Вестник ПСТГУ.
Серия III: Филология.
2021. Вып. 66. С. 11–23
DOI: 10.15382/sturIII202166.11-23

Надежда Николаевна Буйлова
преподаватель НИУ ВШЭ
Российская Федерация, 101000,
г. Москва, Мясницкая ул., д. 20
byjlovs@yandex.ru
ORCID: 0000-0002-0604-6603

Ольга Николаевна Ляшевская
канд. филол. наук
профессор НИУ ВШЭ
Институт Русского языка
им. В. В. Виноградова РАН
Российская Федерация, 101000,
г. Москва, Мясницкая ул., д. 20
olesar@yandex.ru
ORCID: 0000-0001-8374-423X

Лексико-синтаксические маркеры малых литературных жанров

Н. Н. Буйлова, О. Н. Ляшевская

Аннотация: Статья представляет корпусное исследование конструкций поверхностного синтаксиса глаголов русского языка сквозь призму их употребления в различных жанрах массовой литературы. В центре внимания — глагольное управление и конструкции с сирконстантами, лексические ограничения на заполнение слотов конструкций. Было проанализировано пять типов связей — подлежащее, прямое и косвенное дополнение, инфинитивная клауза, а также обстоятельственная клауза. Рассмотрены расширения и усечения конструкций относительно прототипической конструкции глагольного управления, изменения конструкции в результате эллипсиса в подчиненной группе, актантные деривации, мотивированные специализацией значения глагола в том или ином жанре. Для выделения конструкций-маркеров для каждого жанра была использована метрика важности лексико-синтаксических признаков классификации текстов, рассчитанная на основе модели Случайный Лес (random forest).
Составлены конструкционные портреты четырех микрожанров: детективов, любовных романов, фэнтези и научной фантастики. Показано, что тематика произведений, структура повествования, следование массовым штампам влияют не только на выбор лексики, но и на конструкционный потенциал глагола. В целом исследование представляет отправную точку для изучения взаимодействия лексики, семантики и синтаксиса на уровне микрожанра массовой литературы, а полученные данные могут быть использованы в моделях автоматической классификации текстов для электронных библиотек.

Ключевые слова: грамматика конструкций, глагольные конструкции, поверхностный синтаксис, лексико-синтаксические маркеры, автоматическое определение жанра, массовая литература, русский язык.

Определение жанров в литературе — важное направление как теоретических, так и практических исследований. Согласно определению М. М. Бахтина, жанр — это «устойчивый тематически, композиционно и стилистически тип высказывания»[1], однако для современных классификаций такое определение не является достаточным. Современная литература тяготеет к более точным дефинициям, поэтому в нашем исследовании мы будем оперировать понятием микрожанра — более конкретным определением, описывающим тематические особенности того или иного произведения. Например, в любовных романах мы остановились на микрожанре «розового романа» — произведения, сосредоточенного на романтических переживаниях без детективной или фантастической составляющей[2].

С появлением больших массивов неструктурированных данных (электронных сетевых библиотек и корпусов) возникла необходимость в новых методах определения жанра, основанных на статистических закономерностях и в меньшей мере зависящих от пользователя. В компьютерной лингвистике проблему жанровой классификации с успехом решают при помощи машинного обучения, выделяя значимые признаки, в качестве которых могут выступать как низкоуровневые, которые легко вычисляет машина (длины слов и предложений, распределение служебных слов и дискурсивных формул, размер текста[3]), так и сложные оценочные категории, которые определяют эксперты.

Массовая литература представляет особый интерес для исследователя, поскольку слабо поддается классификации — она отделяется от прочих крупных жанров (научной и публицистической литературы[4]) и от качественной литературы, однако отделение любовных романов от детективов — достаточно неоднозначная задача, к решению которой мы подошли через выделение лексико-синтаксических маркеров.

