УДК 81-11 + 81'32 + 81'322.2
DOI 10.25205/1818-7935-2018-16-2-5-18

С. Ю. Пужаева [1], **Е. А. Герасименко** [1], **Е. С. Захарова** [1], **Е. В. Рахилина** [1, 2]

[1] *Национальный исследовательский университет «Высшая школа экономики»*
ул. Старая Басманная, 21/4, Москва, 105066, Россия

[2] *Институт русского языка им. В. В. Виноградова РАН*
ул. Волхонка, 18/2, Москва, 119019, Россия

syupuzhaeva@gmail.com, katgerasimenko@gmail.com
1583253@gmail.com, rakhilina@gmail.com

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ДИСКУРСИВНЫХ ФОРМУЛ ИЗ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ [*]

Статья посвящена проблеме создания модуля автоматического извлечения из текстов русского языка особых единиц – дискурсивных формул. Под дискурсивными формулами (ДФ) мы понимаем неоднословные конструкции, которые, однако, не содержат переменных и выступают в роли ответных реплик на вербальный стимул. Работа над программным модулем включала в себя несколько этапов, в том числе ручную разметку пьес по выявленным в ходе работы критериям. Процесс автоматического извлечения ДФ предусматривает деление текста на синтаксические единицы, соотносимые с клаузой, предсказание принадлежности каждой из единиц к классу ДФ на основании выделенного нами набора признаков и формирование итогового списка ДФ. В качестве алгоритма классификации используется равновесное голосование четырех классификаторов: Random Forest Classifier, Logistic Regression, Ridge Classifier, Support Vector Classifier.

Ключевые слова: дискурсивные формулы, грамматика конструкций, машинное обучение, автоматическое извлечение сущностей.

Введение

Последние десятилетия XX в. ознаменованы появлением перевернувшей многие представления современной лингвистики теории – Грамматики конструкций [Fillmore, 1988; 1989; Fillmore, Kay, 1992]. Сегодня она представляет собой уже множество концепций (см. [Hoffmann, Trousdale, 2013]), базирующихся, тем не менее, на ряде общих принципов, важнейшим из которых оказывается отказ от проведения четкой границы между грамматикой и словарем. Сам по себе этот принцип в разных вариантах уже формулировался в других лингвистических подходах (ср. идею интегрального описания языка, выдвинутую Ю. Д. Апресяном еще в 1980 г. [Апресян, 1980] [1]). Однако понятие конструкции оказалось чрезвычайно продуктивным и придало новый импульс многим исследованиям. Согласно Грамматике конструкций, «C is a CONSTRUCTION iff$_{def}$ C is a form-meaning pair $\langle F_i, S_i \rangle$ such that some aspect of F_i or some aspect of S_i is not strictly predictable from C's component parts or from other

[*] Исследование выполнено при поддержке гранта РНФ № 16-18-02071.

[1] Перепечатано в [Апресян, 1995. С. 8–101].

Пужаева С. Ю., Герасименко Е. А., Захарова Е. С., Рахилина Е. В. Автоматическое извлечение дискурсивных формул из текстов на русском языке // Вестн. Новосиб. гос. ун-та. Серия: Лингвистика и межкультурная коммуникация. 2018. Т. 16, № 2. С. 5–18.

ISSN 1818-7935
Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2018. Том 16, № 2

previously established constructions» [Goldberg, 1995. P. 4], иными словами, конструкция – это «языковое выражение, у которого есть аспект плана выражения или плана содержания, не выводимый из значения или формы составных частей» [Рахилина, Кузнецова, 2010. С. 19]. Как известно, конструкции включают в себя позиции (слоты), которые могут быть заполнены переменными в зависимости от накладываемых на них семантических ограничений. Хрестоматийным примером конструкции служит английская конструкция *let alone*, переводимая на русский язык сочетанием *не говоря* (*уже*) *о*:

(1) F (X A Y let alone X BY)
I doubt (*F*) *that he made* (*X*) *colonel* (*A*) *is in World War II* (*Y*), *let alone general* (*B*)
'Я сомневаюсь (F), что он стал полковником (A) во время Второй мировой войны (Y), не говоря уж о генерале (B)' [Fillmore et al., 1988].

По форме *let alone* представляет собой сочинительный союз, который связывает два равноправных элемента. Однако семантика этой конструкции предполагает наличие некоторой шкалы, «на которой расположены варианты развития событий, причем первая пропозиция (A) представляет более слабое предположение, чем второе (B)». Прагматика конструкции заключается в том, что «здесь отвергается более слабое (т. е. более вероятное) предположение, чем то, в котором заинтересован говорящий» [Рахилина, Кузнецова, 2010. С. 19].

Таким образом, Грамматика конструкций, отражая некомпозициональность языковых выражений, одновременно оставляет для многих из них возможность сложной, но композициональной интерпретации через правила взаимодействия переменных с ограничениями, предъявляемыми самой конструкцией на разных языковых уровнях.

Важная задача, о которой говорил еще Ч. Филлмор [Fillmore, 2008], – выделение перечня конструкций для каждого языка. Она очень трудоемка и решается пока всего для нескольких языков, в число которых входит и русский (для бразильского варианта португальского языка см. [Lage, 2013], для японского языка см. [Ohara, 2015]). Для создания инвентаря конструкций русского языка в рамках совместного проекта Арктического университета (Тромсё, Норвегия) и Высшей школы экономики (Москва, Россия) сейчас ведется работа над специальным ресурсом – базой данных «Конструктикон». Идея этой базы заключается в том, чтобы описать конструкции одного языка так, чтобы был возможен автоматический поиск по их формальным и семантическим признакам [Janda et al., 2018]. В базе есть их синтаксическая и морфологическая разметка, некоторые ограничения на употребления, примеры и примерные толкования. В «Конструктикон» попадают как уже описанные конструкции (которые отражены в учебниках русского языка как иностранного, грамматиках, разнообразных словарях и т. д.), так и новые, выделяемые вручную разметчиками из транскриптов устной речи, пьес, детских книг и т. п. под контролем экспертов.

В настоящее время в базе документировано более 600 конструкций. Ясно, что среди них есть конструкции самых разных синтаксических и семантических типов – они выявлены уже на этапе сбора и предварительной разметки материала. Каждый такой тип (глагольные конструкции, типовые матричные и вставленные предложения, конструкции с легкими глаголами и т. п.) имеет свою специфику и в дальнейшем будет пополнен и описан более подробно в отношении своих семантических и типологических свойств.

