УДК 811.112'373:811.112'25

Богданова Наталия Альбертовна
кандидат филологических наук
г. Минск, Беларусь
Белорусский государственный
университет иностранных языков
e-mail: albertowna@mail.ru

Natallia Bahdanava
PhD in Philology
Minsk, Belarus
Belarusian State University
of Foreign Languages
e-mail: albertowna@mail.ru

РОЛЬ КОЭФФИЦИЕНТА ДАЙСА И КОЛЛОКАЦИОННОЙ ЧАСТОТНОСТИ ДЛЯ РАСЧЕТА ИНДЕКСА СЕМАНТИЧЕСКОЙ ПРОСОДИИ СЛОВА

В статье рассматривается корпусной подход к описанию лексического значения слова, основанный на явлении совстречаемости лексических единиц. Описывается возможность изучения семантической просодии слова – его тональной окраски, вызванной особенностями сочетания узлового слова с коллокатами различной тональности. На основе коэффициента Дайса и коллокационной частности лексических единиц предлагается формула расчета индекса семантической просодии слова (SPI), описывается способ подсчета веса каждого коллоката, взвешенного вклада тональности каждого коллоката и возможности нейтрализации больших коллокационных частот. Предложенная формула апробируется на немецких глаголах с каузальным значением: verursachen, bewirken, auslösen, описывается методика определения тональности слова и рассчитывается индекс семантической просодии для каждого глагола.

Ключевые слова: семантическая просодия; коллокация; узловое слово; коллокат; тональность.

THE ROLE OF THE DICE COEFFICIENT AND COLLOCATION FREQUENCY IN CALCULATING A WORD'S SEMANTIC PROSODY INDEX

The article discusses a corpus-based approach to describing the lexical meaning of a word based on the phenomenon of co-occurrence of lexical units. It describes the possibility of studying the semantic prosody of a word – its tonal colouring caused by the peculiarities of combining a key word with collocations of different tonality. Based on the Dice coefficient and the collocation specificity of lexical units, a formula for calculating the semantic prosody index (SPI) of a word is proposed, and a method for calculating the weight of each collocation, the weighted contribution of the tonality of each collocation, and the possibility of neutralising high collocation frequencies is described. The proposed formula is tested on German verbs with a causal meaning: verursachen, bewirken, auslösen. The methodology for determining the tonality of a word is described, and the semantic prosody index is calculated for each verb.

Keywords: semantic prosody; collocation; key word; collocates; tonality.

Стремительное развитие корпусной лингвистики привело к возникновению новых способов изучения языка и речи и вместе с тем к пересмотру некоторых базовых концепций лингвистики, лексикографии, лингводидактики. Одна из таких инноваций касается подхода к определению границ значения слова. Данные больших текстовых коллекций показывают, что значение слова существует в сложном переплетении контекста, традиций употребления и тонких

смысловых оттенков, которые часто не фиксируются в словарях, хотя могут интуитивно осознаваться носителями языка. В этой связи интересным представляется изучение **семантической просодии** слова – устойчивой **коннотативной ауры** слова или словосочетания, проявляющейся в его типичном окружении. Тенденция слова сочетаться преимущественно с лексикой позитивной или негативной окраски обусловливает восприятие тональности самого слова (изначально нейтрального) в позитивном или негативном ключе, что влияет на восприятие всего высказывания. Невнимание к семантической просодии слова – **скрытый источник коммуникативных неудач и переводческих ошибок**. Игнорирование семантической просодии слова может привести к тому, что носитель языка подсознательно воспримет фразу как стилистически неуклюжую, эмоционально неадекватную или даже оскорбительную, хотя формально она может быть построена грамматически верно.

Концепция семантической просодии была введена в 1990-х годах британским лингвистом Джоном Синклером. В своей работе «Corpus, Concordance, Collocation» Дж. Синклер показал, что значение слова нельзя понять вне его контекста: именно частотность сочетаний формирует скрытую оценку слова [3]. Дж. Синклер и его последователи показали, как слова приобретают устойчивые эмоциональные/оценочные оттенки через типичное окружение. Например, английский глагол *to commit* демонстрирует тенденцию к сочетанию с существительными негативной окраски *(commit a crime, commit suicide, commit an offence*). Употребление его в нейтральном или позитивном контексте *(commit a good deed)* звучит неестественно или иронично.

