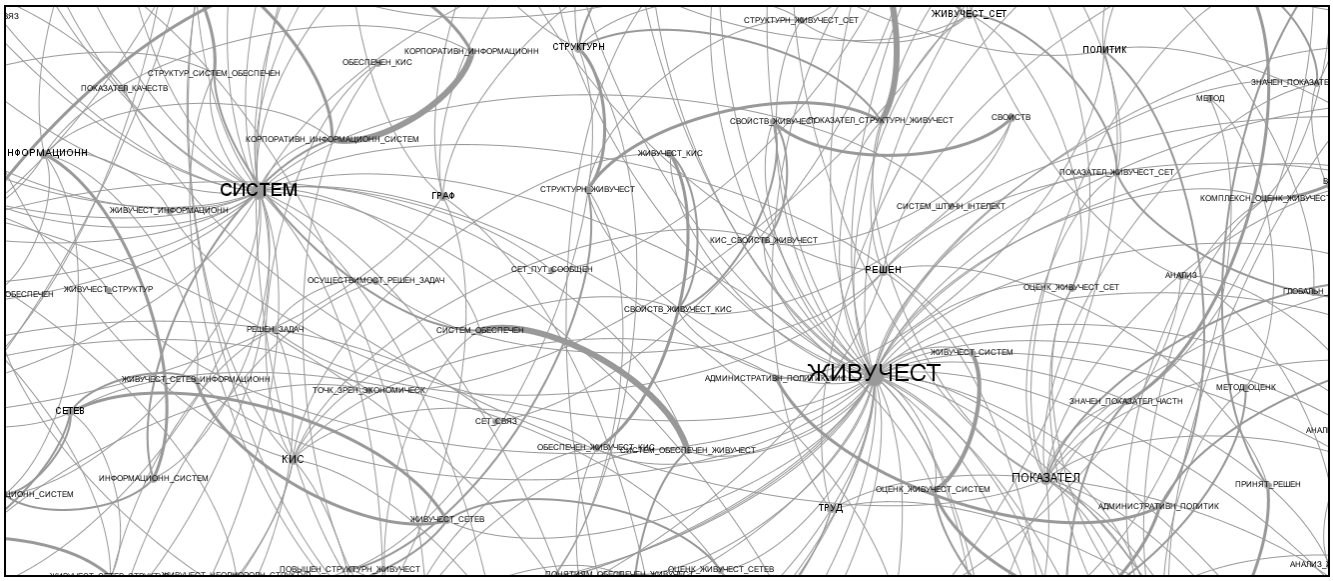

Рис. 3 – Фрагмент СЕИТ размером 200+200+200

Параметры узлов СЕИТ

Для построенных сетей естественных иерархий терминов различных размеров по выбранному тексту было определено распределение исходящих степеней узлов, которое оказалось близким к степенному ($p(k) = Ck^{\alpha}$), т.е. эти сети являются безмасштабными. Оказалось, что коэффициент α для сетей различных размеров (от 20+20+20 до 500+500+500) составляет от 2,1 до 2,3, что вполне соответствует сетям языка (Language Networks) [6].

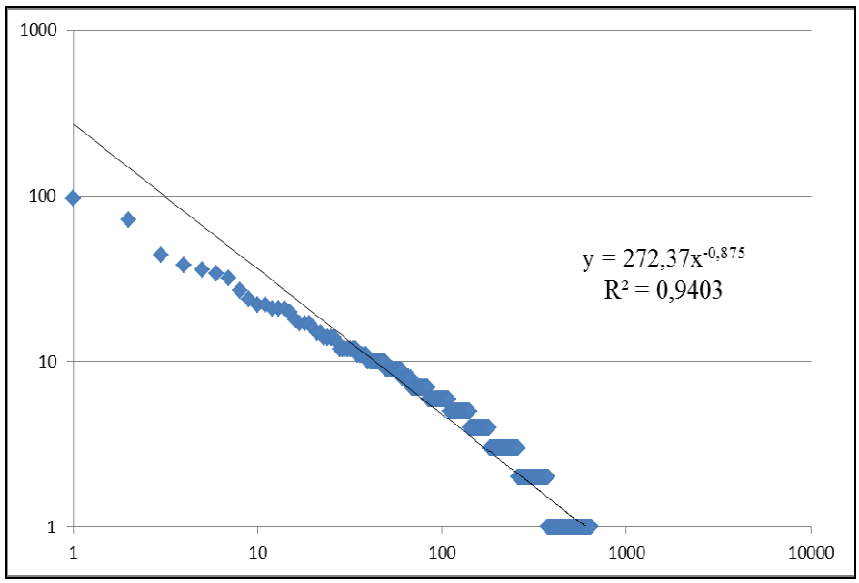

Рис. 4. Ранговое распределение степеней узлов в логарифмической шкале (по оси абсцисс – порядковый номер узла, по оси ординат – степень узла)

В соответствии с предложенным алгоритмом, максимальное количество входных связей для узлов данной сети составляет 5 (для узлов из одного слова – 0 входящих связей, для узлов

из 2 слов – максимально 2 связи, для узлов из 3 слов – максимально 5 связей – три связи от отдельных слов и две от пар слов).