Между тем в ходе такой предварительной работы выяснилось, что среди типов конструкций, представленных в «Конструктиконе», выделяются особые, неканонические с точки зрения теории Грамматики конструкций, неоднословные единицы, – так называемые *дискурсивные формулы* (ДФ), ср.: *Это еще что! Вот оно как! А то!* и др. Несмотря на то что, по интуитивному представлению и разметчиков, и экспертов, они, без сомнения, являются конструкциями русского языка, вопреки нашим теоретическим ожиданиям они не содержат переменных внутри самой единицы. Очевидно, что это уникальное свойство выделяет такие единицы в особую группу, как бы приравнивая их к такому типу безвариантных микротекстов, как поговорки (*Жизнь прожить – не поле перейти*; *Больше живешь – больше видишь*; *Без труда не выловишь и рыбку из пруда* и т. п.). Это свойство создает определенные сложности для их представления в базе «Конструктикон» наряду с другими, каноническими, конструкциями. Действительно, главной целью разметки канонических конструкций и их толкования является описание того способа, которым переменные с определенными свойствами

встраиваются в соответствующую невариативную часть; если переменных нет, то нет и правил, как в поговорках, или они имеют совершенно другую природу. Однако выделенные нами дискурсивные формулы гораздо ближе к стандартным конструкциям. Их главное отличие от канона в том, что соответствующие им слоты не являются частью данного предложения, обычно располагаясь в левом и / или правом контекстах. Более того, эти слоты вообще не являются частью предложения. По-видимому, им соответствует тип речевого акта предложения-соседа.

Таким образом, ясно, что термин *дискурсивные формулы* был выбран неслучайно: с одной стороны, эти особые конструкции имеют прагматическую, дискурсивную природу, а с другой – формульную. К формульным единицам относятся как весьма очевидные идиомы, пословицы и поговорки (см. [Wray, 2005]), так и другие неоднословные единицы, выступающие в качестве формульного выражения [Biber et al., 1999], ср. известные из англоязычной традиции термины *formulae* [Corrigan et al., 2009], *formulaic sequences* и ряд других связанных с ними терминов – *formulaic speech*, *formulas*, *prefabricated routines*, *conventionalised forms*, упоминаемых, в частности, в работе [Schmitt, Carter, 2004].

В функциональном аспекте наше понимание ДФ близко к так называемым коммуникативам [Шаронов, 1996], которые определяются как «особые употребления слов, фразем и коротких предложений в позиции ответных реплик диалога для стереотипного выражения оценки, мнения и эмоции как реакции на высказывание собеседника (например: *Да*; *Нет уж*; *Какое там*; *Обалдеть!*; *На здоровье!* и др.)» [Шаронов, 2016]. Класс коммуникативов много шире, чем нужно для «Конструктикона» – туда входят самые разные (в том числе и однословные) единицы, засвидетельствованные в стандартных словарях, – вплоть до междометий. Есть другой похожий термин: *речевые формулы*, определяющий иной, хотя опять-таки функционально близкий класс языковых единиц – уже неоднословных. Он используется Д. О. Добровольским и А. Н. Барановым при описании русской фразеологии [Баранов, Добровольский, 2000], однако совпадение с ДФ также оказывается лишь частичным, в силу того что речевые формулы представляют собой в первую очередь такие идиоматические сочетания, как пословицы и поговорки.

В отличие от остальных типов конструкций ДФ очень разнообразны и по структуре, и по составу. Их крайне мало в учебниках и словарях, а кроме того, эти формулы быстро устаревают и сменяются новыми. В обычной художественной прозе, которую размечают для поиска конструкций, их не так много, так что оценить их долю в таком языке, как русский, не просто. Однако для того, чтобы системно описать ДФ в соответствии с практикой «Конструктикона», с позиции семантики, структуры, ограничений, налагаемых на слоты конструкции, и в дальнейшем построить их более дробную классификацию, необходимо иметь достаточно полный и представительный список таких единиц. В связи с этим ключевой задачей становится создание модуля автоматического извлечения ДФ, который позволил бы если не решить эту задачу, то хотя бы автоматизировать ее самый сложный, начальный этап: создать как можно более близкий к реальному пилотный список, который затем можно было бы редактировать вручную. Обсуждению этой задачи и посвящена настоящая работа.

Текстовым материалом для машинной и ручной обработки были выбраны пьесы.

Подготовительный этап работы: ручная разметка

На подготовительном этапе ключевая задача – вручную разметить достаточное для построения модели машинного обучения количество текстов, содержащих ДФ. В качестве материала использованы пьесы (разного времени, начиная с XIX в.): в них формульные реакции на предшествующий дискурс разнообразнее и встречаются чаще, так как пьесы, с одной стороны, имитируют устную речь, а с другой – гораздо более удобны для автоматической обработки, чем, к примеру, транскрипты устного дискурса или субтитры. Это объясняется более четкой структурой текста пьес, а также специальными маркерами для реплик говорящих.

Дискурсивные формулы в русском языке

Прежде чем приступить к ручной разметке, было необходимо инструктировать разметчиков, выделить основные черты ДФ, а также более четко очертить круг единиц, которые попадают в этот класс.

В работе [Moon, 1997] выделяются следующие черты формульных единиц, которые автор в данном случае называет просто *multi-word items* (*неоднословные единицы*)*:* устойчивость, некомпозициональность и привычность формы [Ibid. P. 44]. Предположительно, формульные конструкции хранятся как единое целое в ментальном лексиконе говорящих («stored as wholes in the subject's mental lexicon») [Schmitt, Carter, 2004]. По мнению авторов статьи, из этого следует то, что формульная единица воспроизводится с неизменным присущим ей интонационным контуром. Точно так же, как коммуникативы, ДФ синтаксически независимы, поскольку главным образом представляют собой ответные реплики диалога, в их состав тоже входят преимущественно служебные слова (часто десемантизированные) и их сочетания, ср. [Шаронов, 1996].

В связи с этим ДФ были определены нами как неоднословные, легко воспроизводимые некомпозициональные языковые единицы, представляющие собой законченные предложения – реакции на вербальный стимул. Они включают в свой состав преимущественно междометия, частицы, десемантизированные слова, что выдвигает на первый план интонационное оформление и прагматическую функцию. Принципиально важным оказывается отсутствие переменных внутри этих высказываний.