Корпус и компьютерный эксперимент по классификации микрожанров

Исследование строится на корпусе, собранном одним из авторов. Данные были получены полуавтоматически: из открытых интернет-источников были отобраны по 350 текстов, которые составителями библиотек были обозначены метками «детективы», «любовные романы», «фэнтези» и «фантастика». Далее, эти тексты были просмотрены вручную, чтобы уточнить соответствие метки

[1] Бахтин М. М. Эстетика словесного творчества. М.: Искусство, 1986. С. 255.

[2] Вайнштейн О. Розовый роман как машина желаний // Новое литературное обозрение. 1997. № 22. С. 303–331.

[3] Мангалова Е. С., Агафонов Е. Д. О проблеме выделения информативных признаков в задаче классификации текстовых документов // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2013. № 1(22). С. 96–103; Веретенников И. С., Карташев Е. А., Царегородцев А. Л. Оценка качества классификации текстовых материалов с использованием алгоритма машинного обучения «Случайный лес» // Известия Алтайского государственного университета. 2017. № 4 (96). С. 78–83.

[4] См.: Kessler B., Nunberg G., Schutze H. Automatic Detection of Text Genre // 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics: Proceedings. Morristown (N.J.), 1997. P. 32–38.

содержанию. После выбраковки некоторого количества текстов с двойными метками («иронический детектив», «романтическое фэнтези» и т. д.) и текстов-дублей в корпусе осталось 1133 текста, классифицированных в соответствии с искомыми жанрами.

Таблица 1 представляет общие свойства подкорпусов четырех жанров.

Таблица 1

Размер подкорпусов в токенах и количество извлеченных конструкций

Жанр	Размер корпуса	Количество извлеченных конструкций
Любовные романы	17 049 022	130 435
Детективы	14 756 372	203 748
Научная фантастика	32 491 090	195 612
Фэнтези	30 689 024	143 725

Для предобработки данных были использованы скрипты на языке программирования и Python (библиотека os) и программа лексико-грамматического и синтаксического анализа UDPipe[5] (модель СинТагРус для русского языка). Разметка следует принципам базового представления Универсальных Зависимостей (Universal Dependencies[6]). Отметим, что стандарт разметки задает определенные принципы выделения зависимостей — так, зависимое может иметь только одного синтаксического хозяина, и при сочинении предикатов с пересекающимся набором участников соответствующие зависимые относятся к первому по порядку, деепричастный оборот не имеет подлежащего, парцелляция размечается как отдельное предложение и т. д.

Из синтаксически размеченного корпуса были получены текстовые реализации глагольных конструкций: для каждого глагола были извлечены все его зависимые (например, *пить nsubj; obj; obl* для предложения *Они пили чай в полном молчании*). Элементами глагольной конструкции могут быть как обязательные синтаксические аргументы (актанты), так и факультативные (сирконстанты). Порядок следования элементов в конструкции не учитывается.

На следующем этапе был построен классификатор микрожанров с использованием выделенных глагольных конструкций. Мы провели серию экспериментов по автоматической классификации текстов по жанрам с использованием лингвистических признаков разного уровня и ряда стандартных алгоритмов машинного обучения. Как было показано в работе Н. Буйловой[7], автоматическая классифи-

[5] См.: Straka M., Hajic J., Strakova J. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Aachen, 2016, P. 4290–4297.

[6] См.: Nivre J. et. al. Universal dependencies v1: A multilingual treebank collection // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Aachen, 2016. P. 1659–1666.

[7] См.: Byjlova N. Amateur Prose on the Web: Verb Construction as a Feature of Genre Classification // ARANEA 2018: Web Corpora as a Language Training Tool: Proceedings / A. Butašová, V. Benko, Z. Puchovská, eds. Bratislava: Univerzita Komenského v Bratislave, 2018.

кация жанров на данном материале на основе низкоуровневых признаков (длины слов и предложений) работает хуже, чем классификация на основе лексических маркеров, ср., например, глаголы *убить*, *любить*, *заклинать*, *телепортироваться*. Но наилучшие результаты показывает метод, при котором к лексическим маркерам добавляются сведения о синтаксических конструкциях индивидуальных глаголов. Из набора методов машинного обучения, традиционно используемых для жанровой классификации, наиболее успешным оказался метод Случайный лес (random forest)[8]. Для оценки качества построенных моделей мы делили данные на 10 сегментов, из которых девять частей использовались в качестве тестовой выборки и одна — в качестве контрольной, в 10 проходов (10-fold validation). Качество построенной модели оценивалось при помощи F1-меры. Модель классификации Случайный лес на основе лексических маркеров и конструкций поверхностного синтаксиса глаголов показала качество F1 = 0,88.

