**В. Д. Соловьев, В. В. Бочкарев, Л.А. Янда**

**V.D. Solovyev, V.V. Bochkarev, L.A. Janda**

**Динамика частот употребления семантичЕски близких слов[[1]](#footnote-1)**

**SEMANTICALLY CLOSE WORDS**

**FREQUENCY DYNAMICS**

Абстракт. В работе рассматривается новая модель установления семантической тождественности слов через сопоставление частот их словоупотребления на большом временном интервале. Частоты словоупотребления берутся по корпусу Google Books Ngram. Модель применяется к проблематике аспектуальных пар. Показано, что корреляция частот словоупотребления аспектуальных пар с префиксальным образованием глагола совершенного вида выше, чем аспектуальных пар с суффиксальным образованием глагола несовершенного вида. Это неожиданный результат, т.к. исходя из современной теории русского вида, следовало ожидать обратного. Полученный результат перепроверен несколькими способами.

Ключевые слова. Русский язык, аспект, частота словоупотребления, корпус n-грамм

Abstract. We propose a new model for establishing the semantic identity of words by comparing the frequencies of their usage over a large time interval. The word frequencies are taken from the Google Books Ngram corpus. The model is applied to the problematic of aspectual pairs. It is shown that the correlation of the word frequencies for the aspectual pairs with the prefixed formation of the perfective verb is higher than that of the aspectual pairs with the suffixal formation of the secondary imperfective. This is an unexpected result, because from the modern theory of the Russian aspect, one would expect the opposite. The obtained result was rechecked in several ways.

Keywords. Russian language, aspect, word frequency, corpus of n-gram

1. **введение**

Корпус Google Books Ngram позволяет анализировать и сопоставлять динамику частот словоупотреблений и снабжен удобной визуализацией в виде графиков частот (https://books. google.com/ngrams/). В основополагающей работе [Michel et al. 2011] и ряде последующих показаны возможности использования Google Books Ngram для изучения эволюции, как отдельных слов, так и всего лексикона языка в целом, а также культурных трендов в обществе.

Кроме собственно частот словоупотреблений, весьма информативным является форма графиков. Естественным является предположение, что частоты употребления семантически идентичных слов (например, словоизменение внутри одной леммы: *читать* – *читал*) под влиянием внешних факторов меняются схожим образом, т.е. графики частот имеют схожую форму.

Аспектуальная система русского языка находится в процессе становления, многие ее элементы получили неоднозначное освещение в литературе и вызвали большие споры. Неясным остается характер суффиксального образования вторичных имперфективов и префиксального образования перфективов с точки зрения словообразовательного или словоизменительного характера этих процессов. В данной работе изучается степень синхронности изменения частот слов в аспектуальных парах, как один из подходов к рассматриваемой проблеме. Графики частот трактуются как временные ряды. Мы используем терминологию по аспекту, принятую в [Janda et al. 2013].

1. **Данные и методы**

Мы сравниваем корреляции частот словоупотребления внутри следующих групп слов – глаголов: 1) случайно выбранные (в качестве baseline), 2) словоизменительные формы внутри одной леммы, 3) аспектуальные пар с префиксальным образованием перфективов, 4) аспектуальных пар с суффиксальным образованием вторичных имперфективов. Словоизменительная парадигма была взята по ресурсу OpenCorpora (<http://opencorpora.org/>), основанному на словаре Зализняка. Случайная выборка осуществлялась из 6947 глаголов, все формы полной парадигмы которых присутствуют в Google Books Ngram. Рассматриваются частоты словоупотребления за период с 1928 по 2008 гг. Исследование начато с 1928 г., чтобы избежать влияния старинных форм слов, использовавшихся до реформы орфографии 1917 г. Аспектуальные пары брались по базе данных, составленной под руководством Л. Янды (http://emptyprefixes.uit.no/). Считаются коэффициенты корреляции временных рядов по Пирсону. Для контроля вычислялась также корреляция по Спирмену, давшая аналогичные результаты. Следует отметить, что в настоящее время не существует общепринятого и полностью адекватного способа определения степени похожести временных рядов [Koplenig 2015]. Мы используем корреляцию по Пирсону, следую большинству работ.

1. **Результаты и обсуждение**

Получены следующие результаты. В первом и втором случаях коэффициенты корреляции по Пирсону оказались равны 0.0404 и 0.1864 соответственно. Для первой группы глаголов корреляции быть не должно, результат ожидаем. Для второй корреляция оказалась мала (что несколько странно и требует дальнейших исследований), но статистически значима.

Для третьей и четвертой группы коэффициенты корреляции равны соответственно 0.3020 и 0.2435. Это соотношение также неожиданно. Традиционная точка зрения [Маслов 1984], разделяемая многими авторами, гласит, что суффиксальное образование вторичных имперфективов – это словоизменение, а префиксальное образование перфективов – словообразование. Если это так, то это должно приводить к большей семантической схожести и более высокому коэффициенту корреляции перфективов с вторичными имперфективами. Полученный результат можно трактовать как аргумент в пользу того, что оба способа образования аспектуальных пар с точки зрения этого противопоставления имеют примерно одинаковый статус, что согласуется с [Filip 1999] и рядом других публикаций.