Критерии выделения дискурсивных формул на этапе разметки

Безусловно, теоретическое уточнение границ класса ДФ на базе общего представления о них как о квазиконструкциях с ярко выраженной дискурсивной функцией было важным этапом работы, однако на практике, при ручной разметке пьес, выделенные признаки оказались недостаточными в том числе из-за случаев, усложняющих выделение ДФ. Далее будут перечислены сложные случаи разграничения ДФ и омонимичных им синтаксических единиц.

Во-первых, некоторые рассматриваемые нами единицы могут совпадать по форме с вводными сочетаниями, которые в терминологии дискурсивного анализа часто в широком смысле именуют дискурсивными маркерами [Schiffrin, 1988] ср. (2, 3).

(2) Солдат огляделся, снял шинель, вымыл руки и, присев к столу, вытащил расческу. – Ну, Ефим, – сказал он, – вот это житье! А как и что тут, пожалуй, сразу не разберешься! – **Что и говорить!** – крякнул Ефим. – У меня зараз в голове як в той пивной бочке!.. Однако хорошо тут сидеть, только я пойду! (Валентина Осеева. Динка прощается с детством (1969))

(3) Картина, **что и говорить**, мерзкая, в кое-каких деталях он узнавал себя, не испытывая, впрочем, чувства вины, поскольку в описании мерзостей преобладали слова, употребляемые во всех сочинениях на научные темы (Анатолий Азольский. Лопушок // Новый Мир. 1998)

Во-вторых, ДФ по форме могут совпадать со знаменательными единицами, включенными в структуру предложения ср. (4, 5).

(4) [Евлалия Андревна, жен, 30] По какому городу бегать, про каких кавалеров разыскивать? [Марфа Севастьяновна, жен] Да **как же** их… ну, хоть Артемия Васильича **назвать**! Обыкновенно для учтивости кавалерами называешь. [Евлалия Андревна, жен, 30] Зачем ты об Артемии Васильиче говоришь, скажи мне? (А. Н. Островский. Невольницы (1881))

(5) Бестужев же был профессором и доктором наук. – Получается – раздвоение личности? Неужели мировоззрения Бестужева и Лады не пересекались? – **Как же!** Пересекались. Я читал студентам начало и конец курса научного коммунизма, а в середине рассказывал свои идеи. Это было в середине 1960-х, ещё до Праги 1968-го и реакции [Игорь Бестужев-Лада, Илья Кашницкий. Если можно изучать прошлое, почему не исследовать будущее? // Зеркало мира. 2012)

В-третьих, полностью совпадать по форме с ДФ, а в широком смысле к ним могут быть отнесены и такие реплики (мы называем их рефлективными), с помощью которых говорящий

реагирует на свое собственное высказывание, риторически структурируя его, вводя дополнительные аргументы, споря с собой и пр., ср. (6, 7).

(6) Всего этого вместе с толками, сплетнями и пересудами ей было совершенно достаточно, и она объявила, что ни одного дня отец больше не будет работать у Кусиела. **Как так?** Отец устроился, вошёл, можно сказать, в курс, приобрел специальность, через месяц-два станет компаньоном, и, пожалуйста, – бросай дело?! (Анатолий Рыбаков. Тяжелый песок (1975–1977))

(7) Чуть концы не отдал. Можно сказать, уже там был. Зато у меня сын родился через эту яму. – **Как так?** – Я думаю так, что у меня для пацана извести в организме не хватало. – Ну, извести у тебя хватало. – Кроме шуток, – смеётся Ванечка, – может, я научное открытие сделал (Фазиль Искандер. Должники (1968))

В-четвертых, как мы уже говорили, ДФ представляют собой законченные реплики, однако они могут предшествовать обращению (9), другой ДФ (8), располагаться после обращения. Таким образом, признаки начала предложения и наличия финального знака препинания (точка, восклицательный знак, вопросительный знак) с формальной точки зрения не являются для них достаточными.

(8) [Леня, муж] Мы с Марусей подружились. Мы все с ней подружились. Она простая. [Ольга Ивановна, жен] **Ну вот и хорошо**, **вот и все** (Евгений Шварц. Повесть о молодых супругах (1955))

(9) [Ихарев, муж] Как же? и денег не хотите дождаться? [Глов, муж] **Что ж делать**, батюшка? (Николай Гоголь. Игроки (1842))

В-пятых, как ответная реплика ДФ располагается преимущественно в начале комплекса высказываний говорящего, однако бывает и по-другому, в частности переспрос, как в примере (10).

(10) [Дарья Ивановна, жен] Любин... Граф Любин. Ведь ты его знаешь? [Ступендьев, муж] Любина? **Еще бы!** Так ты его ожидаешь? (Иван Тургенев. Провинциалка (1850))

Таким образом, есть контексты, в которых поведение ДФ отклоняется от прототипического, и они существенно осложняют нашу задачу по формальному определению ДФ. В связи с этим мы ввели дополнительные, уточняющие критерии в инструкцию для разметчиков выделять в качестве ДФ единицы, удовлетворяющие следующим требованиям:

1) реплика является ответной на вербальный стимул;
2) можно представить изолированное употребление такого же сегмента в ином контексте;
3) реплика не встроена в синтаксическую структуру остального высказывания;
4) реплика не выступает в функции оценки собственного высказывания (не является вводным сочетанием, рефлективным высказыванием);
5) в реплике преобладают десемантизированные слова, что выводит на первый план прагматическую функцию;
6) реплика не содержит полнозначных слов, аналогичных словам, включенным в синтаксическую структуру левого контекста;
7) последовательность реплик, разделенных не финальным знаком препинания, рассматривается как соединение разных дискурсивных формул, если для каждой из них соблюдаются требования 1–4;
8) реплика может быть размечена как ДФ, если располагается не в абсолютном начале сегмента, но удовлетворяет требованиям 1–4.

По описанным выше формальным признакам ДФ двое разметчиков выделили ДФ в 24-х пьесах полностью вручную, затем еще 46 пьес было размечено полуавтоматически, т. е. проведена корректировка результатов автоматической системы извлечения ДФ. Таким образом, разметка проводилась в несколько итераций.

Система автоматического извлечения дискурсивных формул

Для обработки больших текстовых массивов драматических произведений был создан модуль автоматического извлечения ДФ из текстов. Разработанный нами инструмент состоит из следующих трех компонентов:

1) обработка исходного текста:
   a) удаление обозначений говорящих;
   b) разбиение текста на отдельные единицы, подобные клаузе (псевдоклаузы [2]);
   c) извлечение из каждой единицы признаков, которые дальше используются в модели машинного обучения;
2) получение бинарного предсказания для каждой единицы (является она ДФ или нет);
3) составление финальной таблицы ДФ, которую получает пользователь, что включает в себя выделение левого контекста для каждой ДФ.

