МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
Филологический факультет
Кафедра русского языка
Лаборатория общей и сибирской лексикографии
ИНСТИТУТ РУССКОГО ЯЗЫКА ИМ. В.В. ВИНОГРАДОВА РАН
ГОСУДАРСТВЕННЫЙ ИНСТИТУТ РУССКОГО ЯЗЫКА ИМ. А.С. ПУШКИНА
ПОД ЭГИДОЙ РОССИЙСКОГО ОБЩЕСТВА ПРЕПОДАВАТЕЛЕЙ
РУССКОГО ЯЗЫКА И ЛИТЕРАТУРЫ (РОПРЯЛ)

ЛЕКСИКОГРАФИЯ ЦИФРОВОЙ ЭПОХИ

Сборник материалов Международного симпозиума
(24–25 сентября 2021 г.)

Томск
Издательство Томского государственного университета
2021

Литература

1. *Симони П.* Старинные сборники русских пословиц, поговорок, загадок и проч. XVII–XIX столетий / собрал и приготовил к печати Павел Симони. СПб., 1899. I–XIX + 216 с.
2. *Словарь русских народных говоров* / под ред. Ф.П. Филина и Ф.П. Сороколетова; гл. ред. С. А. Мызников. Вып. 1–51. Л.; СПб., 1965–2019. (издание продолжается).

DOI: 10.17223/978-5-907442-19-1-2021-10

Н.Д. Голев
Nikolay D. Golev

Статистические данные поисковых систем Рунета как лексикографический источник Runet search engine statistics as a lexicographic source

Кемеровский государственный университет, Кемерово –
Kemerovo State University, Kemerovo
ngolevd@mail.ru

Аннотация. В докладе представляются результаты и отмечаются перспективы трех направлений использования компьютерных технологий в лингвистике (и в том числе - лексикографии), разрабатываемых докладчиком в течение многих лет: 1) использование статистических данных поисковых систем Интернета, 2) использование поисковых систем интернета как источника материала, 3) использование обратного машинного перевода.

Summary. The report presents the results and outlines the prospects for three areas of the use of computer technologies in linguistics (including lexicography), developed by the speaker over the years: 1) use of statistical data from Internet search engines, 2) use of Internet search engines as a source of material, 3) use reverse machine translation.

Ключевые слова: поисковые системы интернета, частотность слова, концептуализация лексики, дискурсивный словарь

Keywords: Internet search engines, word frequency, conceptualization of vocabulary, discourse dictionary

Идея использования компьютерной техники и цифровых технологий для решения различных задач лингвистики и лексикографии в частности не нова, уже не новы и сами термины: компьютерная лингвистика, компьютерная (электронная) лексикография. Компьютер – эффективное техническое под-

споре лексикографа, в виртуальном пространстве существует множество словарей, ставятся новые задачи, решение которых невозможно представить без электронной техники, например, создание сводного словаря русской диалектной лексики. Мы не претендуем на обзор и оценку имеющихся результатов в области компьютерной лексикографии. Свою задачу видим в демонстрации возможностей компьютера в сфере инновационной лексикографии – выдвижению и – в ряде случаев – первичной разработки новых лексикографических идей разного уровня и масштаба. Многие из них являются результатом нашей практической лексикографической работы.

Статистические данные поисковых систем Интернета (СДПСИ) по широте, глубине и разнообразию выборки многократно превосходят показатели частотности, которые дают традиционные частотные словари, ср. показатели ряда первых слов из словаря Л.Н. Засориной (1 млн. выборки) с показателями страниц «поисковика» Google, на которых отмечено запрашиваемое слово: *абажур* – 24 / 23.500.000. Частотный словарь, основанный на больших величинах, фиксируемых Интернетом, способен отразить и качественные параметры функционирования лексики. По сути, СДПСИ – это стихийный частотный словарь современной и новейшей русской речи, способный зафиксировать изменения даже в небольшие периоды ее развития.

Полагаем, что специфика СДПСИ заключается не только в количественных характеристиках. Они (данные) специфичны также и в качественном отношении, например, с точки их способности отражать разнообразие дискурсивного пространства. Так, стихийный характер формирования СДСПИ объективно фиксирует сегодняшние социально-экономические отношения, в соответствии с которыми в иллюстрациях СДСПСИ в первую очередь отражаются важные для рынка дискурсивные практики (реклама, социально значимая информация и т.п.), в отличие от НКРЯ, ориентированного на художественную речь. Например: *снегурка* – 32 / 407.000 (первые страницы – фиксируют мультфильмы, песни, фасоны одежды, детские центры и т.п., в названии которых используется слово Снегурка), *снегирёк* – 0 / 124.000 (детский сад, дом отдыха, экобольница, сорт абрикоса). Считаем важным заметить, что эти дискурсы не периферия русской речи, напротив, это речевая реальность, формирующая корпус актуальных прецедентных фраз, а конечном итоге – и языковое сознание молодого поколения носителей русского языка. «Художественноцентризм», характерный для лексикографии прошлого века, сама жизнь отодвигает на второй план, и лексикография новейшего времени, на наш взгляд, не может не реагировать на это обстоятельство.

Частотность слова является не только дискурсивно, но и когнитивно значимой характеристикой слова: чувство частотности рядового носителя языка закономерно верифицируется данными эксперимента. Совокупный коэффициент частотности является отвлеченной (надындивидуальной) устойчивой характеристикой слова, которую объективно фиксирует как языковая интуиция, так и компьютер, отражающий тем самым языковую практику, а – через ее опосредование – и языковое сознание. В этом плане количественный коэффициент является вполне антропоцентричным и по праву должен входить в число характеристик слова в будущем Всеохватном словаре русского языка.

В числе когнитивно значимых аспектов частотности на первом плане, на наш взгляд, находится способность количественного коэффициента слова отражать силу его концептного содержания. В частности, отмеченное выше активное использование слов в рекламном дискурсе (*снегурка, снегирек*) отражает его концептный потенциал на настоящий момент времени.

Считаем, что представленные свойства СДСПИ позволяют использовать их для корректировки позиции слов в синонимических словарях. Например, в «Словаре синонимов русского языка» под ред. А.П. Евгеньевой (1970) приводится синонимический ряд в следующем порядке: *аномалия, отклонение, неправильность, ненормальность, странность*. СДСПИ «корректирует» его таким образом: *отклонение* – 11 000 000, *аномалия* – 3 240 000, *странность* – 1 520, *неправильность* – 412 000, *ненормальность* – 175 000. Синонимический ряд *актер, артист* СДСПИ подтверждают с незначительным перевесом в пользу леммы *актер*: 48 000 000 / 40 000 000.

Мы уже неоднократно обращались к СДСПИ для решения различных вопросов, связанных с описанием словообразовательной системы русского языка в аспекте ее лексической реализации. Главное дериватологическое свойство СДСПИ – их способность фиксировать окказиональную и потенциальную лексику, очень важную также и для деривационной лексикографии. Исследователь и лексикограф могут мысленно предполагать и/или конструировать любое слово и верифицировать свою версию с помощью СДСПИ.

Продолжая линию возможности корректировки с помощью СДСПИ, укажем на возможность верификации данных, имеющихся в диалектных, исторических, жаргонных словарях.