МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА НА РЕПУБЛИКА БЪЛГАРИЯ
КИРИЛО-МЕТОДИЕВСКИ НАУЧЕН ЦЕНТЪР ПРИ БЪЛГАРСКА АКАДЕМИЯ НА НАУКИТЕ
ИЖЕВСКИЙ ГОСУДАРСТВЕННЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ ИМ. М. Т. КАЛАШНИКОВА
НАУЧНОЕ СООБЩЕСТВО “ПИСЬМЕННОЕ НАСЛЕДИЕ”
DIGITAL MEDIEVALIST SCHOLARLY COMMUNITY
ФОНДАЦИЯ „УСТОЙЧИВО РАЗВИТИЕ НА БЪЛГАРИЯ “

Писменото наследство и информационните технологии

El’Manuscript–2014

Материали от V международна научна конференция
Варна, 15–20 септември 2014 г.

София · Ижевск
2014

Сборникът е издаден с финансовата подкрепа на Министерството на образованието и науката на Република България по процедура за подкрепа на международни научни форуми.

Отговорни редактори: проф. дфн В. А. Баранов
доц. д-р В. Желязкова
д-р А. М. Лаврентьев

Редактори: Нели Ганчева, Веселка Желязкова (български текст)
О. В. Зуга, В. А. Баранов (руски текст)
Кевин Хокинс (Kevin Hawkins) (английски текст)

Писменото наследство и информационните технологии [Текст] : материали от V международна науч. конф. (Варна, 15–20 септември 2014 г.) / отг. ред. В. А. Баранов, В. Желязкова, А. М. Лаврентьев. — София ; Ижевск, 2014. — 448 с.

Сборникът съдържа материали от конференция, посветена на разработването и създаването на съвременни средства за съхраняване, описване, обработка, анализ и публикуване на ръкописни и старопечатни книжовни паметници и исторически извори, а също и на въпросите за подготвянето на електронни ресурси в областта на хуманитаристиката и тяхното използване в научните изследвания и преподаването.

ISBN 978–954–9787–25–2

Экспериментальные исследования текста Словаря русского языка XI–XVII вв.

А. Е. Дубашов, Ю. Н. Филиппович

Инфо-когнитивные технологии, историческая лексикография, словарь русского языка XI–XVII вв., автоматизированные системы научных исследований, закон Ципфа, автоматическая обработка текстов

Experimental Studies of Text of the Dictionary of the Russian Language of the 11th–17th Centuries

Aleksey Dubashov, Yuriy Philippovich

This paper presents a new method for estimating the dynamics of the appearance of new word forms in the text of the Dictionary of the Russian Language of the 11th–17th Centuries. Two statistical experiments with the text of citations from the 2nd, 3rd and 4th issues of the dictionary are described. Estimates were calculated according to Zipf's Law. Two factors were evaluated: the total number of new words and the number of new words for the current letter.

Целью исследования является разработка метода получения оценок динамики появления новых словоформ в тексте Словаря русского языка XI–XVII вв. (далее — Словарь или СЛРЯ). Исследования проводились на 2, 3 и 4 выпусках, использовался текст только из цитатного материала [СЛРЯ].

Описание эксперимента 1

Текст Словаря, занесенный предварительно в базу данных, разбивается на определенные промежутки (выборки). В этих промежутках подсчитывается количество новых слов на интересующую букву, которое заносится в таблицу. После этого строятся графики зависимости: количества новых слов от порядкового номера промежутка, логарифмической зависимости количества новых слов от порядкового номера промежутка, график динамики словарного запаса от объема. На основе экспериментальных данных определяется функция и параметры закона проявления новых слов в тексте Словаря. На основе выявленного закона рассчитывается (предсказывается) объем словарного запаса следующего выпуска. Здесь на основе второго выпуска определяется объем четвертого.

Исходные данные. Предмет исследования: цитатный материал словаря. Выпуски: 2, 4. Исследуемые буквы: *А – Я*. Шаг разбивки текста (промежуток выборки): 100 цитат.

Полученные экспериментальные данные представляют собой статистические распределения появления новых слов и динамики пополнения Словаря по выборкам 2 и 2+4 выпусков.

Расчет. В основе метода предсказания объема запаса Словаря лежит гипотеза, что появление новых слов происходит по гиперболическому закону Ципфа-Мандельброта [Филиппович, Прохоров 2002]:

$$f(r) = \frac{i(k,r)}{k} = pr^{-b} \quad (1)$$

$$f(r) = \frac{i(k,r)}{k} = p(r+v)^{-b} \quad (2)$$

r — объем текста, который может быть представлен в различных единицах (количество слов в тексте, предложений, цитат и т.д.), k — общий объем словарного запаса на данной выборке, а p, v и b — параметры закона.