Предсказание производится голосованием обученных на размеченном корпусе моделей машинного обучения. Общий процесс работы системы, включая обучение, представлен на рисунке.

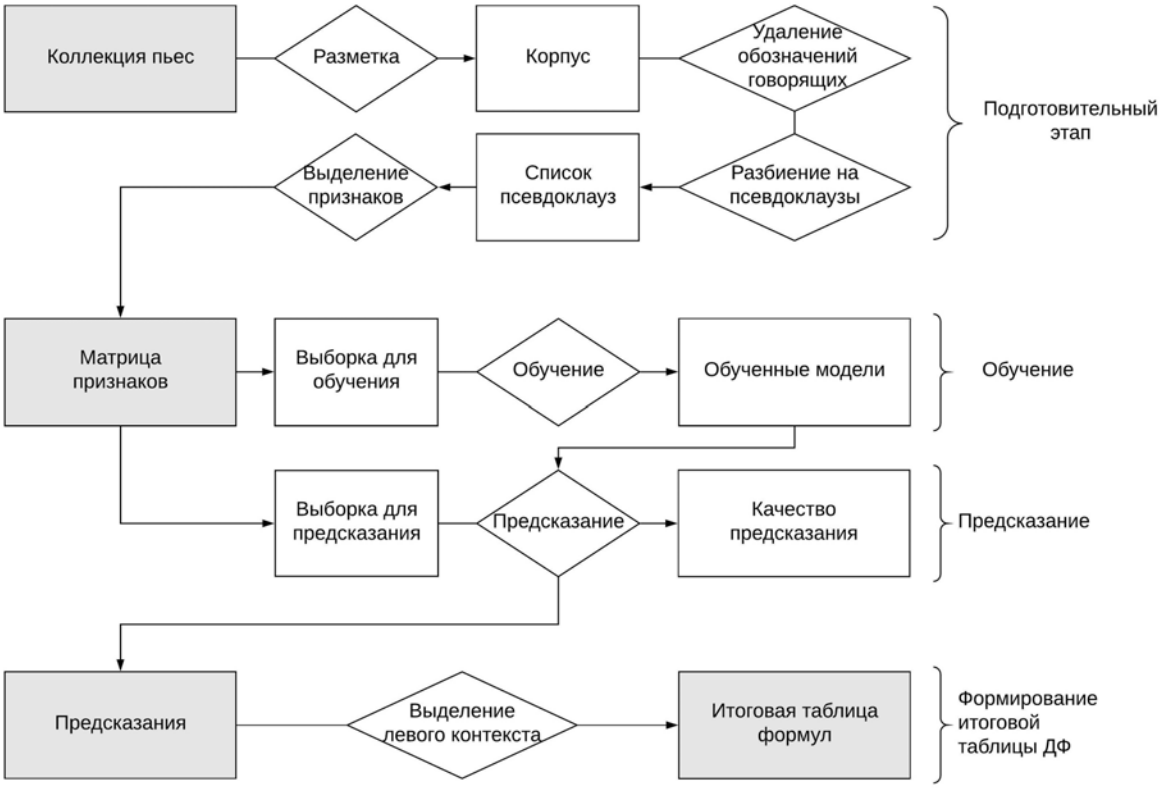

Рис. 1. Автоматическое извлечение дискурсивных формул

Обработка исходного текста

В принципе, для любого семантического анализа драматического текста (пьесы) важно было бы знать, какому персонажу принадлежит та или иная реплика. Однако для нашего автоматического анализа эта информация не только не важна, но, более того, она может снизить качество предсказания, так как по выделенным признакам псевдоклауза с обозначением

[2] См. пояснение понятия «псевдоклауза» далее.

говорящего может быть похожа на ДФ. Поэтому первым этапом предобработки текстов является удаление обозначений говорящих. Далее текст делится на псевдоклаузы – отрезки, разделенные знаками препинания. Такой упрощенный способ выделения псевдоклауз может генерировать некоторое количество неверных результатов (например, несколько именных групп, разделенные запятыми, окажутся отдельными псевдоклаузами), однако в целом подходит для нашей задачи, так как мы ожидаем, что внутри ДФ нет знаков препинания, в то время как сама ДФ отделена знаком препинания. Такое разделение обеспечивает цельность ДФ, а кроме того делает ДФ отдельным объектом для предсказания.

После того как проведено разделение на псевдоклаузы, для каждой из них определяется ряд признаков. До сих пор мы говорили о признаках ДФ, которые использовали разметчики. Для автоматического извлечения ДФ с помощью машинного обучения нужна их большая формализованность. Поэтому, основываясь на тех представлениях о ДФ, которые мы обсуждали в предыдущих разделах, в качестве формальных характеристик, соотносящихся с обоснованными в предыдущем разделе свойствами, мы используем следующие признаки ДФ:

1) часто содержит определенные части речи, например частицы, и конкретные словосочетания и слова, например *да* (в значении усилительной частицы) или *ну*;
2) представляет собой относительно короткую синтаксическую единицу;
3) является восклицанием или вопросом;
4) обладает синтаксической ущербностью (может не содержать предиката или поверхностно выраженных субъекта и объекта);
5) часто содержит глагол в императиве;
6) находится в начале реплики.

В соответствии с этими характеристиками ДФ для каждой псевдоклаузы формируется таблица признаков:

1) сам текст псевдоклаузы в нижнем регистре без знаков препинания, из которого в дальнейшем получаются два набора текстовых признаков:
   a) количество отдельных слов и биграмм в тексте псевдоклаузы;
   b) количество различных символьных 3- и 4-грамм в тексте псевдоклаузы;
2) длина псевдоклаузы в токенах;
3) наличие восклицательного знака в конце псевдоклаузы;
4) наличие вопросительного знака в конце псевдоклаузы;
5) наличие глагола;
6) наличие глагола в императиве;
7) наличие субъекта – слова, определенного автоматическим анализатором как стоящего в именительном падеже и согласующегося с глаголом в лице, числе или роде;
8) наличие объекта – слова без предшествующего предлога, определенного автоматическим анализатором как стоящего в винительном падеже при переходном глаголе;
9) положение псевдоклаузы среди первых трех в реплике;
10) количество слов разных частей речи в псевдоклаузе.

Для морфологического анализа использовался модуль pymorphy2 [Korobov, 2015], части речи обозначены по системе, используемой в этом морфологическом анализаторе.

