DOI 10.15826/izv2.2019.21.4.078
УДК 81'37 + 81'42 + 004.774

А. В. Колмогорова
А. А. Калинин
А. В. Маликова
Сибирский федеральный университет
Красноярск, Россия

КТО И О ЧЕМ *ГОВОРИТ* В «РАДОСТНЫХ» И «ГРУСТНЫХ» ТЕКСТАХ: В ПОИСКАХ ДИСКРИМИНАНТНЫХ ЧЕРТ ТЕКСТОВ РАЗНЫХ ЭМОЦИОНАЛЬНЫХ ТОНАЛЬНОСТЕЙ*

Статья посвящена рассмотрению специфики лексической сочетаемости и синтаксической комбинаторики глагольной лексемы *говорить* в русскоязычных интернет-текстах, принадлежащих различным эмоциональным классам.

Целью публикации является обоснование валидности использования выявленных специфических характеристик сочетаемости и комбинаторики лексемы в качестве дискриминантных черт для автоматического определения 8 эмоциональных тональностей в интернет-текстах на русском языке.

В качестве материала исследования выступает коллекция текстов, отобранных из паблика «Подслушано» в социальной сети «ВКонтакте».

Используя восьмичастную классификацию эмоций, предложенную Г. Лёвхеймом, авторы соотносят каждый из текстов выборки объемом более 1 млн токенов с определенной эмоцией посредством опоры на соответствующие хештеги и эмоциональную разметку текстов, осуществленную 36 асессорами, носителями русского языка от 19 до 45 лет.

Применение метода TF-IDF взвешивания, а также учет значений относительной частотности лексем в 8 сформированных эмоциональных подкорпусах текстов показали, что статус лексемы *говорить* неравноценен в разных подкорпусах: в 4 из них она имеет высокие относительную частотность и показатели статистической специфичности, а в оставшихся 4 подкорпусах — нет.

С помощью использования инструментов корпусной лингвистики доказано, что значимыми для автоматической атрибуции текстов к тому или иному эмоциональному классу оказываются следующие особенности лексической сочетаемости и синтаксической комбинаторики глагола *говорить*: высокий процент субъектных синтаксических связей; частотность конкретных лексем (например, *врач* для класса СТРАХ / УЖАС) и суммарная частотность лексем одной конкретной лексико-семантической группы в позиции субъекта при глаголе; частотность отдельной коллокации (например, *когда люди говорят* для класса ЗЛОСТЬ / ГНЕВ); частотность отдельных синтаксем (например, «*с собой* / *себе* lemma [*говорить*]» — для класса ГРУСТЬ / ТОСКА); частотность конкурирующих синтаксем

* Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект № 19-012-00205 «Разработка классификатора русскоязычных интернет-текстов по критерию их тональности на основе модели эмоций «Куб Лёвхейма»).

«lemma [*говорить*], *что*» и «lemma [*говорить*]: (прямая речь)», маркирующая склонность автора текста фокусироваться на содержании говоримого в форме прямой или косвенной речи.

Будучи применены в качестве параметров, подаваемых на вход компьютерному классификатору текстов, данные дискриминантные черты оказали влияние на точность атрибуции текстов к тому или иному эмоциональному классу.

К л ю ч е в ы е с л о в а: сентимент-анализ; эмоциональная тональность; интернет-тексты; машинное обучение; лексическая сочетаемость; синтаксическая комбинаторика; дискриминантная черта класса текстов.

Ц и т и р о в а н и е: *Колмогорова А. В.*, *Калинин А. А.*, *Маликова А. В.* Кто и о чем *говорит* в «радостных» и «грустных» текстах: в поисках дискриминантных черт текстов разных эмоциональных тональностей. DOI 10.15826/izv2.2019.21.4.078 // Изв. Урал. федер. ун-та. Сер. 2 : Гуманитар. науки. 2019. Т. 21. № 4 (193). С. 219–234.

Поступила в редакцию 13.05.2019
Принята к печати 01.12.2019

Anastasia V. Kolmogorova
Alexander A. Kalinin
Alina V. Malikova
Siberian Federal University
Krasnoyarsk, Russia

WHO AND ABOUT WHAT *SPEAKS* IN "CHEERFUL" AND "SAD" TEXTS: IN SEARCH OF DISCRIMINATION FEATURES IN TEXTS OF DIFFERENT EMOTIONAL TONALITIES

This article focuses on the peculiarities of lexical and syntactical combinability of the Russian verb *говорить* ("to speak") in Russian Internet texts of different emotion classes.

The article aims to substantiate and validate the use of the established specific characteristics of the combinability of the lexeme as discriminant features serving to automatically detect eight emotional tonalities in Internet texts in Russian.

The authors refer to a collection of texts found in the *Подслушано* (*The Overhead*) public page in the *vk.com* social network. Using the eight classes classification of emotions proposed by Lövheim, the researchers correlate each of the texts in their selection whose total volume is over a million tokens with a particular emotion by referring to the corresponding hashtags and the emotion mapping of the texts carried out by 36 assessors, Russian native speakers of 19–45 years old.

The statistical analysis including term-frequency-inverse document frequency measure (TF-IDF) and analysis of lexeme frequency in eight sub-corpora proves that the Russian verb *говорить* does not have the same relevance in all sub-corpora, i.e. in four of them, it demonstrates a high relative frequency and a significant statistical specificity, but in the remaining four others it does not.