Ранжирование узлов в СЕИТ возможно также по свойствам, обуславливаемым сетевой структурой, сылками. Например, для определения авторитетности узла как слова – источника порождения словосочетаний или как составного термина, состоящего из отдельных важных слов, можно анализировать СЕИТ, выбирая при этом наиболее важных «авторов» или «хабов». Для решения этой задачи предлагается использовать известный алгоритм ранжирования веб-страниц, основанных на связях, HITS (hyperlink induced topic search), предложенный Дж. Клейнбергом [7].

Алгоритм HITS обеспечивает выбор из информационного массива лучших «авторов» (узлов, на которые введут ссылки) и «посредников» (узлов, от которых идут ссылки цитирования). Понятно, что в нашем случае термин является хорошим посредником, если от него идут связи на важные словосочетания, и наоборот, термин (словосочетание) является хорошим автором, если на него ведут связи от важных авторов. В соответствии с алгоритмом HITS в нашем случае для каждого узла сети v_j рекурсивно вычисляется его значимость как автора $a(v_j)$ и посредника $h(v_j)$ по формулам:

$$a(v_j) = \sum_i h(v_i);\ h(v_j) = \sum_i a(v_i).$$

В данных формулах суммирование производится по всем узлам, которые ссылаются (или на которые ссылаются – во второй формуле) на данный узел.

Наиболее интересными с семантической точки зрения в рассматриваемой СЕИТ оказались узлы с наибольшим значением авторства и посредничества. В таблице 1. приведены термины, соответствующие таким узлам.

Таблица 1. Термины, соответствующие узлам с наибольшим авторством и посредничеством

№	Термины с наибольшим значением посредничества	Термины с наибольшим значением авторства
1	ЖИВУЧЕСТЬ	ОБЕСПЕЧЕНИЕ ЖИВУЧЕСТИ КИС
2	ПОКАЗАТЕЛЬ	СИСТЕМА ОБЕСПЕЧЕНИЯ ЖИВУЧЕСТИ
3	ОБЕСПЕЧЕНИЕ	ЗНАЧЕНИЕ ПОКАЗАТЕЛЯ ЖИВУЧЕСТИ
4	ОБЕСПЕЧЕНИЕ ЖИВУЧЕСТИ	СРЕДСТВА ОБЕСПЕЧЕНИЯ ЖИВУЧЕСТИ
5	ОЦЕНКА	ОЦЕНКА ЖИВУЧЕСТИ СЕТИ
6	СИСТЕМА	ЗАДАЧА ОБЕСПЕЧЕНИЯ ЖИВУЧЕСТИ
7	ЗНАЧЕНИЕ	ВЫЧИСЛЕНИЕ ПОКАЗАТЕЛЯ ЖИВУЧЕСТИ
8	СЕТЬ	ОЦЕНКА ЖИВУЧЕСТИ СИСТЕМ
9	ПОКАЗАТЕЛЬ ЖИВУЧЕСТИ	ВЕРОЯТНОСТНЫЙ ПОКАЗАТЕЛЬ ЖИВУЧЕСТИ
10	ОЦЕНКА ЖИВУЧЕСТИ	ПОКАЗАТЕЛЬ СТРУКТУРНОЙ ЖИВУЧЕСТИ
11	КИС	СВОЙСТВО ЖИВУЧЕСТИ КИС
12	ХАРАКТЕРИСТИКА	КРИТЕРИЙ ОЦЕНКИ ЖИВУЧЕСТИ
13	ЖИВУЧЕСТЬ КИС	ПОКАЗАТЕЛЬ ЖИВУЧЕСТИ
14	ЖИВУЧЕСТЬ СЕТИ	ЧАСТНАЯ ХАРАКТЕРИСТИКА ЖИВУЧЕСТИ
15	СТРУКТУРА	ЗНАЧЕНИЕ УРОВНЯ ЖИВУЧЕСТИ

Представления об информационной значимости наборов терминов для построения СЕИТ, степени их важности для отражения смысла научного текста 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» [8].

Ассоциативные связи на основе СЕИТ

Рассматриваемые в предложенной модели СЕИТ связи являются направленными и могут рассматриваться как отношения «общее-частное» при построении общей онтологии. Вместе с тем, построенная сеть СЕИТ может рассматриваться как основа для формирования других связей между ее узлами. Так, например, если два термина-узла данной сети порождают третий термин, который также входит в данную сеть, то можно считать, что такие термины связаны ассоциативной связью. На рис. 5 приведен фрагмент сети СЕИТ, где жирными кривыми выделены такие связи, например, слово «решение» ассоциативно связано со словом «задача», что объясняется наличием общего узла-термина «решение задач».