Кроме таблицы признаков, формируется и бинарная целевая переменная – каждой псевдоклаузе приписывается класс 1, если псевдоклауза является ДФ, или 0 в противном случае.

Получение предсказаний

Модель машинного обучения была обучена на 34-х пьесах с наиболее высоким уровнем согласованности разметчиков. Согласованность разметчиков измерялась с помощью каппы Коэна [Cohen, 1960]. Полученное значение каппы Коэна по всем псевдоклаузам для выбранных 37-ми пьес составило 0,565, что по шкале, предложенной в [Landis, Koch, 1977], относится к классу «moderate agreement». Согласованность разметчиков могла нарушаться из-за неоднородности ДФ с точки зрения структуры, а также нетривиальности определения границ класса ДФ. Так, выражения *черт побери*, *черт бы его побрал*, *попытка не пытка*, *зуб даю* и многие другие могут быть рассмотрены и как периферийные элементы класса ДФ, и как речевые формулы [Баранов, Добровольский, 2000]. Кроме того, спорные случаи представля-

ют собой конструкции вида *ну да*, *конечно!*, поскольку их можно интерпретировать двояко: как одну ДФ или как соседство ДФ и коммуникатива, что во многом зависит от интонационного оформления подобных единиц.

Объем обучающей выборки составил 100 002 псевдоклаузы, из них 2 515 ДФ, для тестирования были использованы 3 пьесы объемом 13 931 псевдоклауза, из них 337 ДФ. ДФ составляют всего около 2,5 % от общего числа псевдоклауз. Несбалансированное распределение этих классов затрудняет классификацию, так как алгоритмы машинного обучения построены так, чтобы увеличить долю верных ответов (accuracy) на обучающей выборке за счет минимизации значения функции потерь, однако при этом они не учитывают цену ошибки. Для повышения качества таким алгоритмам будет выгоднее отнести все объекты к более многочисленному классу, так как это повысит долю верных ответов. Таким образом, показатели качества (точность и полнота) внутри немногочисленного класса могут быть низкими, притом что общая оценка точности алгоритма будет высокой. В данном случае, если отнести абсолютно все псевдоклаузы к обычным и не выделить ни одной ДФ, доля верных ответов составит около 97,5 % – подобный высокий показатель не отражает действительного качества предсказания. Поэтому для оценки качества предсказаний использовались точность (precision), полнота и сбалансированная F-мера (F1-мера, среднее гармоническое точности и полноты) внутри класса ДФ. Следует отметить, что формулы не являются выбросами и не определяются алгоритмами выявления выбросов – применение алгоритма IsolationForest показало F1-меру 0 на тестовой выборке.

В результате преобразования текста клауз в набор бинарных признаков полная матрица признаков состоит из 28 696 признаков, большинство из которых являются текстовыми и, предположительно, не несут ценной для предсказания информации. Поэтому одним из вариантов предобработки данных стал выбор признаков, который производится с помощью логистической регрессии с L1-регуляризацией. В результате обнуления весов незначимых признаков общее количество признаков снизилось до 2 142, что уменьшает время обучения модели и предсказания значений.

Так как общей целью является составление полного списка ДФ, необходимо сохранять относительно высокое значение полноты, чтобы бóльшая часть формул не была утеряна и попала в список. В данной работе мы максимизируем F1-меру до тех пор, пока полнота остается приемлемой для данной задачи.

Наилучшее качество показало равновесное голосование четырех классификаторов из библиотеки scikit-learn для Python 3 – Random Forest Classifier, Logistic Regression, Ridge Classifier и Support Vector Classifier, обученных на полных данных. Во всех классификаторах подобран параметр class_weight для учета дисбаланса классов. По теоретическим соображениям к предсказаниям применяется одно правило: все псевдоклаузы дальше третьей в реплике помечаются нулем, т. е. не являются ДФ. В табл. 1 представлены показатели каждого классификатора и их голосования, обученных на полной матрице признаков. Голосование классификаторов показывает максимальную F1-меру и приемлемое соотношение точности и полноты.

Таблица 1

Сравнение классификаторов

Классификатор	Точность	Полнота	F1-мера
Random Forest	0,27	0,73	0,39
Logistic Regression	0,28	0,73	0,40
Ridge Classifier	0,26	**0,80**	0,39
SVC	0,28	0,75	0,41
Vote	**0,30**	0,73	**0,42**

В табл. 2 представлено сравнение качества работы разных моделей. Baseline – это базовый подход: правило, по которому ДФ считается псевдоклауза, которая входит в первые три псевдоклаузы реплики, после которой стоит восклицательный или вопросительный знак

и в которой содержатся союзы или частицы. NoText – это голосование четырех классификаторов на всех признаках кроме текстовых (25 признаков). NoSelection – голосование классификаторов на полном наборе признаков, Selection – голосование классификаторов на отобранных признаках.

Таблица 2

Сравнение базового подхода и моделей,
обученных на разных подмножествах признаков

Модель	Точность	Полнота	F1-мера
Baseline	0,18	0,25	0,21
NoText	0,12	**0,86**	0,20
NoSelection	**0,30**	0,73	**0,42**
Selection	0,29	0,74	0,41

Генерация итоговой таблицы

Для всех формул, предсказанных голосованием модели, с помощью регулярного выражения определяется левый контекст. Это необходимо для того, чтобы в дальнейшем вручную определить, действительно ли псевдоклауза является ДФ. Кроме того, составляется список уникальных ДФ, который требуется для составления общего списка ДФ. Для использования системы создано веб-приложение [3], выделяющее ДФ в одной пьесе, и оффлайн-версия, которая позволяет получить список ДФ по нескольким пьесам. Исходный код онлайн- и оффлайн-версий системы находится в открытом доступе [4].

Ошибки, **допущенные системой**

Мы рассмотрим ошибки двух типов: ложноположительные и ложноотрицательные результаты. Среди положительных ответов системы около двух третей являются ложноположительными, т. е. ДФ, выделенные системой, не являются ДФ на самом деле. В большинстве случаев неверное решение принимается на основе формальных признаков псевдоклаузы. Среди размеченных вручную ДФ система не определяет около четверти ДФ, что зачастую связано с формальными характеристиками псевдоклауз, которые отличаются от прототипических признаков ДФ.