Referring to the tools of corpus linguistics, the authors prove that to automatically attribute texts to a certain emotion class, it is essential to take into account the following peculiarities of lexical and syntactic combinability of the verb *говорить*: a high percentage of subjective syntactic connections, the frequency of particular lexemes (e.g. *врач* for the classes СТРАХ / УЖАС), and the total frequency of the lexemes belonging to one particular lexico-semantic group functioning as subject of the verbs; the frequency of separate collocations (e.g. *когда люди говорят* for the ЗЛОСТЬ / ГНЕВ class); the frequency of separate syntaxemes (e.g. "*с собой* / *себе* lemma [*говорить*]" for the ГРУСТЬ / ТОСКА class); the frequency of competing syntaxemes in the specific lexemes and collocations in the position of its subject, the frequency of the syntaxemes "lemma [*говорить*], *что*" и "lemma [*говорить*]: (direct speech)", marking the author's proneness to focus on the content of what is being said in the form of direct and reported speech.

After having been applied as parameters to run the classifier, the discriminate features increased the accuracy of classification for some emotion classes of texts.

K e y w o r d s: sentiment analysis; emotional tonality; Internet texts; machine learning; lexical combinatorics; syntactical combinations; text class feature.

Acknowledgements
The research is supported by the *Russian Foundation for Basic Research*, project 19-012-00205 "Design of Sentiment Classifier for Internet Texts in Russian with Reference to Lövheim's Cube of Emotion Model".

C i t a t i o n: Kolmogorova, A. V., Kalinin, A. A., & Malikova, A. V. (2019). Kto i o chem *govorit* v "radostnykh" i "grustnykh" tekstakh: v poiskakh diskriminantnykh chert tekstov raznykh emotsional'nykh tonal'nostei [Who and about What *Speaks* in "Cheerful" and "Sad" Texts: In Search of Discrimination Features in Texts of Different Emotional Tonalities]. *Izvestia. Ural Federal University Journal. Series 2: Humanities and Arts, 21*, 4 (193), 219–234. doi: 10.15826/izv2.2019.21.4.078

Submitted on 13 May, 2019
Accepted on 01 December, 2019

1. Введение

При создании технологий сентимент-анализа русскоязычных текстов методы лингвистического анализа доказали свою эффективность для решения задач по выявлению дискриминантных черт (принятый англоязычный термин — *features*) текстов разных эмоциональных классов [Polyakov, Kalinina, Pleshko; Пазельская, Соловьев]. Для машинного обучения по прецедентам подобные признаки, наряду с качественной обучающей выборкой, играют важную роль.

На данном этапе исследовательской группой ведется работа по созданию прототипа классификатора русскоязычных интернет-текстов по критерию их эмоциональной тональности. Подобные попытки создания многоклассового классификатора текстов предпринимались на основании различных типологий эмоций [Chaffar, Inkpen; Ovesdotter, Roth, Sproat]. В настоящем проекте

классификация производится по 8 классам эмоций (согласно классификации Г. Лёвхейма).

В качестве дискриминантных черт текстов 8 эмоциональных тональностей используются 18 параметров текстов, репрезентирующих как вербальные (лексическо-семантические, грамматические), так и невербальные их характеристики (эмоджи). Отметим, что наибольшую эффективность для повышения точности классификации на данный момент демонстрируют лексические и синтаксические параметры, такие как, например: частотность лексем из лексико-семантических полей «Болезнь», «Смерть», «Семья», «Одиночество»; наличие в тексте каритивных конструкций с соматизмами; частотность синтаксем «усилительное наречие с прилагательным» (ADV-интенсификатор + ADJ) и «усилительное наречие с другим наречием» (ADV-интенсификатор + ADV).

Точность классификации на данный момент не превышает 48 % (значение меры F1 = 0,48), а для некоторых классов текстов (например, тексты, отражающие эмоцию стыда или эмоцию отвращения) — ниже данного значения. Возникает проблема расширения первичной базы параметров, принимаемых в расчет классификатором.

Целью данной публикации является обоснование вывода о том, что в ряде случаев дискриминантной чертой эмоционального класса текстов, которая будет эффективна в качества параметра работы классификатора, может являться не частотность лексемы или отдельной словоформы, а частотность конкретной лексемы в определенной синтаксической позиции или в определенной комбинаторной «связке». Предметом исследования, нашедшим свое отражение в данной публикации, является специфика субъектно-объектных отношений, которые характерны для глагола *говорить* в каждом из 8 эмоциональных классов текстов из нашей размеченной тренировочной выборки. Полученные описания комбинаторных характеристик лексемы затем моделируются в качестве параметров для работы классификатора и валидируются при помощи подачи их на «вход» программы и последующей оценки точности и полноты осуществленной с их помощью классификации.

2. Материал и методы

В качестве теоретической основы для выделения эмоциональных классов текстов взята так называемая модель «Куб Лёвхейма». Шведский нейрофизиолог Гуго Лёвхейм разработал модель корреляции базовых эмоций с комбинацией уровней трех моноаминов — серотонина, дофамина и норадреналина — в виде куба [Lövheim, p. 342]. Анализ большого количества медицинской и психологической литературы позволил исследователю визуализировать данную корреляцию в виде куба на координатной плоскости с осями 5-HT (серотонин), NE (норадреналин), DA (дофамин) (см. рис. 1). В зависимости от сочетания уровня данных гормонов в крови субъекта эмоции исследователь предложил восьмичленную классификацию эмоций, где первая номинация класса

