УДК 519.6 doi:10.15421/321604

В.А. Громов, А.М. Мигрина, А.А. Новаковский
Днепропетровский национальный университет имени Олеся Гончара

СТАТИСТИЧЕСКИЕ ХАРАКТЕРИСТИКИ ГРАФОВ ОДНОВРЕМЕННОГО ПОЯВЛЕНИЯ СЛОВ В ТЕКСТАХ ЕСТЕСТВЕННЫХ ЯЗЫКОВ: УСТНАЯ И ПИСЬМЕННАЯ РЕЧЬ

В рамках теории сложных сетей проведен сравнительный анализ статистических характеристик графов одновременного появления слов в текстах русского и английского языков для устной и письменной речи (в качестве репрезентативной выборки для первой использованы сообщения социальной сети «Twitter»), установлен тип распределения для данных характеристик – степенное – и оценены параметры указанных распределений.

У межах теорії складних мереж проведено порівняльний аналіз статистичних характеристик графів одночасної появи слів у текстах російської та англійської мов для усного та письмового мовлення (як репрезентативну вибірку для першого брали повідомлення з соціальної мережі «Twitter»), установлено тип розподілу для даних характеристик – степеневий – та оцінено параметри визначених розподілів.

The paper provides (in the framework of complex networks theory) a comparative analysis for statistical characteristics of co-occurrence graphs for texts of English and Russian languages both for oral and written speech (the former is substituted by set of twits), ascertains probability distribution type for these characteristics (actually, power ones) and estimates parameters of these distributions.

Ключевые слова: теория сложных сетей, графов одновременного появления слов, язык и речь, степенные распределения.

Введение. Сложность естественного языка как системы обусловливает необходимость его анализа с самых различных точек зрения. В настоящее время здесь, наряду с классическими подходами, предполагающими анализ языковых явлений в рамках языковедческих и литературоведческих парадигм, разработан подход, связанный с анализом естественных языков в рамках теории сложных систем. В частности, одним из бурноразвивающихся направлений является анализ графов естественных языков как сложных сетей [1].

В настоящей работе для русского и английского языков проведена идентификация распределений основных статистических характеристик графов одновременного появления слов (co-occurrence graphs) в текстах естественного языка; для каждого из языков исследование проводилось отдельно на двух выборках: первая – это корпус текстов литературного языка, вторая – множество сообщений в социальной сети «Twitter». Мы полагаем, что первая выборка репрезентативна для письменной речи, вторая – much bolder assumption – для устной речи; в пределе – это различие между языком и речью (langue et parole) [10]. Отметим, что в настоящем исследовании базовой единицей является текст, и граф одновременного появления слов характеризует именно одновременное появление слов в одном тексте.

Выбор языков – английский и русский – обусловлен, во-первых, наличием достаточно обширных выборок текстов (как корпуса литературных текстов, так и твитов), во-вторых, серьёзным отличием их грамматических строёв: русский язык является флективным языком синтетического типа и обладает свободным порядком слов в предложении, английский – напротив, является аналитическим языком и имеет строгий порядок слов в предложении; при этом оба языка относятся к группе индоевропейских.

Дальнейшее изложение структурировано следующим образом: в первом разделе приведён обзор существующих исследований, посвящённых статистическим характеристикам графов естественных языков, в третьем – описание метода проверки соответствия полученных выборок степенному распределению. Четвёртый раздел посвящён описанию и анализу полученных результатов; наконец, в последнем разделе сформулированы выводы.

Обзор литературы. Языковые сети (language networks) т. е. графы, связанные с описанием естественных языков, могут быть подразделены на несколько категорий в зависимости от того, какой аспект языка – синтаксический, семантический или теоретико-текстовой – подвергается исследованию [12].

На синтаксическом уровне при построении языковых сетей обычно опираются на различного рода синтаксические конструкции [5] и сеть, в этом случае, отражает синтаксические зависимости между словами предложения. В работе [12] в рамках теории сложных сетей исследуются характеристики графа синтаксических связей и графа совместного появления слов в текстах; подвергается исследованию вопрос о влиянии универсальных законов организации и эволюции сложных сетей, являющихся носителями языковых процессов, на эволюцию естественных языков.

Семантический уровень предполагает анализ связей между семантическими концепциями, связанными со словами и\или словосочетаниями [6].