Далее мы использовали метод выделения значимых признаков (variable importance), основанный на данных классификатора Случайный лес. Мы выделили глаголы с конструкциями, которые вносят наибольший вклад в машинное обучение — т. е. единицы, обладающие наибольшей жанроразличительной силой. Данные единицы могут быть названы маркерными конструкциями малых литературных жанров, и на их основе можно построить лексико-синтаксический портрет микрожанра.

Таблица 2 представляет верхнюю часть списков лексико-синтаксических признаков, отсортированных по метрике важности[9] для каждого из четырех жанров.

Таблица 2

Лексико-синтаксические маркеры жанров, ранжированные по важности на основе модели классификации Случайный лес

Фэнтези			**Научная фантастика**		
Глагол	**Конструкция**	**Важность**	**Глагол**	**Конструкция**	**Важность**
уметь	xcomp;	176,18	превращаться	nsubj;obl;	19,55
буркнуть	nsubj;	118,36	буркнуть	nsubj;	17,49
заставлять	xcomp;	83,05	вздыхать	nsubj;	11,46
интересовать	obj;nsubj;	32,74	помнить	obl;nsubj;	9,90
говорить	obl;obl;	30,23	заключаться	nsubj;obl;	6,65
становиться	nsubj;obl;	23,55	проходить	obj;obl;	5,20
касаться	obl;	23,20	проверять	obj;	4,98

Детективы			**Любовные романы**		
Глагол	**Конструкция**	**Важность**	**Глагол**	**Конструкция**	**Важность**
звонить	obl;nsubj;	617,46	краснеть	nsubj;	515,56

[8] См.: Ho T. K. Random Decision Forests // Proceedings of the 3rd International Conference on Document Analysis and Recognition. Los Alamitos (Calif.), 1995. Vol. 1. P. 278–282.

[9] Метрика важности не нормирована.

Детективы			Любовные романы		
Глагол	**Конструкция**	**Важность**	**Глагол**	**Конструкция**	**Важность**
интересовать	obj;nsubj;	214,28	уезжать	nsubj;	202,58
стать	nsubj;xcomp;advcl;	130,21	хотеться	nsubj;xcomp;	168,54
подозревать	nsubj;obj;	115,97	касаться	obl;	164,78
просить	nsubj;advcl;	105,85	говорить	obl;obl;	66,10
говорить	obl;obl;	89,68	вскрикнуть	nsubj;advcl;	64,76
заказывать	obj;	52,33	розоветь	nsubj;	60,57

Метод анализа лексико-синтаксических маркеров микрожанра

Для анализа использовались следующие виды связей.

Нами были отобраны пять типов связей — три актантных (обязательных для построения грамматически и семантически правильного предложения) и две сирконстантных (факультативных), а также их сочетания:

Nsubj (*объект, подлежащее*) — *Ты зря* ***прохаживаешься****...*
Obj (*субъект, прямое дополнение*) — *Даже как вас* ***зовут****, и то не знаю!*
Xcomp (*инфинитивная клауза*) — ***Хотел*** *ее встретить и заблудился.*
Obl (*косвенное дополнение*) — ***Ляпнул*** *Леня и тут же пожалел об этом.*
Advcl (*клауза со значением причины, обстоятельства, следствия*) — *И даже попугай сегодня* ***не показывался****, хоть и обожал скандалы.*

Мы получили список конструкций, ставших наиболее важными признаками для машинного обучения, и определили коэффициент логарифмического подобия для каждой из них. В случае если LLh был больше 3.5, конструкция считалась значимой. Полученный список был проанализирован следующим образом.