отражает наименее выраженную степень интенсивности эмоции-аффекта, а вторая — ее высшую точку (исключение составляет лишь эмоция удивления, имеющая одночленную номинацию): ИНТЕРЕС / ВОЗБУЖДЕНИЕ (Interest / Excitement); УДОВОЛЬСТВИЕ / РАДОСТЬ (Enjoyment / Joy); УДИВЛЕНИЕ (Surprise); ГРУСТЬ / ТОСКА (Distress / Anguish); ГНЕВ / ЯРОСТЬ (Anger / Rage); СТРАХ / УЖАС (Fear / Terror); ПРЕЗРЕНИЕ / ОТВРАЩЕНИЕ (Contempt / Disgust); СТЫД / УНИЖЕНИЕ (Shame / Humiliation). Таким образом, целью проводимого исследования является разработка алгоритма для автоматической атрибуции русскоязычных интернет-текстов к одному из 8 вышеуказанных эмоциональных классов.

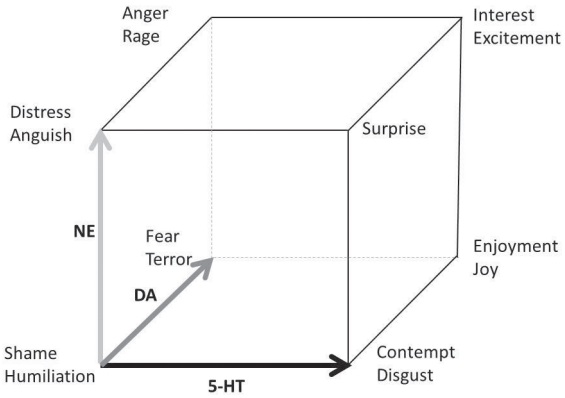

Рис. 1. Куб Лёвхейма [Lövheim, p. 342]
Fig. 1. Lövheim's cube of emotion [Lövheim, p. 342]

Достижение подобной цели возможно в рамках использования метода машинного обучения по прецедентам, требующего для обучения классификатора прежде всего размеченной по классам коллекции данных (текст — вербализованная в нем ведущая эмоция).

На основе обучающей выборки строится статистический или вероятностный классификатор. Разметка данных в таких алгоритмах может быть выполнена как авторами классификатора, так и сторонними лицами (асессорами).

Источником данных для нашей обучающей выборки послужил паблик «Подслушано» в социальной сети «ВКонтакте» — проект, в котором пользователи анонимно делятся каждый день своими откровениями, рассказывая о пережитых в разных ситуациях эмоциях и чувствах. В целях извлечения данных 8 эмоциональных классов по Лёвхейму были соотнесены с хештегами [Davidov, Tsur, Rappoport], под которыми размещаются посты пользователей на том основании, что: 1) ряд хештегов, расставляемых командой редакторов паблика, а не самими авторами текстов, прямо указывают на вербализуемое эмоциональное состояние: #Подслушано_счастье, #Подслушано_страшное, #Подслушано_стыдно, #Подслушано_фууу; 2) тексты из каждого класса были рандомизированно

оценены 36 асессорами, носителями русского языка в возрасте от 19 до 45 лет, обоих полов: каждый асессор должен был отнести каждый из предложенных ему текстов к одной или нескольким (из 8) эмоций. Подобная разметка показала, что хештеги соответствуют тем эмоциям, с которыми они были соотнесены. Продемонстрировавший при этом значительный разброс оценок корпус данных под хештегом #Подслушано_наблюдения был затем полностью повторно размечен 15 информантами на одной из краудсорсинговых платформ (см. табл. 1).

Таблица 1

Подкорпусы и их соотнесение с хештегами

Эмоциональный класс текстов (подкорпус)	Хештег в «Подслушано»
ГРУСТЬ / ТОСКА	#Подслушано_одиночество
ИНТЕРЕС / ВОЗБУЖДЕНИЕ	#Подслушано_успех
УДОВОЛЬСТВИЕ / РАДОСТЬ	#Подслушано_счастье
СТРАХ / УЖАС	#Подслушано_страшное
БРЕЗГЛИВОСТЬ / ОТВРАЩЕНИЕ	#Подслушано_фууу
ЗЛОСТЬ / ГНЕВ	#Подслушано_БЕСИТ
СТЫД / УНИЖЕНИЕ	#Подслушано_стыдно
УДИВЛЕНИЕ	#Подслушано_наблюдения #Подслушано_странное

Когда обучающая выборка была получена, мы применили инструментарий корпусной лингвистики для поиска вербальных дискриминантных черт текстов каждого эмоционального класса. В качестве такого инструмента использовался корпусный менеджер Sketch Engine. Корпусный менеджер — это поисковая система для работы с данными корпуса, получения статистической информации и предоставления пользователю результатов в удобной форме [Николаев, Митренина, Ландо, с. 141]. При помощи Sketch Engine были установлены лексические единицы, коллокации, синтаксемы, которые могут быть рассмотрены в качестве потенциальных параметров для обучения классификатора. Дополнительно с этой же целью привлекался метод TF-IDF взвешивания, представляющий собой статистическую меру, используемую для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса.

Согласно технологии машинного обучения по прецедентам, после того как установлены потенциальные параметры для обучения классификатора, выбирается один из существующих алгоритмов машинного обучения. В качестве алгоритма классификации нами выбран метод опорных векторов (SVM), поскольку это наиболее быстрый метод нахождения решающих функций [Wiebe, Riloff; Witten, Frank]. Программный код реализован на языке программирования Python.