Ложноположительные результаты для класса ДФ

Наиболее часто алгоритм ошибочно считал ДФ псевдоклаузы, которые имеют следующий вид: *тут что-то не так*, *с чего ж и жить*, *вот тут что-то*, *может быть*, *сколько можно* и др. Все перечисленные клаузы обладают некоторыми формальными признаками, свойственными ДФ: они относительно коротки (длиной в несколько токенов), содержат слова служебных частей речи и десемантизированные слова или целиком состоят из таких слов. Некоторые псевдоклаузы являются омонимичными ДФ; иными словами, в некоторых контекстах они являются ДФ, а в других, к примеру, вводными сочетаниями. Рассмотрим несколько примеров из текстов выборки.

Псевдоклауза не является ДФ и не омонимична ей:

(11) [Алексей, муж] Да живут теперь много; все номера почти заняты. [Ихарев, муж] **Кто же именно?** (Николай Гоголь. Игроки (1842))

[3] http://web-corpora.net/wsgi3/discourse-formulas/
[4] https://github.com/kategerasimenko/discourse_formulas

(12) [Федор Иванович, муж] Вроде нет ничего. Сейчас посмотрю. (Роется в пиджаке, посматривая на жену.) [Таисия Петровна, жен (глядя в сторону)] Только у нас ведь не сберкасса! [Иванов, муж] Отдам же, ну. [Федор Иванович, муж] **Вот тут что-то...** Два рубля (Людмила Петрушевская. Уроки музыки (1989))

Выделенные псевдоклаузы являются вводными конструкциями, а не ДФ:

(13) Да? У меня, **может быть**, другие планы (Людмила Петрушевская. Уроки музыки (1989))

(14) Что ж вы, **в самом деле**, меня уж за дурака считаете? (Николай Гоголь. Игроки (1842))

Псевдоклауза включена в структуру предложения и не является ДФ:

(15) [Федор Иванович.] Просто сколько можно-то? [Николай.] **Сколько можно**, столько и можно (Людмила Петрушевская. Уроки музыки (1989))

Помимо этого, довольно часто алгоритм считал ДФ псевдоклаузы, выражающие реакцию говорящего на собственные слова, а не на слова собеседника:

(16) Все маме скажу, **вот увидишь!** (Людмила Петрушевская. Уроки музыки (1989))

В целом, чаще всего алгоритм совершает ошибки первого рода в случаях, когда принадлежность клаузы к классу ДФ определяется не свойствами самой клаузы, а свойствами ее контекста, прагматической функцией и типом семантической связи с соседними клаузами.

Вероятно, дополнение используемого набора признаков признаками, характеризующими наличие и тип связи между псевдоклаузами в тексте, а также признаками, характеризующими контекст, улучшит качество классификации, уменьшив число ошибок первого рода. Качество также может повысить усовершенствование алгоритма деления текста на псевдоклаузы.

Ложноотрицательные результаты для класса ДФ

Клаузы, являющиеся ДФ, но не выявленные алгоритмом, демонстрируют отклонение от прототипического поведения ДФ по одному или нескольким признакам:

1) полноценная или близкая к полноценной синтаксическая структура: наличие подлежащего и / или сказуемого, выраженного глаголом в финитной форме сослагательного наклонения (*а вот посмотрим*, *а я откуда знаю*, *вот тут ты ошибаешься*, *правду говорю*);

2) относительно большая длина – 4 и больше токенов (*это вы зря так думаете*, *а на кой он нам нужен*, *мое дело вообще сторона*);

3) состоит в основном из полнозначных слов (*мое дело вообще сторона*, *прошу прощения*, *опять верно*);

4) стоит не на первом месте в реплике:

(17) Сейчас инструмент раскурочит. **Так его! Так его!** (Людмила Петрушевская. Уроки музыки (1989))

(18) Как хочешь, мне-то что. **Мое дело вообще сторона**. Тебя мать пригласила (Людмила Петрушевская. Уроки музыки (1989))

В примере (18) *как хочешь* и *мне-то что* также являются ДФ, но, в отличие от *мое дело вообще сторона*, были верно отнесены к классу ДФ, между тем последняя также является ДФ, синонимичной первым двум.

Заключение и перспективы

В статье рассмотрена проблема автоматического извлечения дискурсивных формул – особых конструкций, характеризующихся отсутствием переменных внутри себя и семантико-синтаксической опорой на тип речевого акта предшествующего высказывания. Языковым материалом для решения этой задачи послужили прозаические драматические произведения

на русском языке XIX–XXI вв. Классификация той или иной единицы по ее принадлежности к ДФ представляет собой непростую задачу как в теоретическом, так и в техническом плане.

Наиболее перспективным направлением развития системы автоматического выявления ДФ, на наш взгляд, является улучшение лингвистической предобработки текстов: более сложная система деления текста на псевдоклаузы, которая уменьшит количество «шума» в рассматриваемых объектах, а также добавление признаков, основанных на левом контексте псевдоклаузы. Левый контекст для ДФ – это стимул, который вызывает ее появление, поэтому опора на определенные лексические маркеры и знаки препинания в контексте слева может улучшить качество выявления ДФ.

Кроме того, хороший результат может дать использование дополнительных технологических решений – word2vec для пословного преобразования псевдоклауз в наборы векторов и последующая классификация этих наборов с использованием свёрточной нейронной сети. Технология word2vec основывается на дистрибутивной семантике и позволяет получить представление о том, в каких контекстах чаще всего употребляется то или иное слово, и о контекстной близости пары слов. Мы предполагаем, что ДФ как единицы, обладающие общими функциями, будут появляться в схожих контекстах.

Идея применения свёрточных нейронных сетей, изначально разработанных для анализа изображений, в задачах обработки естественного языка описана в статье [Kim, 2014]. Использование нескольких свёрточных фильтров вместе с объединяющим слоем с фиксированной формой выходных данных решает проблему произвольной длины набора векторов (разное количество слов в псевдоклаузах), позволяя алгоритму на каждом шаге получать более абстрактное представление о семантике псевдоклаузы.

Обе эти стратегии – лингвистическую и технологическую – мы предполагаем использовать в последующих версиях программы.

Список литературы

Апресян Ю. Д. Типы информации для поверхностно-семантического компонента модели «Смысл ⇔ Текст» // Wiener Slawistischer Almanach, Sonderband 1. Wien: Institut für Slawistik der Universität Wien, 1980.

Апресян Ю. Д. Избранные труды. М.: Языки русской культуры, 1995. Т. 2: Интегральное описание языка и системная лексикография. 352 с.

Баранов А. Н., Добровольский Д. О. Речевые формулы в диалоге // Тр. Междунар. семинара «Диалог-2000» по компьютерной лингвистике и ее приложениям. М.: Наука, 2000. Т. 1. С. 25–31.