– Для каждой конструкции определялся семантический класс глагола-вершины.

– Определялась ядерная (прототипическая) конструкция, представляющая модель управления глагола в прототипическом значении. Количество элементов ядерной конструкции определяется количеством ядерных участников обозначаемой ситуации; такая конструкция представляет грамматически корректное и семантически целостное высказывание[10].

– Если конструкция, представленная в тексте, отличается от ядерной, можно говорить об усечении или расширении конструкции. В отличие от словарного подхода к диатезе[11], в нашем исследовании текстовых реализаций конструкций принимаются во внимание как актантные, так и сирконстантные участники.

– Из банка примеров извлекались предложения, соответствующие схеме.

– Примеры анализировались с опорой на тематику микрожанра.

[10] См.: Храковский В. С. Диатеза // Лингвистический энциклопедический словарь. М., 1990.

[11] Плунгян В. А. Залог и актантная деривация // Плунгян В. А. Общая морфология: Введение в проблематику: учеб. пособ. 2-е изд., испр. М.: Едиториал УРСС, 2003. С. 191–224; Тестелец Я. Г. Диатеза, залог, актантная деривация // Тестелец Я. Г. Введение в общий синтаксис. М.: РГГУ, 2001. С. 411–436.

Анализ лексико-синтаксических маркеров

В целом было отмечено, что для детективов и любовных романов выделяется больше лексико-синтаксических маркеров, чем для фэнтези и научной фантастики. Как следует из табл. 1, это не связано ни с объемом корпусов, ни с количеством выделенных глагольных конструкций. Вместе с тем это может быть объяснено с точки зрения мощности словаря и разнообразия конструкций: выдуманные миры могут диктовать свои требования к богатству языка, следовательно, предлагать другие лексемы для описания различных действий.

Предсказуемо в числе маркерных конструкций оказываются ядерные конструкции глаголов-маркеров тех или иных жанров. Например, для любовных романов таким маркером становятся конструкции глаголов <*уезжать*+nsubj> и <*краснеть*+nsubj>:

(1) ***Я уезжаю***.
(2) *И, увидев, как Эмлин* ***покраснела****, поняла, что ее догадка верна — подруга затеяла все эти пикники и игры, чтобы как можно больше времени проводить с мистером Смитсоном,*
для научной фантастики — *заключаться* и *помнить* в конструкции V+nsubj+obl:
(3) *Что за собой заметать надобно, Пантёлкин и сам понимал, но мудрость* ***заключалась*** *не в этом*.
(4) *Про камни я* ***помню***.

Однако лингвистический интерес представляют системные отличия конструкций от ядерных, а именно, усечение и расширение конструкции, а также конструкции, маркирующие изменение значения глагола.

Грамматически обусловленное усечение конструкций

Среди лексико-синтаксических маркеров микрожанров наблюдается много конструкций с опущением подлежащего. Одной из основных причин опущения подлежащего можно считать употребление глагола в причастной или деепричастной клаузе, а также в инфинитивной группе, подчиненной матричному глаголу[12]. Типичный случай — глагол *пожимать* в конструкции V+obj (ядерная конструкция V+nsubj+obj+iobj), маркер любовных романов. Глагол употребляется в специализированном значении «приветствовать кого-л., сжимая руку или часть руки» и обозначает сопровождение речевого или иного основного действия. Подлежащее здесь выражается в главной клаузе, при этом большая часть употреблений приходится на деепричастные обороты, что обусловливает отсутствие в конструкции субъекта-подлежащего:

(5) *Как поживаете, сэр? — проговорил Джулиан,* ***пожимая*** *руку.*

Другим фактором опущения подлежащего можно назвать сочинительную связь между однородными клаузами. По правилам синтаксического представления предложения, у зависимых может быть только один синтаксический хозяин,

[12] См.: Ляшевская О. Н., Кашкин Е. В. Типы информации о лексических конструкциях в системе ФреймБанк // Труды Института русского языка им. В. В. Виноградова. 2015. № 6. С. 464–556; Русская грамматика: в 2 т. М., 1980.