3. Статистическая значимость леммы *говорить* в 8 классах текстов

Использование метода TF-IDF взвешивания показало, что в классах текстов ЗЛОСТЬ / ГНЕВ, УДИВЛЕНИЕ, ИНТЕРЕС / ВОЗБУЖДЕНИЕ, СТЫД / УНИЖЕНИЕ лемма *говорить* входит в 10 наиболее статистически специфических для данного класса слов.

Кроме того, в классе ЗЛОСТЬ / ГНЕВ кроме леммы *говорить* в число статистически значимых форм входит форма *говорят* (9 ранг, значение TF-IDF 7,4779). Это наблюдение нашло свое дальнейшее подтверждение при анализе 8 подкорпусов текстов при помощи инструментария корпусного менеджера Sketch Engine.

Таким образом, неравноценный по результатам TF-IDF взвешивания и по значениям частотности статус лексемы *говорить* в 8 эмоциональных классах текстов дал основания сформулировать гипотезу о необходимости тестирования специфики ее комбинаторного поведения в качестве параметра классификатора.

4. Синтаксические связи ключевого слова *говорить* в 8 эмоциональных классах текстов

4.1. Субъект речи

Использование инструмента Word Sketch корпусного менеджера Sketch Engine позволило получить данные об удельном количестве и качественных характеристиках синтаксических связей ключевого слова (далее — КС) *говорить*. В фокусе внимания в данной публикации находятся два синтаксических отношения КС: субъект (кто *говорит*) и объект (что *говорит*).

Наибольший процент субъектных связей демонстрирует класс УДИВЛЕНИЕ (табл. 2), несколько меньшее (в порядке убывания) — классы СТРАХ / УЖАС (5-я позиция по частотности лексемы), ЗЛОСТЬ / ГНЕВ, УДОВОЛЬСТВИЕ / РАДОСТЬ (5-я позиция по частотности лексемы). Наименьший процент таких связей — класс СТЫД / УНИЖЕНИЕ.

Таблица 2

Процент субъектных и объектных связей леммы *говорить* в 8 эмоциональных классах текстов

Эмоциональный класс	Субъект при КС	Прямое дополнение после КС	Косвенные дополнения с предлогами *о, про* при КС
ГРУСТЬ / ТОСКА	8,62	0	0,71
ИНТЕРЕС / ВОЗБУЖДЕНИЕ	8,88	2,30	0,30
УДОВОЛЬСТВИЕ / РАДОСТЬ	**12**	**4**	0,90
СТРАХ / УЖАС	**13**	1,73	0,30
БРЕЗГЛИВОСТЬ / ОТВРАЩЕНИЕ	6,25	**4,17**	0,31
ЗЛОСТЬ / ГНЕВ	**12,20**	2,58	**1,10**
СТЫД / УНИЖЕНИЕ	8,11	0	0,90
УДИВЛЕНИЕ	**14,60**	2,78	**2,70**

4.1.1. Субъект речи, номинированный полнозначной лексемой

Наиболее полно лексически позиция субъекта заполняется в классах: УДИВЛЕНИЕ (45 лексем), СТРАХ / УЖАС (25), ИНТЕРЕС / ВОЗБУЖДЕНИЕ (18), ЗЛОСТЬ / ГНЕВ (17). В остальных же четырех классах субъект речи в ближайшем к глаголу контексте обозначается лишь при помощи дейктиков — приглагольных местоимений *он*, *она*, *они*.

Анализ лексических заполнителей позиции субъекта речи позволяет зафиксировать некоторые специфические черты для упомянутых выше четырех классов текстов.

В классе УДИВЛЕНИЕ позицию субъекта речи, в отличие от других эмоциональных классов, могут занимать не только одушевленные, но и неодушевленные существительные (*гороскоп*, *вещь)*, зачастую с абстрактной семантикой (*сознание*):

(1) *Сознание говорит*, что это самая главная причина нашего знакомства (УДИВЛЕНИЕ).

В классе СТРАХ / УЖАС позицию субъекта речи чаще других лексем (относительная частота 30,34 ipm) занимает лексема *врач*:

(2) Раз на третий после химии *врачи говорят*, что болезнь не вылечить (СТРАХ / УЖАС).

Во всех анализируемых классах текстов в качестве заполнителей обсуждаемой синтаксической позиции присутствуют термины родства (*бабушка*, *дедушка*, *мать*, *отец*, *родители*, *муж*), но класс ЗЛОСТЬ / ГНЕВ маркирован номинациями лиц женского пола, имеющих семантический элемент «неродственные социальные связи» (*одноклассницы*, *подруга*, *подружка*, *женщина*, *коллега*):

(3) С тех пор, как закончила универ, мамины *подруги говорят* мне одно и то же на любой праздник: «желаю скорее найти жениха», потому что «время-то уже у тебя подходит» и «Наташенька-то моя в 25 уже Вовочку родила!» (ЗЛОСТЬ / ГНЕВ).

Для текстов из класса ИНТЕРЕС / ВОЗБУЖДЕНИЕ в позиции субъекта превалируют (8 лексем из 18, или 42 ipm) институциональные номинации людей (*руководительница*, *таксист*, *администратор*, *кассир*, *преподаватель*):

(4) Вернулся он, а *администратор говорит*: была уборка и номер заняли, но он настаивал и ему дали зайти, спросив у нового постояльца (ИНТЕРЕС / ВОЗБУЖДЕНИЕ).

Кроме того, для текстов из категории ЗЛОСТЬ / ГНЕВ особую роль играет коллокация (*не понимаю* / *ненавижу* / *терпеть не могу* / *бесит*) *когда люди говорят* (табл. 3).