Доступ к огромным электронным базам текстов позволил объективно выявлять статистически значимые закономерности сочетаемости слов, лежащие в основе семантической просодии. Анализ миллионов коллокаций показал, что многие слова обладают устойчивой, интуитивно улавливаемой носителями языка, но часто неочевидной для неносителей языка эмоциональной окраской.

Работ по семантической просодии слов на данном этапе пока еще крайне мало, в основном авторы фиксируют свое внимание на выявлении отрицательной семантической просодии через анализ наиболее частотных коллокатов исследуемых слов, а также устанавливают ограничения по семантической просодии на основании закрепленности тональности не за леммой в целом, а за определенной грамматической формой [1, 2, 4, 5].

Думается, что коль скоро в корпусных менеджерах имеются специальные встроенные метрики, указывающие не только на частотность вхождения в корпуса определенных лемм по отдельности и в составе коллокациях, но и метрики, указывающие на силу связи коллокатов, можно попробовать формализовать процедуру исследования семантической просодии слова, предложив формулу для вычисления индекса семантической просодии слова (SPI).

Для составления формулы индекса семантической просодии слова обратимся к данным корпуса DWDS, а именно к его разделу «Профиль слова» (Wortprofil). Текущая версия этого раздела (2025 года) основана на корпусах объемом около 13 миллионов текстов, содержащих 283 миллиона предложений и 7 миллиардов слов. Профили слов предоставлены для 400 000 лемм, для в

общей сложности 22,5 миллиона различных коокуорренций с почти 1 миллиардом вхождений в корпусе. Эмпирическую базу составили, с одной стороны, пользующиеся общенациональным признанием актуальные ежедневные и еженедельные газеты, а с другой стороны, тексты художественной литературы, научно-популярной литературы, научных работ и прессы из корпуса «Базовый корпус», восходящие к началу XX века. При этом важно отметить, что в профиле слов отображены только статистически значимые словосочетания, то есть имеющие минимальную частоту вхождения F = 5 и значение logDice > 0 (Рис.1)

hat Akkusativ-Objekt		logDice	Freq.
1. Schaden	M W A	9.4	12414
2. Unfall	M W A	9.0	7141
3. Kosten	M W A	8.7	12760
4. Verkehrsunfall	M W A	7.7	1381
5. Sachschaden	M W A	7.5	1177
6. Mehrkosten	M W A	7.5	1185
7. Lärm	M W A	7.3	1269
8. Schmerz	M W A	7.2	1504
9. Aufregung	M W A	7.1	1135
10. Wirbel	M W A	7.1	893

Рисунок 1 – Наиболее частотные коллокаты глагола «verursachen»

В профиле слова корпуса DWDS представлены две метрики: logDice и F. LogDice – это статистическая мера силы связи между узлом и коллокатом в корпусе, которая рассчитывается по формуле:

logDice = 14 + $\log_2([2*f(_{X,Y})]/[f(_X)+f(_Y)])$, где: $f(_{X,Y})$ – частота совместной встречаемости слов X и Y; $f(_X)$, $f(_Y)$ – частотность каждого слова отдельно.

F – это показатель абсолютной частоты встречаемости данной коллокации в корпусе. Чем больше значение F, тем чаще встречается эта коллокация в корпусе.

SPI, таким образом, должен учитывать следующие показатели:

1. S(i)Тональность коллоката. Под тональностью мы будем понимать оценку по шкале «хорошо» – «нейтрально» – «плохо» со значениями -1, 0, +1 соответственно. В практике сентимент-анализа разработаны алгоритмы анализа тональности фраз и текстов, а также специальные сентимент-словари, по которым можно проверить тональность слова. Для немецкого языка существуют словарь SentiWS, в котором перечислены слова с положительной и отрицательной полярностью, взвешенные в интервале [-1; 1]. Он имеет ряд ограничений, однако будет полезен при определении тональности коллокатов.

2. L(i) – logDice. Это главный показатель силы связи между узлом и коллокатом. Чем больше значение logDice, тем крепче привязан коллокат к узлу в языке.