Рахилина Е. В., Кузнецова Ю. Л. Грамматика конструкций: теория, сторонники, близкие идеи // Лингвистика конструкций / Под ред. Е. В. Рахилиной. М.: Азбуковник, 2010. С. 18–79.

Шаронов И. А. Коммуникативы как функциональный класс и как объект лексикографического описания // Русистика сегодня. 1996. №. 2. С. 89–111.

Шаронов И. А. Дискурсивные слова и коммуникативы // Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегодной международной конференции «Диалог» (Москва, 1–4 июля 2016 г.). М.: Изд-во РГГУ, 2016. Вып. 15 (22). С. 605–615.

Biber D., Johansson S., Leech G., Conrad S., Finegan E. Longman Grammar of Spoken and Written English. Harlow: Pearson Education, 1999. 1204 p.

Cohen J. A coefficient of agreement for nominal scales // Educational and psychological measurement. 1960. Vol. 20. No. 1. P. 37–46.

Corrigan R., Moravcsik E. A., Ouali H., Wheatley K. (Eds.). Formulaic language. Amsterdam: John Benjamins Publishing, 2009. Vol. 1: Distribution and historical change. 315 p.

Fillmore Ch. J. The mechanisms of "construction grammar" // Annual Meeting of the Berkeley Linguistics Society. Berkeley, 1988. Vol. 14. P. 35–55.

Fillmore Ch. J. Grammatical construction theory and the familiar dichotomies // North-Holland Linguistic Series: Linguistic Variations. 1989. Vol. 54. P. 17–38.

Fillmore Ch. J. Border conflicts: FrameNet meets construction grammar // Proceedings of the XIII EURALEX International Congress. Barcelona: Universitat Pompeu Fabra, 2008. P. 49–68.

Fillmore Ch. J., Kay P., O'Connor M. C. Regularity and idiomaticity in grammatical constructions: The case of LET ALONE // Language. 1988. Vol. 64. No. 3. P. 501–538.

Fillmore Ch. J., Kay P. Construction Grammar Course Book. Berkeley: University of California, 1992. 113 p.

Goldberg A. Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press, 1995. 265 p.

Hoffmann T., Trousdale G. (Eds.). The Oxford handbook of construction grammar. Oxford: Oxford University Press, 2013. 586 p.

Janda L. A., Lyashevskaya O., Nesset T., Rakhilina E., Tyers F. M. A Constructicon for Russian: Filling in the Gaps // Constructicography: Constructicon development across languages / Ed. by B. Lyngfelt, L. Borin, K. H. Ohara, & T. T. Torrent. Amsterdam: John Benjamins, 2018.

Kim Y. Convolutional Neural Networks for Sentence Classification // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, 2014. P. 1746–1751.

Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages // International Conference on Analysis of Images, Social Networks and Texts. Cham, 2015. P. 320–332.

Lage L. M. Frames e construções: a relevância de um constructicon para o português brasileiro // Revista Gatilho (PPGL / UFJF. Online). 2013. Vol. 16.

Landis J. R., Koch G. G. The measurement of observer agreement for categorical data // Biometrics. 1977. Vol. 33. No. 1. P. 159–174.

Moon R. Vocabulary connections: Multi-word items in English // Vocabulary: Description, acquisition and pedagogy / Ed. by N. Schmitt, M. McCarthy. Cambridge: Cambridge University Press, 1997. P. 40–63.

Ohara K. H. Constructicon Building as a Practical Implementation of Construction Grammar and Frame Semantics: Japanese FrameNet // Poster at the 13th International Cognitive Linguistics Conference. Newcastle: Northumbria University, 2015.

Schiffrin D. Discourse markers // Studies in Interactional Sociolinguistics. 1988. No. 5. 364 p.

Schmitt N., Carter R. Formulaic sequences in action // Formulaic sequences: Acquisition, processing and use / Ed. by N. Schmitt. Amsterdam: John Benjamins, 2004. P. 1–22.

Stefanowitsch A., Gries S. Th. Collostructions: investigating the interaction between words and constructions // International Journal of Corpus Linguistics. 2003. Vol. 8, No. 2. P. 209–243.

Tomasello M. Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge, MA: Harvard University Press, 2003. 388 p.

Wray A. Formulaic language and the lexicon. Cambridge: Cambridge University Press, 2005. 348 p.

Материал поступил в редколлегию 31.03.2018

Svetlana Yu. Puzhaeva [1], **Ekaterina A. Gerasimenko** [1]
Elena S. Zakharova [1], **Ekaterina V. Rakhilina** [1, 2]

[1] *National Research University – Higher School of Economics*
21/4 Staraya Basmannaya Str., Moscow, 105066, Russian Federation

[2] *Vinogradov Institute of Russian Language RAS*
18/2 Volkhonka Str., Moscow, 119019, Russian Federation

syupuzhaeva@gmail.com, katgerasimenko@gmail.com
1583253@gmail.com, rakhilina@gmail.com

AUTOMATIC EXTRACTION OF FORMULAIC EXPRESSIONS FROM RUSSIAN TEXTS

The present paper describes automatic extraction of linguistic items we call formulaic expressions from the Russian drama texts. Particularly, by formulaic expressions (FE) we mean multiword constructions that contain no variables and are used as reactions to verbal stimuli. We consider FE

to be a specific kind of constructions in the framework of Construction Grammar. Therefore, they are to be described in the Constructicon project, which is a web-platform where the constructions of one language are presented in a special way for automatic search by various aspects. To facilitate the compilation of comprehensive FE list, we developed a module for automatic FE extraction. Implementation of the module consisted of several stages, including manual annotation of dramatic texts. The first step involved describing the features of FE and their difference compared to other syntactic items such as parenthetical words, lexical verbs and meaningful parts of sentence. Afterwards, two annotators marked Fes in 24 dramatic texts and 46 texts were annotated semi-automatically. Subsequently, we used 34 dramatic texts with the highest inter-annotator agreement. The process of FE extraction involves splitting the text into the special fragments corresponding to clauses, predicting whether each fragment is an FE corresponding to a particular feature set and compiling the final list of FEs. For prediction, we use a uniform weight vote of four classifiers (Random Forest Classifier, Logistic Regression, Ridge Classifier, Support Vector Classifier), which showed the best performance compared to rule-based baseline and classifiers outside the ensemble. We also compared the prediction quality of systems based on different feature sets and used the one with all the features. The best quality currently achieved is precision 0.30 and recall 0.73 (F1-score 0.42). Further development includes improving the preprocessing stage and employing left context, where FE stimulus is located. We also consider using distributional semantic models like word2vec for word embedding and neural networks.