Таблица 3

Относительная частотность (ipm) коллокации *когда люди говорят* в 8 классах текстов

Эмоциональный класс	*когда люди говорят*
ГРУСТЬ / ТОСКА	0
ИНТЕРЕС / ВОЗБУЖДЕНИЕ	1,73 ipm
УДОВОЛЬСТВИЕ / РАДОСТЬ	0
СТРАХ / УЖАС	0
БРЕЗГЛИВОСТЬ / ОТВРАЩЕНИЕ	0
ЗЛОСТЬ / ГНЕВ	68,41 ipm
СТЫД / УНИЖЕНИЕ	0
УДИВЛЕНИЕ	0

Объектом негативных эмоций могут становиться:

а) сама манера произнесения (6), лексического оформления высказывания (5), т. е. языковой вкус «обобщенного говорящего»:

(5) Терпеть не могу, *когда люди говорят* все слова в уменьшительно-ласкательной форме (ЗЛОСТЬ / ГНЕВ);

(6) Безумно бесит, *когда люди говорят* очень тихо и неэмоционально (ЗЛОСТЬ / ГНЕВ);

б) тематика сообщаемого:

(7) Ненавижу, *когда люди говорят* со мной о своих счастливых отношениях. Это очень больно (ЗЛОСТЬ / ГНЕВ);

в) некое «типичное высказывание», передаваемое обычно при помощи прямой речи или ксенопоказателей:

(8) Бешусь, *когда люди говорят*: «я гей / лесбиянка и поэтому я особенный(ая)» (ЗЛОСТЬ / ГНЕВ).

4.1.2. Особые типы субъектов речи

При работе с подкорпусами обратили на себя внимание 4 вида форм, сопряженных с субъектами речи особого типа (автокоммуникант; обобщенный коммуникант; неопределенный коммуникант; коммуникант, чье высказывание становится объектом пересказа), каждая из которых имеет особое дискурсивно-модальное значение:

а) синтаксема «*с собой* / *себе* lemma [*говорить*]» как маркер автокоммуникации:

(9) Бабушка, тебе 83, а мне 30, я помогла тебе, потому что мне торопиться некуда, и я уже давно *сама с собой говорю*;

б) коллокация *мне говорили* в качестве показателя «общественного мнения»:

(10) *Мне говорили*, что у меня помутнение и я просто преследовала его, напридумав себе наши отношения;

в) дискурсивный маркер неточности, приблизительности информации *говорят*:

(11) *Говорят*, золото нельзя воровать;

г) ксенопоказатель [Левонтина] *мол* в синтаксеме «lemma [*говорить*], *мол*»:

(12) Пришла ко мне соседка, *говорит*, *мол*, пошли купим жвачку.

Данные (табл. 4) показывают, что «разговор с самим собой» является специфическим речевым жанром для эмоционального состояния тоски, а для страха и ужаса свойственно внимание к общественному мнению (*мне говорили*). К неточной информации особенно «чувствительны» авторы текстов, в которых представлена эмоция удивления, несколько менее чувствительны — текстов, где проявляются страдание и интерес. Фокус на передаче чужого слова заметен в «удивительных» и «интересных» текстах. Наблюдения за комбинаторикой 4 типов дискурсивно-модальных форм зафиксированы в следующих «равенствах»:

ГРУСТЬ / ТОСКА = автокоммуникация + внимание к недостоверной информации;

ИНТЕРЕС / ВОЗБУЖДЕНИЕ; УДИВЛЕНИЕ = внимание к недостоверной информации + стремление к передаче чужого слова;

СТРАХ / УЖАС = важность общественного мнения + внимание к чужой речи;

БРЕЗГЛИВОСТЬ / ОТВРАЩЕНИЕ = «безразличие» к данным аспектам коммуникации;

УДОВОЛЬСТВИЕ / РАДОСТЬ; СТЫД / УНИЖЕНИЕ = незначимость данных аспектов коммуникации.

Таблица 4

Относительная частотность (ipm) 4 дискурсивно-модальных форм глагола *говорить*, сопряженных с субъектами речи особого типа в 8 эмоциональных классах текстов

Эмоциональный класс	*с собой* / *себе* lemma [говорить]	*мне говорили*	, *говорят*,	lemma [*говорить*], *мол*
ГРУСТЬ / ТОСКА	121,51	33,25	121,51	0
ИНТЕРЕС / ВОЗБУЖДЕНИЕ	0	12,00	18,60	54,01
УДОВОЛЬСТВИЕ / РАДОСТЬ	0	0	13,30	26,60
СТРАХ / УЖАС	0	18,06	9,03	13,54
БРЕЗГЛИВОСТЬ / ОТВРАЩЕНИЕ	0	0	0	0
ЗЛОСТЬ / ГНЕВ	0	0	8,12	0
СТЫД / УНИЖЕНИЕ	0	0	0	0
УДИВЛЕНИЕ	0	0	28,11	17,55

4.2. Объект речи

Объектные отношения в целом слабо выражены у анализируемого ключевого слова (табл. 2). Во всех классах в основном это тривиальные сочетания: *говорить фразу, комплимент, слово, бред*.

