ATLANTIS PRESS *Advances in Social Science, Education and Humanities Research (ASSEHR), volume 122*
45th International Philological Conference (IPC 2017)

К вопросу о мерах лексического сходства частотных словарей

Гребенников Александр Олегович

Санкт-Петербургский государственный университет (СПбГУ), Санкт-Петербург, Россия; Университет ИТМО, Санкт-Петербург, Россия

On the measures of lexical similarity between frequency dictionaries

Alexander Grebennikov

Saint Petersburg State University (SPSU), St. Petersburg, Russia;
ITMO University, St. Petersburg, Russia

Аннотация

Анализируется использование существующих мер лексического сходства частотных словарей в стилеметрических исследованиях на материале серии частотных словарей русских писателей. Показан их недостаточный потенциал при решении стилеразличительных задач. Также рассматривается численный метод выделения лексических маркеров, использование которого, напротив, представляется перспективным.

Abstract

Two measures of lexical similarity between frequency dictionaries are analyzed. The examples are drawn from frequency dictionaries of short stories by Russian writers. Their potential in stylometric researches is shown.

Ключевые слова: лингвостатистика, частотный словарь, силеметрия.

Keywords: statistical linguistics, frequency dictionary, stylometry.

(1) Последние десятилетия ознаменованы ростом количества частотных словарей как языка в целом, так и различных подъязыков. Они могут активно использоваться в решении лингвистических задач различного типа, в частности –играть большую рольпри исследовании вопросов стилистики вообще и «авторской» стилистики, в частности. Представляется интересным проанализировать использование существующих мер лексического сходства частотных словарей с точки зрения их возможного использования в стилеметрических исследованиях. Учеными Института русского языка им В. В. Виноградова РАН предлагается использовать в качестве простейшей меры лексического сходства двух частотных словарей следующую формулу:

$$C_{xy} = \sum \min\{px_i, py_i\}, \qquad (1)$$

где $\Sigma\ px_i$ и $\Sigma\ py_i$ – вероятности, т.е., в нашем случае, относительные частоты единиц словаря; *min* — минимальная из двух частот [Шайкевич А. Я.].

Показатель приобретает значение 1 при сравнении частотных словарей одного и того же текста; он близок к 0 при сравнении частотных словарей использующих разную графику.

(2) Работоспособность формулы в стилеметрических исследованиях предлагается проверить на материале частотных словарей рассказов выдающихся русских писателей (А. И. Куприна, А. П. Чехова, Л. Н. Андреева), создаваемых на кафедре математической лингвистики СПбГУ [Частотный словарь … , 2006; Частотный словарь … , 1999; Частотный словарь … , 2003]. Данные словарные материалы обеспечивают хорошую базу для сравненияв силу одинакового объёма (около 200 000 словоупотреблений), единства жанра текстов, лежащих в их основе,и времени их создания, принципов отбора материала и составления. В результате получены следующие значения: для пары словарей А. П. Чехова и Л. Н. Андреева значение *C* составило 0,96; для пары А. П. Чехов – А. И. Куприн – 0,941; для пары Л. Н. Андреев – А. И. Куприн – 0,925. Полученные значения подтверждают принадлежность исследуемых тестов к одному языку (что представляется само собой разумеющимся), однако стилеразличительный потенциал формулы если и имеет место, то выражен весьма незначительно. Возможно, дальнейшие исследования, учитывающие, например, динамику роста объема словаряс ростом объема выборки, а также аппроксимация полученных результатов окажутся полезными с точки зрения уточнения полученных результатов.

(3) Поскольку весьма вероятно, что наибольший вклад в общую сумму вносят самые частые слова, то формула 1 может быть модифицирована следующим образом:

$$C_{xy} = \frac{\sum \min\{px_i, py_i\}}{0{,}5\left(\sum px_i + \sum py_i\right)},$$

где $\Sigma\ px_i$ и $\Sigma\ py_i$ вычисляются для выбранной зоны рангового частного словаря (первых *i* единиц).

Тогда, например, для первых ста наиболее частых лексем значение *C* составляет: для пары словарей Чехов – Андреев – 0,948; для пары Чехов – Куприн – 0,945; для пары Андреев – Куприн – 0,930.

Если же мы рассмотрим первую тысячу наиболее частых лексем,то полученные данные приобретают следующий вид:для пары Чехов – Андреев – 0,957; для пары Чехов – Куприн – 0,959; для пары Андреев – Куприн – 0,945.