Вероятно, одним из влияющих факторов является то, что при образовании вторичного имперфектива часто привносится дополнительное значение многократности действия (как в рассматривавшемся Ю.Д. Апресяном [Апресян 1995] примере: *пить – выпить – выпивать*). Другим фактором является выборка глаголов: в единственной существующей аспектуальной базе для русского языка содержатся только тройки базовый имперфектив – естественный перфектив – вторичный имперфектив. Эта база данных не включает пары, в которых вторичный имперфектив образуется из специализированного перфектива. Учет таких пар может привести к увеличению соответствующего коэффициента корреляции.

В целом не очень высокие значения коэффициентов корреляции легко объяснить многозначностью слов. В настоящее время у нас нет хорошего способа автоматически выделять различные значения слов и получать для них различные графики.

Для нивелирования трудностей с многозначностью мы выделили 22 глагола, не имеющих совершенно различных значений (*делать, работать, казаться, играть, просить, верить, ставить, звать, звонить, хранить, рисовать, прятать, влечь, плевать, жечь, растить, мерзнуть, жрать, копать, нюхать, красть, щупать*). Рассмотрен более узкий временной интервал – с 1950, с устоявшимся современным русским языком и большим количеством изданных книг по сравнению с предшествующим периодом.

Для выбранных 22 глаголов среднее значение корреляции внутри словоизменения оказалось равно 0,568, а корреляция между базовым имперфективом и естественным перфективом – 0,758. Далее для этих слов были выбраны некоторые специализированные перфективы (*переделать* и т.д.), подсчитаны коэффициенты корреляции между ними и соответствующими вторичными имперфективами. Среднее значение оказалось равно 0,633. Таким образом, и после устранения явной многозначности и учета специализированных перфективов все равно корреляция между базовым имперфективом и естественным перфективом оказалась выше.

Этот результат указывает на то, что префиксальный способ образования естественного перфектива не в меньшей степени сохраняет значение исходного слова, чем суффиксальное образование вторичных имперфективов.

1. **Заключение**

Google Books Ngram дает в руки исследователей совершенно новую методологию и позволяет по новому взглянуть на аспектуальную систему русского языка. Полученные результаты являются аргументом в пользу известной общей гипотезы Дж. Бэбби о том, что словоизменение и словообразование не полярные противоположности, а, скорее, точки на континуальной школе лексико-грамматических процессов.

**Литература**

1. Апресян Ю. Д. Трактовка избыточных аспектуальных парадигм в толковом словаре // Апресян Ю. Д. Избранные труды. Т.II: Интегральное описание языка и системная лексикография. М., 1995.

2. Маслов Ю.С. Очерки по аспектологии. Л.: ЛГУ. 1984.

3. Filip H. Aspect, Eventuality Types, and Nominal Reference. London. 1999.

4. Janda L. A. et al. Why Russian Aspectual prefixes aren’t empty. Prefixes As Verb Classifiers. Bloomington: Slavika Publishers Indiana University. 2013.

5. Koplenig A. Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. Digital Scholarship in the Humanities. Oxford University Press. 2015. doi:10.1093/llc/fqv030.

6. Michel J. et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*. 14 January 2011. Vol. 331. № 6014. pp. 176-182.

**References**

1. Apresjan Yu. D. Traktovka izbytochnyh aspektualnyh paradigm v tolkovom slovare [Interpretation of redundant of aspectual paradigms in the explanatory dictionary]. Apresjan Yu. D. Izbrannye trudy. T. II. Integralnoe opisanie yazyka i sistemnaya leksikografiya [Selected papers. Vol. II: Integral description of the language and systemic lexicography]. Moscow. 1995. pp. 102-113.

2. Maslov Yu. S. Ocherki po aspektologii [Essays on Aspectology]. Leningrad: LGU. 1984

3. Filip H. Aspect, Eventuality Types, and Nominal Reference. London. 1999.

4. Janda L. A. et al. Why Russian Aspectual prefixes aren’t empty. Prefixes As Verb Classifiers. Bloomington: Slavika Publishers Indiana University. 2013.

5. Koplenig A. Why the quantitative analysis of diachronic corpora that does not consider the temporal aspect of time-series can lead to wrong conclusions. Digital Scholarship in the Humanities. Oxford University Press. 2015. doi:10.1093/llc/fqv030.

6. Michel J. et al. Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*. 14 January 2011. Vol. 331. № 6014. pp. 176-182.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Соловьев Валерий Дмитриевич**

Казанский федеральный университет (Россия)

**Solovyev Valery**

Kazan Federal University (Russia)

***E-mail: maki.solovyev@mail.ru***

**Бочкарев Владимир Владимирович**

Казанский федеральный университет (Россия)

**Bochkarev Vladimir**

Kazan Federal University (Russia)

***E-mail: vbochkarev@mail.ru***

**Янда Лора**

Университет Тромсё (Норвегия)

**Janda Laura**

University of Tromsø (Norway)

***E-mail: laura.janda@uit.no***

1. Работа выполнена при поддержке РФФИ, грант № 15-06-07402 [↑](#footnote-ref-1)