Номинация предмета разговора, как правило, вводится предлогами *о*/*об* и *про*. Темы наиболее разнообразны в текстах, отражающих эмоцию удивления и эмоцию ЗЛОСТЬ / ГНЕВ. При этом тематика разговоров в последнем эмоциональном классе текстов более конкретна и материальна (табл. 5) (*доплата*, *секс*, *ребенок*, *польза*, *габариты*), нежели в «удивительных» текстах, где большинство существительных, вербализующих тему обсуждения, относятся к категории абстрактных: *любовь*, *космос*, *прошлое*, *культура*, *чувство*.

Таблица 5

Дополнения глагола *говорить*, вводимые предлогами *о*/*об*, *про* в 5 эмоциональных классах текстов[1]

Эмоциональный класс	Дополнения
СТРАДАНИЕ / ТОСКА	о вещах
СТРАХ / УЖАС	об оплате
БРЕЗГЛИВОСТЬ / ОТВРАЩЕНИЕ	про страсть
ЗЛОСТЬ / ГНЕВ	о меркантильности, о габаритах, о доплате, о мелочи, о красоте, о сексе, о ребенке, про год, про пользу
УДИВЛЕНИЕ	про недостатки, про нее, про встречу, про любовь, о космосе, об объятии, о косметике, о прошлом, о болезни, о культуре, о чувстве

В классах текстов ГРУСТЬ / ТОСКА, СТРАХ / УЖАС и БРЕЗГЛИВОСТЬ / ОТВРАЩЕНИЕ встретилось только по одному случаю упоминания предмета разговора, а в классах ИНТЕРЕС / ВОЗБУЖДЕНИЕ, УДОВОЛЬСТВИЕ / РАДОСТЬ и СТЫД / УНИЖЕНИЕ — ни одного употребления.

Интересным представляется и способ передачи содержания говоримого: при помощи прямой речи (поисковая формула в Sketch Engine: сочетание леммы *говорить* и двоеточия после нее) или косвенной (поисковая формула в Sketch Engine: сочетание леммы *говорить* с запятой и союзом *что*) (см. табл. 6).

Так, для эмоциональных классов ИНТЕРЕС / ВОЗБУЖДЕНИЕ и ЗЛОСТЬ / ГНЕВ важна объективация содержания говоримого в целом, будь то в форме прямой или косвенной речи; для ряда классов существует дифференциация между формами сообщения говоримого: ГРУСТЬ / ТОСКА тяготеет к косвенной речи, передаваемой придаточным предложением; СТРАХ / УЖАС и УДИВЛЕНИЕ — наоборот, к прямой речи (рис. 2).

[1] В таблице не представлены классы текстов с незаполненными позициями дополнений данного типа.

(13) Затем врач спокойным голосом, поедая печеньку и допивая чай, *говорит мне*: «У Вас гепатит С» (СТРАХ / УЖАС);

(14) И тут она останавливается и серьезно так *говорит*: «Беременным отказывать нельзя!» (УДИВЛЕНИЕ);

(15) Сядет медленно на свой стул перед зеркалом и начинает *говорить, что* вот там, в зеркале, она ненастоящая, а когда-то была настоящая и хорошенькая (СТРАДАНИЕ / ТОСКА).

Таблица 6

Относительная частотность (ipm) синтаксем с леммой *говорить* в 8 классах текстов

Эмоциональный класс	lemma [*говорить*], *что*	lemma [*говорить*]:
ГРУСТЬ / ТОСКА	247,92	70,83
ИНТЕРЕС / ВОЗБУЖДЕНИЕ	423,74	380,28
УДОВОЛЬСТВИЕ / РАДОСТЬ	293,71	270,22
СТРАХ / УЖАС	43,30	260,00
БРЕЗГЛИВОСТЬ / ОТВРАЩЕНИЕ	174,48	218,00
ЗЛОСТЬ / ГНЕВ	440,85	433,25
СТЫД / УНИЖЕНИЕ	284,77	327,49
УДИВЛЕНИЕ	235,89	655,63

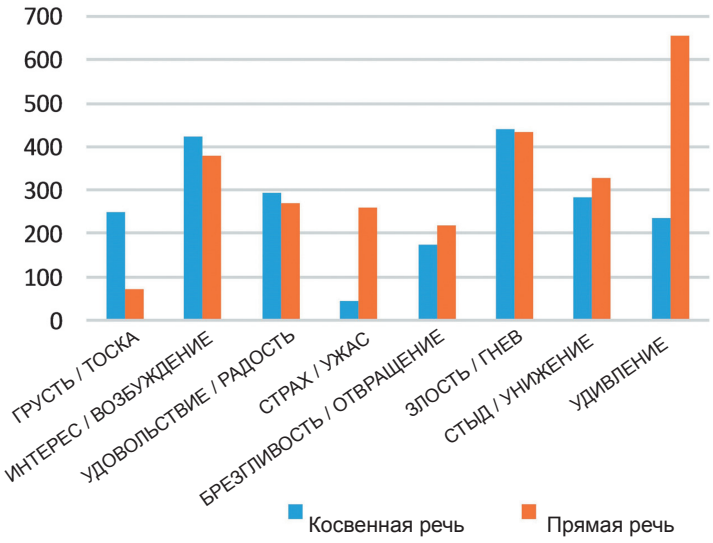

Рис. 2. Соотношение прямой и косвенной речи после леммы *говорить*

Fig. 2. Comparative analysis of direct and indirect speech frequency after the *говорить* lemma

5. Предварительная валидация результатов

При добавлении в список параметров ряда рассмотренных комбинаций с лексемой *говорить*, подаваемых на вход классификатору, суммарная точность классификации по всем классам выросла на 15 % по сравнению с точностью, достигнутой при подаче на вход базовых 18 параметров. В особенности это коснулось классов БРЕЗГЛИВОСТЬ / ОТВРАЩЕНИЕ, СТРАДАНИЕ / ТОСКА и УДОВОЛЬСТВИЕ / РАДОСТЬ. Подчеркнем, что в табл. 7 указаны данные только для тех параметров, которые дали прирост значений F1-score. Однако параметры, не показавшие своей эффективности, будут в дальнейшем тестироваться в комбинациях с другими.