3. F(i) – частотность заданной коллокации, которая показывает сколько раз этот конкретный узел и этот коллокат встретились вместе в большом собра-

нии текстов. Так как мы имеем дело с корпусными данными, где частотность может быть очень высокой, то для расчета SPI используем логарифм абсолютной частоты по основанию 10, чтобы большие значения F не подавляли остальные показатели. Поскольку в наших данных минимальная частотность коллоката F(i) >=5, то $\log_{10}(5) = 0, 699$, что является допустимым числом и проблем с $\log_{10}(0)$ не возникнет.

4. Далее важно учесть, что каждый коллокат вносит определенный вклад в семантическую просодию слова. Вес коллоката W(i) вычислим как произведение L(i) и $\log_{10}(F(i))$. Это комбинированный вес или важность коллоката для определения семантической просодии слова. Если связь очень сильная, но коллокация встречается редко (то есть большое значение logDice, но низкая частотность), вес будет умеренным. Если же связь сильная и коллокация частотная (большое значение logDice и высокая частотность), вес будет очень большим. Если же связь слабая, то даже высокая частотность коллокации не даст большого значения W(i). Умножение балансирует силу связи и распространенность коллокации, таккак только высокочастотные коллокации с сильной связью оказывают существенное влияние на семантическую просодию слова.

5. Взвешенный вклад тональности коллокации: S(i)*W(i). Это вклад конкретной коллокации в общую семантическую просодию слова, учитывающий ее тональность и вес. При этом если тональность коллоката положительная, то и взвешенный вклад будет положительным, если отрицательная, то, соответственно, отрицательным, но если тональность нейтральная (S=0), то взвешенный вклад тональности коллоката обнулится вне зависимости от силы связи между узлом и коллокатом и частотности коллокации. Это, однако, не означает, что данная колокация вообще никак не повлияет на семантическую просодию слова, так как она будет учтена далее при нормировании показателей.

6. Числитель формулы SPI представим как сумму всех взвешенных вкладов по всем коллокатам: Σ (S(i)*W(i)). Если сумма большая положительная, то семантическая просодия слова положительная. Если большая отрицательная, то семантическая просодия слова будет также отрицательной. Если значения колеблются в районе нуля, то такую просодию можно считать нейтральной.

7. Знаменатель формулы SPI представим как сумму всех весов по всем коллокатам Σ W(i). Это нормировочный множитель. Он нужен, чтобы привести итоговый индекс SPI к удобному диапазону [-1, +1], независимо от того, насколько велики сами веса. Здесь и будут учтены те коллокации, которое не вносят вклад в положительную или отрицательную семантическу просодию слова, оставаясь нейтральными.

Учитывая все вышесказанное, формула SPI выглядит следующим образом:

$$SPI\ (i) = \Sigma\ (S(i)\ *L(i)*\ \log_{10}\ (F(i)) / \Sigma\ (L(i)*\ \log_{10}\ (F(i)))$$

Для примера возьмем немецкие глаголы со значением причинности: *verursachen, erwecken, bewirken* и рассчитаем SPI для каждого из них. Для этого найдем в корпусе DWDS по 20 коллокатов, отсортированных по значению logDice от большего к меньшему, присвоим каждому коллокату показатель тональности: 1 (позитивный), 0 (нейтральный), (-1) – негативный, рассчитаем

взвешенный вклад тональности каждой коллокации, суммируем полученные результаты и вычислим SPI.

Глагол *verursachen* демонстрирует самые сильные связи с коллокатами *Schaden, Unfall, Kosten, Verkehrsunfall, Sachschaden, Mehrkosten, Lärm, Schmerz, Aufregung, Wirbel, Brand, Krebs, Stau, Leid, Millionenschaden, Chaos, Emission, Elfmeter, Ärger, Störung*. Здесь очевидно преобладание коллокатов негативной тональности. В случае сомнений мы обращались к словарю SentiWS, а при отсутствии слова в словаре к словарным дефинициям. Так, неоднозначно определение тональности существительных *Aufregung, Elfmeter, Wirbel*. В словаре SentiWS существительное *Aufregung* имеет негативную окраску, слова *Elfmeter* тональный словарь не содержит. Словарная дефиниция слова *Elfmeter* показала также негативную тональность существительного, так как данная мера в футболе является штрафной и связана с нарушениями в игре. В случае со словом *Wirbel*, которое отсутствует в словаре SentiWS и в словарной дефиниции явно не представлена информация, дающая основания для обозначения тональности, существительному была присвоена нейтральная тональность и вклад коллокации был нейтрализован. В результате расчета $SPI_{(verursachen)} = – 0{,}9533$, что можно трактовать как сильную негативную семантическую просодию.