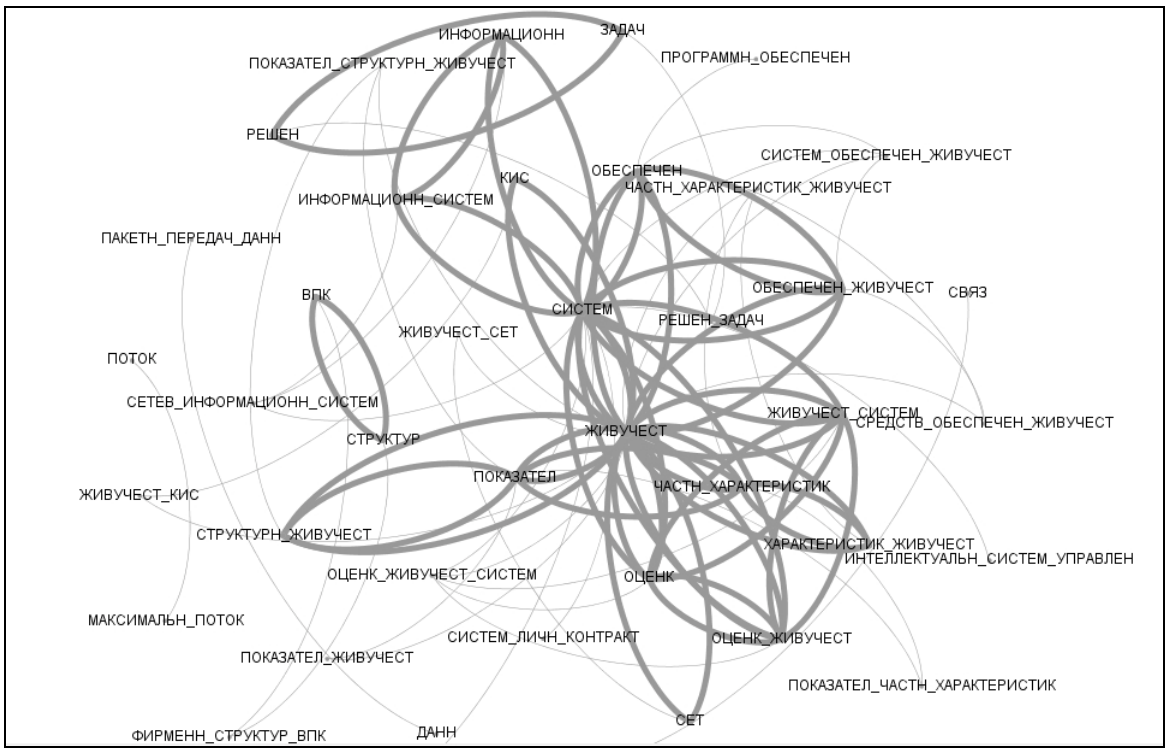

Рис. 5. Фрагмент СЕИТ размером 20+20+20 с ассоциативными связями

Выводы

Таким образом, в данной работе:

- Предложен алгоритм построения сетей естественных иерархий терминов на основе анализа текстов.
- На основании этого алгоритма по текстам научных статей по проблематике живучести информационных и технических сетей построена сеть естественной иерархии терминов.
- Сеть естественных иерархий терминов оказалась скейл-фри по исходящим связям.
- Предложен алгоритм построения ассоциативных связей между терминами в СЕИТ.
- Предложено использование алгоритма HITS для выбора наиболее важних элементов в СЕИТ.
- Выбраны программные средства визуализации СЕИТ.

Сеть языка, построенную с помощью предложенной методики, можно использовать в качестве базы для построения онтологии предметной области (в рассмотренном примере – по проблематике живучести), использовать на практике в качестве готового к применению средства навигации в информационных массивах, а также для организации контекстных подсказок пользователям информационно-поисковых систем.

1. *Yagunova E., D. Lande D.* Dynamic Frequency Features as the Basis for the Structural Description of Diverse Linguistic Objects // CEUR Workshop Proceedings. Proceedings of the 14th All-Russian Scientific Conference "Digital libraries: Advanced Methods and Technologies, Digital Collections" Pereslavl-Zalessky, Russia, October 15-18, 2012. – P. 150-159.
2. *Lande D.V., Snarskii A.A., Yagunova E.V., Pronoza E.V.* The Use of Horizontal Visibility Graphs to Identify the Words that Define the Informational Structure of a Text // 12th Mexican International Conference on Artificial Intelligence, 2013. – P. 209-215.
3. *Lande D.V.* Building of Networks of Natural Hierarchies of Terms Based on Analysis of Texts Corpora // E-preprint ArXiv 1405.6068
4. Salton G., McGill M.J. Introduction to Modern Information Retrieval. – New York : McGraw-Hill, 1983. – 448 p.
5. *Luque B., Lacasa L., Ballesteros F., Luque J.* Horizontal visibility graphs: Exact results for random time series // Phys. Review E, 2009. – P. 046103-1 – 046103-11.
6. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. – М.: МИЭМ, 2011. – 272 с.
7. *Kleinberg J.* Authoritative sources in a hyperlinked environment // In Processing of ACM-SIAM Symposium on Discrete Algorithms, 1998, 46(5):604-632.
8. *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. – Пермь, 2010. – Вып. 1. – С. 85-91.