При подаче описанных выше коллокаций и синтаксем на вход классификатору в качестве признаков-предикторов классов показатель weighted average F1-score, представляющий собой взвешенное по доле каждого класса гармоническое среднее значений точности и полноты классификации, вырос на 2 %. Вопрос о том, комбинация данных черт с какими другими дискриминантными чертами из уже имеющихся в распоряжении исследователей даст наиболее оптимальные результаты, остается пока открытым.

Таблица 7

Оценка точности работы классификатора (F1-score) при добавлении ряда параметров на основе глагола *говорить*

Эмоция	База из 18 параметров	+ *когда люди говорят*	+ себе [*говорить*]	+ lemma [*говорить*], *мол*	+, *говорят*,
Гнев	0,48	0,48	0,48	0,48	0,48
Отвращение	0,04	0,04	0,05	0,06	0,06
Тоска	0,45	0,45	0,46	0,46	0,48
Радость	0,10	0,10	0,07	0,12	0,12
Интерес	0,34	0,34	0,32	0,32	0,32
Страх	0,48	0,48	0,47	0,48	0,48
Стыд	0,10	0,10	0,14	0,14	0,14
Удивление	0,23	0,23	0,19	0,24	0,24
micro avg	0,35	0,35	0,34	0,36	0,36
macro avg	0,27	0,27	0,27	0,28	0,28
weighted avg	0,30	0,30	0,32	0,32	0,32

Заключение

Лексема *говорить* как своеобразная «маркирующая субстанция» пронизывает все тексты. Но синтаксическое и комбинаторное поведение лексемы отличается внутри текстов различных эмоциональных классов, так что ряд характеристик можно рассматривать как дискриминантные черты отдельных эмоциональных тональностей в тексте.

В качестве кандидатов в маркеры классов есть основания рассматривать:

1) высокий процент субъектных связей лексемы *говорить*;

2) высокий процент объектных связей (вводимых предлогами *о*/*об*, *про*) лексемы *говорить*;

3) частотность конкретных лексем, коллокаций или суммарную частотность лексем одной конкретной лексико-семантической группы в позиции субъекта при глаголе *говорить*: лексема *врач* — для класса СТРАХ / УЖАС; коллокация *когда люди говорят* и ЛСГ «Номинации лиц женского пола, имеющие семантический элемент “неродственные социальные связи”» — для класса ЗЛОСТЬ / ГНЕВ; ЛСГ «Институциональные номинации человека» + *говорить* — для класса ИНТЕРЕС / ВОЗБУЖДЕНИЕ;

4) частотность 4 форм выражения нетипичного субъекта речи: синтаксемы «*с собой*/ *себе* lemma [*говорить*]» — для класса ГРУСТЬ / ТОСКА; коллокация *мне говорили* — для класса СТРАХ / УЖАС; синтаксема «lemma [*говорить*], *мол*» — для классов ИНТЕРЕС / ВОЗБУЖДЕНИЕ и УДИВЛЕНИЕ;

5) частотность конкурирующих синтаксем «lemma [*говорить*], *что*» и «lemma [*говорить*]: (прямая речь)», маркирующую склонность автора текста фокусироваться на содержании говоримого в форме прямой или косвенной речи: предпочтение косвенной речи прямой — для класса ГРУСТЬ / ТОСКА, предпочтение прямой речи косвенной — для классов УДИВЛЕНИЕ и СТРАХ / УЖАС.

В заключение отметим «на полях», что наибольшую «сензитивность» к характеристикам говорения показали классы текстов ЗЛОСТЬ / ГНЕВ, УДИВЛЕНИЕ, ИНТЕРЕС / ВОЗБУЖДЕНИЕ и ТОСКА / СТРАДАНИЕ. Интересно, что согласно «Кубу Лёвхейма» именно эти 4 эмоции характеризуются высоким уровнем норадреналина.

Исследования

Левонтина И. Б. Пересказывательность в русском языке // Компьютерная лингвистика и интеллектуальные технологии : по материалам ежегодной междунар. конф. «Диалог» (Бекасово, 26–30 мая 2010 г.). Вып. 9 (16). М. : Рос. гос. гуманитар. ун-т, 2010. С. 284–288.

Николаев И. С., *Митренина О. В.*, *Ландо Т. М.* Прикладная и компьютерная лингвистика. М. : ЛЕНАНД, 2016.

Пазельская А. Г., *Соловьев А. Н.* Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии : по материалам ежегодной междунар. конф. «Диалог». М. : Изд-во РГГУ, 2011. С. 510–522.

Chaffar S., *Inkpen D.* Using a heterogeneous dataset for emotion analysis in text // Canadian Conference on Artificial Intelligence. Berlin ; Heidelberg : Springer, 2011. P. 62–67.

Davidov D., *Tsur O.*, *Rappoport A.* Enhanced Sentiment Learning Using Twitter Hashtags and Smileys // Proceedings of the 23rd International Conference on Computational Linguistics: Posters / Association for Computational Linguistics. 2010. P. 241–249.