Определив параметры этого закона, можно рассчитывать объем словарного запаса в зависимости от объема текста. Используя статистики по второму тому словаря, были определены параметры закона появления новых словоформ: b = 0,2349 и p = 0,0304.

Объем словарного запаса второго тома, полученного экспериментальным путем, равен K_2 = 38659 слов, при N_2 = 69 замерах в выборке (1 замер = 100 цитатам), а рассчитанный на основе формулы (1) S_2 — 39264. Погрешность результата составила 1,54 %.

Далее были рассчитаны характеристики словарного запаса обоих томов Словаря на основе второго тома: K_{2+4} = 77034, N_{2+4} = 170, S_{2+4} = 78273. Погрешность результата составила 1,58 %, она ухудшилась на 0,04 % при предсказании, на один выпуск вперед.

Столь высокая погрешность на первом этапе, где даже о предсказании еще не идет речь, связана с аппроксимацией. Ее можно улучшить, используя для аппроксимации не закон Ципфа (1), а уточнение этого закона Мандельбротом (2). Здесь же была предпринята попытка улучшить результат предсказания за счет подведения рассчитанного закона к экспериментально полученному результату объема словарного запаса.

Оценка эффективности. Для оценки эффективности предсказания было проведено сравнение результата предсказания описанным методом с методом, предложенным Ю. К. Орловым [Орлов 1978а]. Характеристика словарного запаса двух томов рассчитывалась методом последовательных приближений [Орлов 1978б]. Ошибка в прогнозе метода Орлова составила 1,756 % от экспериментально полученных данных.

Описание эксперимента 2

Отличие второго эксперимента от первого состояло в том, что исследовались три выпуска вместо двух (2, 3 и 4) и рассматривались только слова, начинающиеся на букву *У*. Экспериментальные данные аналогичны первому эксперименту.

Расчет: На основе 4-го тома была сделана попытка предсказать словарный запас всех трех томов на букву У.

Сначала был произведен расчет по словоформам. По статистике 4-го тома словаря были определены параметры закона появления новых словоформ на букву *У*, а затем объем словарного запаса на букву *У*: b = 0,2052, p = 0,02013; K_4 = 1229, N_4 = 102; K_{2+3+4} = 2183, N_{2+3+4} = 198; S_{2+3+4} = 2082. Погрешность результата составила 4,62 %. По сравнению с предыдущим экспериментом, она ухудшилась в 100 раз, хотя также не выходит за рамки разумного.

Затем был произведены расчет по словам (лемматизированным словоформам) и оценка эффективности. Результаты: b = 0,58, p = 0,06021; K_4 = 480, N_4 = 102; K_{2+3+4} = 688, N_{2+3+4} = 198; S_{2+3+4} = 634. Погрешность результата составила 7,82 %. По сравнению с предыдущим экспериментом, она ухудшилась еще в два раза. Ошибка в прогнозе метода Орлова составила 3,393 % от экспериментально полученных данных.

Это говорит о том, что не удалось с достаточной точностью определить закон появления новых слов для конкретной буквы. Видно, что точность прогноза падает при снижении количества рассматриваемых слов в выборке. Это дает основания предполагать, что при увеличении выборки точность определения закона и, соответственно, точность предсказания будет расти.

Подводя итог результатам экспериментов, укажем преимущества и недостатки предложенного метода. *Преимуществами* являются высокая точность предсказания словарного запаса на больших объемах текста и теоретическая обоснованность предсказываемого результата, а *недостатками* — необходимость знания динамики появления новых слов, для чего требуется проведение динамического исследования текста, что в свою очередь подразумевает использование специализированных программных средств, а также существенно больший объем материала.

Литература

Орлов 1978а — *Орлов Ю. К.* Статистическое моделирование речевых потоков. / Под ред. Р. Г. Пиотровского. М.; Л., 1978. (Вопросы кибернетики. Выпуск 41. Статистика речи и автоматический анализ текста).

Орлов 1978б — *Орлов Ю. К.* Модель частотной структуры лексики. Исследования в области вычислительной лингвистики. Вып. 2. Ч. 1. М, 1978.

СЛРЯ — *Словарь русского языка XI–XVII вв.* Вып.1–29– . М.: Наука, 1975–2013– .

Филиппович, Прохоров 2002 — Филиппович Ю. Н., Прохоров А. В. Семантика информационных технологий: Опыты словарно-тезаурусного описания. М.: МГУП, 2002. 368 с.