Глагол *erwecken* демонстрирует самые сильные связи с коллокатами *Eindruck, Anschein, Fossil, Mitleid, Misstrauen, Assoziation, Verdacht, Neid, Argwohn, Aufmerksamkeit, Illusion, Vertrauen, Sympathie, Neugier, Gefühl, Tote, Interesse, Schein, Hoffnung, Bewunderung*. Тут уже нет явных предпочтений по коллокатам негативной или позитивной тональности. Кроме того, есть нейтральные колокаты *Eindruck, Assoziation, Anschein, Schein*, которые могут приобретать различную тональность в зависимости от контекста употребления. В целом, среди выделенных коллокатов преобладают коллокаты с положительной тональностью, однако эта картина «размывается» нейтральными коллокатами, поэтому $SPI_{(erwecken)} = 0{,}058544$ и семантическая просодия глагола *erwecken* может быть описана как нейтрально-положительная.

Глагол *bewirken* сочетается с такими существительными как *Gegenteil, Veränderung, Wunder, Umdenken, Positiv, Gut, Sinneswandel, Wandel, Verhaltensänderung, Groß, Änderung, Verbesserung, Umschwung, Rückgang, Verschiebung, Effekt, Stimmungsumschwung, Anstieg, Meinungsumschwung, Wende.* Здесь ситуация совсем иная: 16 из 20 коллокатов нейтральны, что не даст результатов по явной положительной и отрицательной семантической просодии глагола, однако за счет четырех коллокатов позитивной тональности *(Gut, Positiv, Verbesserung, Wunder*) и при одновременном отсутствии коллокатов негативной тональности $SPI_{(bewirken)} = 0{,}246105$, что позволяет трактовать семантическую просодию глагола как положительно-нейтральную.

Безусловно, это только предварительный этап изучения семантической просодии слова, имеющий ряд спорных моментов:

- неотработанная методика определения тональности. Распределения весов 1, 0, -1 явно недостаточно для описания тональности различных слов. И хотя разница в тональности может интуитивно ощущаться носителем и неносителем языка, нужны четкие и максимально объективные критерии ее определения.

- ограниченный набор коллокатов. Для примера расчета SPI были взяты первые 20 коллокатов глагола, отобранные по критерию logDice. Возможно, при учете большего количества коллокатов индекс семантической просодии слова может измениться.

- учет одной меры силы связи. Кроме logDice существуют и другие меры описания коллокационной связи (MI и его варианты, t-score, log-likelihood), их привлечение также может помочь усовершенствовать подсчет SPI.

Прояснение этих и других вопросов будет способствовать повышению объективности в изучении семантической просодии слов.

ЛИТЕРАТУРА

1. Сергеева М. В. Денисова И. В. Изучение семантической просодии слова на базе анализа корпусов // Вестник нижегородского государственного лингвистического университета им. Н.а. добролюбова. Серия: Языкознание. – 2012. – № 10 (697). – С. 86–91. URL: https://elibrary.ru/item.asp?id=17778203 (дата обращения: 01.01.2025).

2. Hauser D., Schwarz N. Semantic Prosody: How Neutral Words With Collocational Positivity/Negativity Color Evaluative Judgments // Journal of Experimental Psychology: General. – 2023. URL: https://www.researchgate.net/publication/368725854_Semantic_Prosody_How_Neutral_Words_With_Collocational_Positivity Negativity_Color_Evaluative_Judgments (accessed: 01.01.2025).

3. Sinclair J. Corpus, Concordance, Collocation. – Oxford: Oxford University Press, 1991. – 179 p.

4. Stewart, D. Semantic Prosody: A Critical Evaluation. – London: Taylor & Francis, 2010. – 193 p.

5. Stubbs M. Words and Phrases: Corpus Studies of Lexical Semantics. – Oxford: Blackwell Publishing, 2001. – 287 p.