На теоретико-текстовом уровне мы встречаемся с графами совместного появления слов в текстах [5]; в некоторых случаях используются ориентированные графы [11], возможен учёт синтаксических отношений и близости между лексическими понятиями (lexicalized concepts).

В работе [12] представлены характеристики языковых сетей всех трёх типов как сложных сетей, полученные различными исследователями на различных выборках: авторы подчёркивают, что все три типа сетей демонстрируют свойства малого мира (small word).

Наличие у сети указанного свойства обычно сопряжено с тем, что статистические характеристики данной сети подчиняются степенному закону распределения. В работах [7; 8] приведены обзоры природных и социальных явлений, демонстрирующих степенные законы распределения; часть указанных обзорных работ посвящена языковым системам. Здесь следует также отметить, что оценка параметров степенных распределений и проверка статистических гипотез о соответствии наблюдаемых данных степенному распределению (goodness-of-fit tests) является отдельной непростой задачей математической статистики, требующей для своего решения методов, отличных от методов, принятых в гауссовой статистике [3]. В настоящей работе при проверке такого рода статистических гипотез мы опирались на работу [2].

Постановка задачи. Рассмотрим выборки литературных текстов и сообщений социальной сети «Twitter» (аналог устной речи) для английского и русского языков.

Взвешенный неориентированный граф совместного появления слов в тексте (для данной выборки) будем дефинировать как $\overline{G} = (V, E)$, где V – непустое множество слов языка (словарь), а $E = \{w_{ij}\}$, где вес ребра w_{ij} – вес ребра, равный числу раз, которое пара слов, соответствующая вершинам i и j, одновременно встречалась в текстах выборки. Если $w_{ij} = 0$, то ребро между вершинами отсутствует. Вместе с графом $\overline{G}$ рассмотрим его невзвешенную реплику G.

Для каждой из рассмотренных в работе характеристик была построена эмпирическая функция распределения, вычислена оценка математического ожидания и дисперсии и оценены параметры расспределения.

В данной работе были вычислены следующие характеристики графа G [1]:

- Степень вершины (degree of a vertex): определяется как количество рёбер, инцидентных данной вершине.

- Вес ребра (weight of edge): число раз, которое пара слов, соответствующая вершинам инцидентных ребру, встречается в тексте.

- Длина наименьшего пути (shortest pass length).

- Нагрузка вершины (stress centrality): количество наименьших путей, которые проходят через данную вершину.

- Кластерный коэффициент (clustering coefficient): определяется как соотношение числа связей между соседями вершины на их максимально возможное количество указанных связей.

- Коэффициент ассортативности (assortativity coefficient): сеть называют ассортативной, если вершины, которые имеют большую степень, (хабы) предпочтительно связаны с такими же вершинами с большими степенями. Если, напротив, такие вершины предпочтительно связаны с вершинами с малыми степенями (связаны между собой через цепочки из вершин с малой степенью), то сеть называется дисассортативной. Определяется как коэффициент парной корреляции Пирсона между степенями вершин, инцидентных одному и тому же ребру.

- **Коэффициент клуба богатых** (rich-club coefficient): показывает, насколько часто хабы в сети взаимосвязаны и образуют хорошо связанный подграф. Определяется как соотношение числа связей между узлами со степенью, большей, чем k на их максимально возможное количество указанных связей.

Математические выражения для всех вышеуказанных характеристик приведены в [1]. Поскольку используемые выборки были существенно неодинаковых размеров, то для проведения сравнения характеристик, полученных для разных выборок, были вычислены их относительные аналоги, определяемые как отношения полученного значения величины к соответствующему значению для полносвязного графа с таким же числом вершин. В дальнейшем изложении данные величины мы будем называть относительными, неотнесённые – абсолютными.