и это первый элемент сочинения. Так, у глагола *целовать* в любовных романах выделяется бессубъектная конструкция-маркер obj;obl;advcl; (полная конструкция V+nsubj+obl). Анализ корпуса показывает, что в этом случае *целовать* системно употребляется в качестве второго члена сочинения:

(6) *Не задумываясь, она* ***поднялась*** *на цыпочки и* ***поцеловала*** *его в щеку, коснувшись губами прохладной кожи, отдающей свежим ветром и потом.*

Маркерная конструкция глагола *приглашать* для любовных романов и детективов — V+obj+obl (ядерная конструкция V+nsubj+obj+xcomp). Здесь опущение субъекта может быть связано в том числе с неопределенно-личной конструкцией, см. пример (7).

(7) *Нет балов и вечеров, по крайней мере, меня на них не* ***приглашают****.*

Изменение конструкции

Вышеуказанные примеры демонстрируют стратегии опущения подлежащего, однако также возможны и другие изменения конструкций, связанные с эллипсисом. В конструкции-маркере глагола *приглашать* вместо инфинитивного зависимого (xcomp) употребляется предложная именная группа (obl), см. пример (8).

(8) *Сергей очень любил всевозможные технические новинки, — пояснила мне Мария, распахивая дверь и* ***приглашая*** *меня в дом.*

Наблюдается эллипсис инфинитива (ср. *приглашая меня зайти в дом*), в результате которого косвенное дополнение становится синтаксическим актантом глагола *приглашать.* Можно говорить о специализации значения глагола *приглашать* в жанрах любовного романа и детектива: это всегда приглашение к перемещению того или иного рода (зайти, заехать и т. п.).

Расширение конструкции

Отмечаются не только опущения, но и дополнения конструкции: добавляются в основном сирконстанты со значением места, времени, условия и т. д. Вышеупомянутая конструкция глагола *целовать* интересна не только с точки зрения ее усечения относительно ядерной, но и с точки зрения ее усложнения. Так, элемент конструкции *advcl* в примере (6) является вершиной деепричастной клаузы и обозначает действие, сопровождающее поцелуй. Учитывая, что выражен и такой достаточно периферийный участник фрейма, как точка контакта (*в щеку*, obl), можно высказать гипотезу, что авторы массовых любовных романов стремятся к детальности проработки сцены. В целом усложнение конструкции глагола может быть связано со следующей спецификой жанра: нормативный словарь для описания любовных сцен достаточно скуден, и усложнение синтаксиса служит противовесом простоте лексики.

Актантные деривации, связанные с системным сдвигом значения

Выше мы уже отмечали, что в отдельных жанрах может происходить специализация лексического значения глагола. В более общем ключе можно утверждать, что жанр задает construal[13], способ видения определенной ситуации, и это может иметь предсказуемые синтаксические эффекты.

Вернемся к маркерным конструкциям с усеченным составом актантов. Например, глагол *выезжать* (прототипическая конструкция: V+nsubj+obl+obl) в конструкции, характерной для детективов и любовных романов, имеет только одно косвенное дополнение:

(9) *Сев за руль, я резко стартанула и вскоре уже* ***выезжала*** *из города.*
(10) *Одним нравится* ***выезжать*** *на балы, другим приглашать гостей к себе.*

Глагол *выезжать* — яркий пример реализации жанрохарактеризующей лексики. В одной и той же конструкции реализуются два значения глагола, первое из которых, в детективных романах, более общее, второе, в любовных романах, более частное. В первом случае мы имеем дело с общим значением перемещения, и зависимое (obl) обозначает либо начальную, либо конечную точку, в зависимости от фокуса повествования. Во втором случае выражается более локальное значение, которое, к тому же, четко маркирует время происходящих в любовных романах событий — *выезжают* почти исключительно *в свет* или *на бал*.