Lövheim H. A new three-dimensional model for emotions and monoamine neurotransmitters. DOI 10.1016/j.mehy.2011.11.016 // Medical Hypotheses. 2012. № 78. P. 341–348.

Ovesdotter C., *Roth D.*, *Sproat R.* Emotions from text: machine learning for text-based emotion prediction // Proceedings of the Joint Conference on Human Language Technology/Empirical Methods

in Natural Language Processing (HLT/EMNLP) / Association for Computational Linguistics. 2005. P. 579–586.

Polyakov P. Yu., Kalinina M. V., Pleshko V. V. Automatic Object-oriented Sentiment Analysis by Means of Semantic Templates and Sentiment Lexicon Dictionaries // Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialogue 2015". 2015. № 14. T. 2. P. 44–52.

Wiebe J., Riloff E. Creating Subjective and Objective Sentence Classifiers from Unannotated Texts // Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science. Vol. 3406. P. 486–497.

Witten I. H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 2nd ed. Burlington : Morgan Kaufmann, 2005.

References

Chaffar, S., & Inkpen, D. (2011). Using a Heterogeneous Dataset for Emotion Analysis in Text. In *Canadian Conference on Artificial Intelligence* (pp. 62–67). Berlin; Heidelberg: Springer.

Davidov, D., Tsur, O., & Rappoport, A. (2010). Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters* (pp. 241–249). Association for Computational Linguistics.

Levontina, I. B. (2010). Pereskazyvatel'nost' v russkom iazyke [Retelling in Russian]. *Komp'iuternaia lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoi mezhdunarodnoi konferentsii "Dialog" (Bekasovo, 26–30 maia 2010 g.)* [Computational Linguistics and Intellectual Technologies "Dialogue 2010"] (Iss. 9 (16), pp. 284–288). Moscow: Russian State University for the Humanities. (In Russian)

Lövheim, H. (2012). A New Three-Dimensional Model for Emotions and Monoamine Neurotransmitters. *Medical Hypotheses*, 78, 341–348. doi: 10.1016/j.mehy.2011.11.016

Nikolaev, I. S., Mitrenina, O. V., & Lando, T. M. (2016). *Prikladnaia i komp'iuternaia lingvistika* [Applied and Computational Linguistics]. Moscow: LENAND. (In Russian)

Ovesdotter, C., Roth, D., & Sproat, R.(2005). Emotions from Text: Machine Learning for Text-Based Emotion Prediction. In *Proceedings of the Joint Conference on Human Language Technology/ Empirical Methods in Natural Language Processing (HLT/EMNLP)* (pp. 579–586). Association for Computational Linguistics.

Pazelska, A. G., & Solovjev, A. N. (2011). Metod opredeleniia ehmotsij v tekstakh na russkom iazyke [Sentiment Analysis of Texts in Russian]. *Komp'iuternaia lingvistika i intellektual'nye tekhnologii: po materialam ezhegodnoj mezhdunarodnoj konferentsii "Dialog"* [Computational Linguistics and Intellectual Technologies "Dialogue 2011"] (pp. 510–522). Moscow: RGGU Press. (In Russian)

Polyakov, P. Yu., Kalinina, M. V., & Pleshko, V. V. (2015). Automatic Object-oriented Sentiment Analysis by Means of Semantic Templates and Sentiment Lexicon Dictionaries. In *Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialogue 2015"* (14, Vol. 2, pp. 44–52).

Wiebe, J., & Riloff, E. (2005). Creating Subjective and Objective Sentence Classifiers from Unannotated Texts. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing. CICLing 2005. Lecture Notes in Computer Science* (Vol. 3406, pp. 486–497).

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). Burlington: Morgan Kaufmann.

Колмогорова Анастасия Владимировна
доктор филологических наук, профессор,
заведующий кафедрой романских
языков и прикладной лингвистики
Сибирский федеральный университет
660041, Красноярск, пр. Свободный, 82а
E-mail: nastiakol@mail.ru

Kolmogorova, Anastasia Vladimirovna
Dr. Hab. (Philology), Professor
Head of the Romance Languages and Applied
Linguistics Department
Siberian Federal University
82a, Svobodny Ave., 660041 Krasnoyarsk, Russia
Email: nastiakol@mail.ru
ORCID: 0000-0002-6425-2050
ResearcherID: D-9618-2017
Scopus AuthorID: 56642774800

Калинин Александр Андреевич
старший преподаватель кафедры
романских языков и прикладной
лингвистики
Сибирский федеральный университет
660041, Красноярск, пр. Свободный, 82а
E-mail: verbalab@yandex.ru

Kalinin, Alexander Andreevich
Senior Lecturer, Romance Languages
and Applied Linguistics Department
Siberian Federal University
82a, Svobodny Ave., 660041 Krasnoyarsk, Russia
Email: verbalab@yandex.ru
ORCID: 0000-0002-0012-1692
Scopus AuthorID: 57203318829

Маликова Алина Вячеславовна
преподаватель кафедры романских
языков и прикладной лингвистики
Сибирский федеральный университет
660041, Красноярск, пр. Свободный, 82а
E-mail: malikovaav1304@gmail.com

Malikova, Alina Vyacheslavovna
Lecturer, Romance Languages and Applied
Linguistics Department
Siberian Federal University
82a, Svobodny Ave., 660041 Krasnoyarsk, Russia
Email: malikovaav1304@gmail.com
ORCID: 0000-0002-3438-1839
ResearcherID: D-9625-2017
Scopus AuthorID: 57204770224