Мы видим некоторую корректировку полученных результатов, обратно пропорциональную количеству охваченных лексем, при этом их возможности для стилеразличения по-прежнему минимальны и требуют вышеупомянутого дальнейшего анализа с привлечением расширенного диапазона данных.

(4) Одновременно в последние десятилетия, с приходом в языкознание вообще и авторскую лексикографию, в частности, понятий «языковая картина мира», «концепт» и т.п. возникла проблемасоотнесения единиц словаря языка писателя с основными темами его творчества, индивидуально-авторской картиной мира. Возникают и разрабатываются

понятия «ключевых слов», «текстем», «идиоглосс» и др. [Поцепня Д. М.] Для анализа реальных лексических различий между единицами в частотных словарях предлагается использовать следующую формулу:

$$S = \frac{(x - m - 1)}{\sqrt{m}}$$

где x – частота слова, m – математическое ожидание этой частоты [Шайкевич]. Данная формула справедлива при сравнении словарей принадлежащих одному корпусу с числом подкорпусов более двух, что выполняется в случае используемого нами материала (корпус – русский рассказ рубежа веков, подкорпруса – .массивы текстов одного автора).

Предлагается считать все слова, S которых превысил некоторый порог (например, $S > 3$), лексическими маркерами. Тогда, количество лексических маркеров и их доля в анализируемых словарях составит:для словаря Чехова –895 лексем (включая слова с частотой 27); для словаря Андреева – 931 лексема (включая слова с частотой 25); для словаря Куприна –также 931 лексема (включая слова с частотой 25).

Если же считать $S > 2$, тогда количество лексических маркеров составит соответственно: для словаря Чехова – 1069 лексем (включая слова с частотой 23); для словаря Андреева – 1032 лексемы (также включая слова с частотой 23); для словаря Куприна – 1637 лексем (включая слова с частотой 22).При этом изменение в абсолютном количестве охватываемых лексем сопровождается лишь незначительным сдвигом в частотным спектре.

(5) Теперь возможно проверить, сколькоиз этих маркеров (для удобства и единообразия сравнения – изверхней тысячи лексем, расположенных в порядке убывания частот) являются индивидуальными, т.е. присутствующими только в одном из пары сравниваемых словарей и отсутствующими в другом. Для пары словарей Чехов – Андреев насчитывается 65 подобных лексем; для пары Чехов – Куприн – 67; для пары Андреев – Куприн – 62 лексемы. Подобный подход представляется целесообразным использовать в качестве одного из методов при выделении идеологически значимых слов в словаре языкаписателя. Его применение, также, может быть в дальнейшем сопоставлено с выделенными ранее показателями лексической концентрации и рассеяния [Гребенников А. О., 1998].

(6) Дальнейший содержательный и сопоставительный анализ выделенных единиц представляется перспективным в свете наших недавних исследований стилеразличительного потенциала частотных словарей языка автора [Гребенников, 2015].

Литература

Гребенников А.О. Исследование устойчивости лексико-статистических характеристик текста. Дис. … канд. филол. наук. СПб, 1998. 210 с.

Гребенников А.О. Индивидуально-авторский характер различных зон распределения в частотных словарях языка писателя// Структурная и прикладная лингвистика. 2015. № 11. С. 100-110.

Поцепня Д.М. Словари языка писателя. // Лексикография русского языка: учеб. для вузов РФ / под ред. Д.М. Поцепни. СПб:Филологический фак-т СПбГУ, 2013.704 с. – С. 583-654.

Частотный словарь рассказов А.И.Куприна. Авт.-сост. А.О. Гребенников; под ред. Г.Я.Мартыненко. СПб: Изд-во С.-Петерб. ун-та, 2006. 551 с.

Частотный словарь рассказов А.П.Чехова. Авт.-сост. А.О.Гребенников; под ред. Г.Я.Мартыненко. СПб: Изд-во С.-Петерб. ун-та, 1999. 172 с.

Частотный словарь рассказов Л.Н.Андреева. Авт.-сост. А.О. Гребенников; под ред. Г.Я.Мартыненко. СПб: Изд-во С.-Петерб. ун-та, 2003. 396 с.

Шайкевич А.Я. Меры лексического сходства частотных словарей// Труды международной научной конференции «Корпусная лингвистика -2015». 22–26 июня 2015 г. 2015». – СПб: С.-Петербургский гос. университет, Филологический факультет, 2015. 456 с. — С. 422-429.

References