Другой схожий по конструкции глагол — это *пить,* реализующий свою полную конструкцию (V+nsubj+obj). Данный глагол употребляется в двух конструкциях, близких по значению, но различающихся по смыслу. Вышеуказанная маркерная конструкция представляет собой устойчивое сочетание с существительным «чай» (*Думаю, мы не будем пить чай?*). Вторая группа — это коллокация «X пьет за Y» (из-за особенностей модели прямые и косвенные дополнения могут размечаться одинаково), в которой глагол «пить» обозначает ситуацию «употреблять алкогольные напитки»:

(11) *Они* ***пьют*** *за упокой ее грешной души*.

Глаголы с несколькими маркерными конструкциями

Отдельное внимание стоит уделить глаголам с несколькими маркерными конструкциями. Это позволяет выделить многозначные глаголы и, следовательно, точнее обозначить тематику произведения. К таким относится, к примеру, глагол *касаться*, в своей ядерной конструкции (V+nsubj+obj) характерный для любовных романов (12) и детективов (13). Эта конструкция в большинстве случаев реализует переносное значение глагола ‘иметь отношение к чему-либо’:

(12) *Джойслин делался невероятно упрямым, когда дело* ***касалось*** *его жены.*
(13) *Пока наш интерес* ***касался*** *только угона «шестерки» с автостоянки.*

[13] Langacker R. W. Grammar and conceptualization. Berlin; New York: Walter de Gruyter, 2010. (Cognitive Linguistics Research; 14.)

Прототипическое значение глагола *касаться* часто отмечается в описании романтических сцен. В обоих значениях полная аргументная конструкция одинакова (V+nsubj+obj), но маркером любовных романов и фэнтези становится усеченная конструкция с опущением подлежащего:

(14) *Легко касаясь его тела, она вновь узнавала его.*
(15) *Слышали в гостях что-нибудь интересное? — спросил он, поигрывая прядкой на ее виске,* ***касаясь*** *жемчужной серьги.*

Объяснение состоит в том, что физический контакт сопровождает другое действие — опущение субъекта грамматически обусловлено употреблением глагола в качестве деепричастия (15) или инфинитива при матричном глаголе (14).

Конструкционный потенциал лексических групп глагола

В качестве обобщения в следующем разделе хочется обсудить некий «глагольный портрет жанра» с точки зрения семантических сдвигов, конструкционного потенциала и набора тематик.

В наших данных мы видим некоторые закономерности, связанные с основными темами исследуемых жанров. Например, глаголы речи более характерны для детективов и любовных романов, которые сконцентрированы на взаимодействии между героями. При этом отдельные глаголы также переходят в класс «сопровождение речи», что также подчеркивает значение коммуникации в этих романах. Общий конструкционный профиль таких глаголов разнообразен, но в целом глаголы речи придерживаются прототипической конструкции «X сообщает Y-у о Z-е» при косвенной речи и «сообщил X» при прямой (в ремарках и словах автора).

Бытийные глаголы тяготеют к конструкциям с полностью заполненными валентностями; такие конструкции характеризуют по большей части любовные романы, фэнтези и научную фантастику (возможно, из-за некоторой тяги этих жанров к пассивным конструкциям).

Глаголы движения с любыми конструкциями нехарактерны для научной фантастики (кроме глагола *проходить* с невысоким LLh) — одним из объяснений может быть как раз тематическая направленность этого жанра, сконцентрированной на активных перемещениях между локациями. В итоге мы получаем множество глаголов движения, которые не имеют высокого коэффициента — и в таком случае необходимо говорить не об отдельных лексемах, а о классе в целом.

Глаголы контакта ожидаемо маркируют любовные романы (что соответствует нашим ожиданиям от топика) и при этом зачастую используются в составе причастного оборота, т. е. в качестве дополнительного действия при основном.

Ментальные глаголы с полным заполнением валентностей и высоким LLh маркируют детективные романы, что также может быть связано с темой расследований и преступлений, а также с постоянным напоминанием читателю, кто и что расследует/обдумывает/решает.

Модальные глаголы с комплиментарной клаузой характерны для любовных романов, что соответствует нашим ожиданиям от этого микрожанра (X хочет, чтобы Y).