Метод оценки параметров степенного распределения. Для вычисления оценки показателя степени степенного распределения и проверки статистической гипотезы о соответствии данных указанному распределению использован метод, предложенный в [2]: будем предполагать, что распре-

деление является дискретным степенным, тогда с учётом условий нормировки для функции распределения имеем

$$F(x) = \frac{\varsigma(\alpha, x)}{\varsigma(\alpha, x_{\min})},$$

где $\varsigma(\alpha, x_{\min})$ – дзета-функция Гурвица

$$\varsigma(\alpha, x_{\min}) = \sum_{n=0}^{\infty} (x + x_{\min})^{-\alpha}.$$

Для отыскания показателя распределения α использован метод максимального правдоподобия, согласно которому

$$\frac{\varsigma'(\alpha, x_{\min})}{\varsigma(\alpha, x_{\min})} = -\frac{1}{n}\sum_{n=1}^{\infty} \ln x_i.$$

Для решения трансцендентного уравнения использовался генетический алгоритм.

Статистические характеристики. Рассмотрены выборки, полученные из корпусов литературных текстов английского и русского языков; объёмы выборок составили 3000 и 2200 текстов соответственно. Сообщения социальной сети «Twitter» отвечали периоду с 07.09.2014 по 07.10.2014 (для англоязычных твитов) и периоду с 07.09.2014 по 07.10.2014 (для русскоязычных); объёмы выборок здесь составили порядка $2 \cdot 10^7$ и $1.2 \cdot 10^7$ соответственно.

Таблица 1

Абсолютные значения статистических характеристик для графов совместного появления слов в текстах

	Twitter Англ.	Текст Англ.	Twitter Рус.	Текст Рус.
Длина наим. пути	3176 (1416)	1972 (856)	1736 (766)	1846 (643)
Нагрузка	1627 (548)	1548 (672)	1762 (592)	1487 (601)
Степень вершин	438 (396)	317 (201)	376 (278)	409 (269)
Вес ребра	4212 (2122)	3827 (1889)	3921 (1945)	3624 (1741)
Кластерный коэффициент	0,537 (0,37)	0,619 (0,218)	0,485 (0,278)	0,583 (0,265)
Коэффициент ассортативности	0,24 (0,26)	0,38 (0,06)	0,32 (0,144)	0,296 (0,126)
Коэффициент клуба богатых	0,27 (0,431)	0,18 (0,37)	0,388 (0,541)	0,372 (0,487)

Все построенные эмпирические функции распределения имеют ярко выраженный степенной характер (рис. 1).

В табл. 1 представлены полученные абсолютные характеристики для всех выборок (в скобках указано значение среднеквадратического уклонения), в табл. 2 – относительные, в табл. 3 – значения показателя степени α.

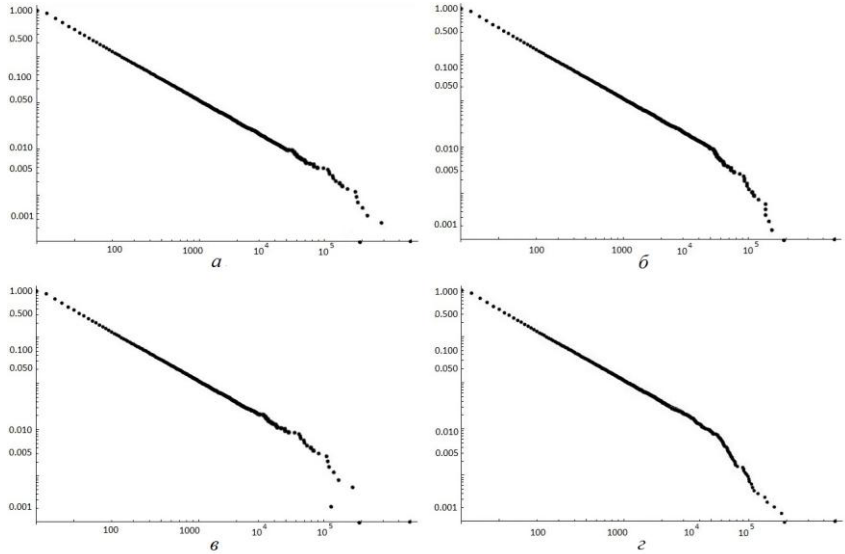

Рис. 1. Функции распределения степеней вершин (в двойном логарифмическом масштабе) для (а) – лит. текстов русского языка, (б) – сообщений «Twitter» русского языка, (в) – лит. текстов английского языка, (г) – сообщений «Twitter» английского языка

Таблица 2

Относительные значения статистических характеристик для графов совместного появления слов в текстах