Keywords: formulaic expressions, construction grammar, machine learning, automatic entity extraction.

References

Apresyan Yu. D. Tipy informatsii dlya poverkhnostno-semanticheskogo komponenta modeli «Smysl ⇔ Tekst» [Information Types for Surface Semantic Components of the Model «Meaning ⇔ Text»]. *Wiener Slawistischer Almanach*, Sonderband 1. Wien, Institut für Slawistik der Universität Wien, 1980. (in Russ.)

Apresyan Yu. D. Izbrannye trudy [Selected Studies]. Moscow, Yazyki russkoy kul'tury, 1995, vol. 2: Integral Language Description and Systematic Lexicography, 352 p. (in Russ.)

Baranov A. N., Dobrovol'skiy D. O. Rechevye formuly v dialoge [Speech formulas in the dialogue]. *Trudy mezhdunarodnogo seminara Dialog-2000 po komp'yuternoy lingvistike i ee prilozheniyam* [*Proceedings of the international workshop Dialogue-2000 on computational linguistics and its applications*]. Moscow, Nauka, 2000, vol. 1, p. 25–31. (in Russ.)

Biber D., Johansson S., Leech G., Conrad S., Finegan E. Longman Grammar of Spoken and Written English. Harlow, Pearson Education, 1999, 1204 p.

Cohen J. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 1960, vol. 20, no. 1, p. 37–46.

Corrigan R., Moravcsik E. A., Ouali H., Wheatley K. (Eds.). Formulaic language: Volume 1. Distribution and historical change. Amsterdam, John Benjamins Publishing, 2009, 315 p.

Fillmore Ch. J. The mechanisms of “construction grammar”. *Annual Meeting of the Berkeley Linguistics Society*. Berkeley, 1988, vol. 14, p. 35–55.

Fillmore Ch. J. Grammatical construction theory and the familiar dichotomies. *North-Holland Linguistic Series: Linguistic Variations*, 1989, vol. 54, p. 17–38.

Fillmore Ch. J. Border conflicts: FrameNet meets construction grammar. *Proceedings of the XIII EURALEX international congress*. Barcelona, Universitat Pompeu Fabra, 2008, p. 49–68.

Fillmore Ch. J., Kay P., O’Connor M. C. Regularity and idiomaticity in grammatical constructions: The case of LET ALONE. *Language*, 1988, vol. 64, no. 3, p. 501–538.

Fillmore Ch. J., Kay P. Construction Grammar Course Book. Berkeley, University of California, 1992, 113 p.

Goldberg A. Constructions: A Construction Grammar Approach to Argument Structure. Chicago, University of Chicago Press, 1995, 265 p.

Hoffmann T., Trousdale G. (Eds.). The Oxford handbook of construction grammar. Oxford, Oxford University Press, 2013, 586 p.

Janda L. A., Lyashevskaya O., Nesset T., Rakhilina E., Tyers F. M. A Constructicon for Russian: Filling in the Gaps. B. Lyngfelt, L. Borin, K. H. Ohara, & T. T. Torrent. (Eds.). *Constructicography: Constructicon development across languages*. Amsterdam: John Benjamins, 2018.

Kim Y. Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (*EMNLP*). Doha, 2014, p. 1746–1751.

Korobov M. Morphological analyzer and generator for Russian and Ukrainian languages. *International Conference on Analysis of Images*, *Social Networks and Texts*. Cham, 2015, p. 320–332.

Lage L. M. Frames e construções: a relevância de um constructicon para o português brasileiro. *Revista Gatilho* (PPGL / UFJF. Online), 2013, vol. 16.

Landis J. R., Koch G. G. The measurement of observer agreement for categorical data. *Biometrics*, 1977, vol. 33, no. 1, p. 159–174.

Moon R. Vocabulary connections: Multi-word items in English. N. Schmitt & M. McCarthy (Eds.). *Vocabulary: Description*, *acquisition and pedagogy*. Cambridge: Cambridge University Press, 1997, p. 40–63.

Ohara K. H. Constructicon Building as a Practical Implementation of Construction Grammar and Frame Semantics: Japanese FrameNet. *Poster at the 13th International Cognitive Linguistics Conference.* Newcastle, Northumbria University, 2015.

Rakhilina E. V., Kuznetsova Yu. L. Grammatika konstruktsiy: teoriya, storonniki, blizkie idei [Introduction. Construction Grammar: Theories, Adherents, Similar Approaches]. E. V. Rakhilina (Ed.). *Lingvistika konstruktsiy* [*Construction linguistics*]. Moscow, Azbukovnik, 2010, P. 18–79. (in Russ.)

Schiffrin D. Discourse markers. *Studies in Interactional Sociolinguistics*, 1988, no. 5, 364 p.

Schmitt N., Carter R. Formulaic sequences in action. N. Schmitt (Ed.). *Formulaic sequences: Acquisition*, *processing and use*. Amsterdam, John Benjamins, 2004, p. 1–22.

Sharonov I. A. Kommunikativy kak funktsional'nyy klass i kak ob"ekt leksikograficheskogo opisaniya [Communicatives as a functional class and an object of lexicographic description]. *Rusistika segodnya* [*Russian studies today*], 1996, no. 2, p. 89–111. (in Russ.)

Sharonov I. A. Diskursivnye slova i kommunikativy [Discursive Words and Communicatives]. *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy mezhdunarodnoy konferentsii «Dialog»* (*Moskva*, *1–4 iyulya 2016 g.*) [*Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference "Dialogue 2016"* (*Moscow*, *1–4 July 2016*)]. Moscow, RGGU Press, 2016, issue 15 (22), p. 605–615. (in Russ.)

Stefanowitsch A., Gries S. Th. Collostructions: investigating the interaction between words and constructions. *International Journal of Corpus Linguistics*, 2003, vol. 8, no. 2, p. 209–243.

Tomasello M. Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge, MA, Harvard University Press, 2003, 388 p.

Wray A. Formulaic language and the lexicon. Cambridge, Cambridge University Press, 2005, 348 p.

For citation:

Puzhaeva Svetlana Yu., Gerasimenko Ekaterina A., Zakharova Elena S., Rakhilina Ekaterina V. Automatic Extraction of Formulaic Expressions from Russian Texts. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2018, vol. 16, no. 2, p. 5–18. (in Russ.)

DOI 10.25205/1818-7935-2018-16-2-5-18