	Twitter Англ.	Текст Англ.	Twitter Рус.	Текст Рус.
Длина наим. пути	0,125	0,12	0,141	0,125
Нагрузка	0,047	0,066	0,086	0,069
Степ. вершин	0,013	0,015	0,02	0,019

Окончание табл. 2

Относительные значения статистических характеристик для графов совместного появления слов в текстах

Вес ребра $*10^{-4}$	1,22	1,63	2,13	2,36
Кластерный коэффициент	0,537	0,619	0,485	0,583
Коэффициент ассортативн.	-0,24	0,38	-0,32	0,296
Коэффициент клуба богатых	0,27	0,18	0,388	0,372

Таблица 3

Значения показателей степени

	Twitter Англ.	Текст Англ.	Twitter Рус.	Текст Рус.
Нагрузка	2,339	2,301	2,407	2,277
Степ. вершин	2,552	2,47	2,502	2,24
Вес ребра	2,543	2,378	2,482	2,296

Выводы. Анализ полученных данных позволяет сформулировать следующие выводы.

1. Основные статистические характеристики графов одновременного появления слов в обоих языках в устной и письменной речи подчиняются степенным распределениям с показателем $\alpha \sim 2.0 \div 2.5$.
2. Графы одновременного появления слов в текстах письменного языка (как для русского, так и английского языков) представляют собой ассортативные сложные сети, графы одновременного появления слов в сообщениях устного языка – дисассортативные.
3. Анализ остальных статистических характеристик позволяет сделать вывод, что характеристики, полученные для русской письменной и устной речи ближе друг другу, чем к аналогичным характеристикам для письменной и устной речи английского языка.

Библиографические ссылки

1. **Barrat, A.** Dynamical Processes on Complex Networks [Text] / A. Barrat, M. Barthelemy, A. Vespignani. – Cambridge, 2008. – 348 p.
2. **Clauset, A.** Power-Law Distributions In Empirical Data [Text] / A. Clauset, C.R. Shalizi, M.E.J. Newman // SIAM Rev. – 2002. – № 51 (4). – P. 661–703.

3. **Embrechts, P.** Modelling Extremal Events for Insurance and Finance [Text] / P. Embrechts, C. Klueppelberg, T. Mikosch. – N.-Y., 1997. – 644 p.

4. **Ferrer i Cancho, R.** Patterns in syntactic dependency networks [Text] / R. Ferrer i Cancho, R. Koehler, R.V. Sole // Phys. Rev. E. – 2004. – № 69. – P. 327–367.

5. **Ferrer i Cancho, R.** The Small-World of Human Language [Text] / R. Ferrer i Cancho, R.V. Sole // Proc. R. Soc. Lond. Series B. – 2001. – № 268. – P. 2261–2266.

6. **Holanda, A.J.** Thesaurus as a complex network [Text] / A.J. Holanda, I. Torres Pisa, O. Kinouchi, A. Souto Martinez, E. Seron Ruiz // Physica A. – 2004. – № 344. – P. 530–536.

7. **Kello, C.T.** (2010) Scaling laws in cognitive sciences [Text] / C.T. Kello, G.D.A. Brown, R. Ferrer-i-Cancho, J.G. Holden, K. Linkenkaer-Hansen, T. Rhodes, G.C. Van Orden // Trends in Cognitive Sciences. – 2010. – Vol. 14, № 5. – P. 511–531.

8. **Kwapień, J.** Physical approach to complex systems [Text] / J. Kwapień, S. Drożdż // Physics Reports. – 2012. – Vol. 515. – P. 115–226.

9. **Resnick, S.I.** Heavy-Tail Phenomena: Probabilistic and Statistical Modeling [Text] / S.I. Resnick. – N.-Y., 2007. – 404 p.

10. **Saussure, de F.** Course in general linguistics (3rd ed.) [Text] / F. de Saussure. – Chicago, 1986. – 320 p.

11. **Sole, R.V.** Global Organization of Scale-F ree Language Networks [Text] / R.V. Sole, A. Corominas, B. Valverde // SFI Working Paper. – 2005. – Vol. 1. – P. 5–12.

12. **Sole, R.V.** Language Networks: their structure, function and evolution [Text] / R.V. Sole, B.C. Murtra, S. Valverde, L. Steels // Complexity. – 2010. – Vol. 15, № 6. – P. 20–26.

Надійшла до редколегії 06.07.2016