Посессивные глаголы более распространены в детективах, что может быть обусловлено временной спецификой: большая часть детективов описывает наше время и современные товарно-денежные отношения, в отличие от фантастических или романтизированных реальностей иных жанров.

Заключение

Машинное обучение зачастую воспринимается как практический инструмент для решения прикладных задач. Однако использование некоторых его особенностей может быть полезно и для задач теоретических. В работе предложен метод сбора материала глагольных конструкций, идущий от корпусных данных (bottom-up method) и использующий технологии автоматической классификации текстов. С его помощью были получены новые данные по конструкциям-маркерам жанров, которые стали предметом последующего качественного исследования.

Жанр произведения представляет различные тематики: преступления, романтические переживания, магический мир, достижения науки. Выбор темы накладывает ограничения на выбор лексики и лексической конструкции. В корпусных данных это проявляется в высокой частоте жанроспецифичных глаголов (ср. *изучать, целовать, кастовать, аннигилировать*). Кроме того, с тематикой произведения связано отдельное значение (фрейм) лексической единицы, а также выбор его синтаксической конструкции. Один и тот же глагол в различных синтаксических конструкциях может быть маркером разных жанров.

Авторство также накладывает отпечаток на выбор лексики и конструкций. В нашем исследовании мы постарались минимизировать вклад стиля, ограничив количество произведений одного автора пятью книгами.

Разные жанры имеют разные схемы развития сюжета. К примеру, любовным романам присуща линейная структура повествования с ключевыми сценами (эмоционально насыщенными актами взаимодействия между главными героями, ср. *целовать V+obj+obl+advcl*) и сценами-филлерами (необходимыми для развития сюжета, с более скупым лексическим оформлением и более простым синтаксисом глагольных групп, ср. *вспоминать+obl*). Это обусловливает наличие двух групп глаголов с различными синтаксическими конструкциями.

В более широком ключе можно предположить, что различные жанры имеют разную структуру нарратива, диалога и других литературных форм, что проявляется в наблюдаемых сдвигах распределения лексико-синтаксических единиц.

Не стоит недооценивать литературные тренды и связанные с ними языковые штампы (или, шире, паттерны, которым неосознанно следует или подражает автор). Этот фактор также может определять лексико-синтаксическое своеобразие жанра. Важным поджанром детективов 1990–2000-х гг. можно назвать «женские» детективы (литературная маска М. Серова), в которых выбор лексики смещен в сторону топика «расследование» в противовес «крутым» детективам (топики «насилие» и «вооруженные столкновения»).

В нашем исследовании рассматриваются четыре литературных микрожанра (детективы, любовные романы, фэнтези и фантастика). Не останавливаясь

специально на лексических различиях жанров, мы показали, что конструкции поверхностного синтаксиса отдельных глаголов русского языка и лексических групп системно отличаются от жанра к жанру. В частности, некоторые привычные штампы массовой литературы, конфигурации конструкций уровня клаузы в тексте могут объяснять грамматически обусловленные опущения синтаксического субъекта, приращение конструкции за счет обстоятельственных аргументов и другие изменения в текстовых реализациях конструкций.

Безусловно, предлагаемый метод не свободен от недостатков. Ограничением метода является качество автоматической разметки, а также объем, состав корпусной выборки и сам набор сопоставляемых микрожанров. Однако в целом, это исследование представляет отправную точку для изучения взаимодействия лексики, семантики и синтаксиса на уровне микрожанра массовой литературы.

В прикладном отношении полученные данные о роли поверхностно-синтаксических паттернов могут быть использованы в моделях автоматической классификации текстов для электронных библиотек.

Список литературы

Бахтин М. М. Эстетика словесного творчества. М.: Искусство, 1986.

Вайнштейн О. Розовый роман как машина желаний // Новое литературное обозрение. 1997. № 22. С. 303–331.

Веретенников И. С., Карташев Е. А., Царегородцев А. Л. Оценка качества классификации текстовых материалов с использованием алгоритма машинного обучения «Случайный лес» // Известия Алтайского государственного университета. 2017. № 4 (96). С. 78–83.

Ляшевская О. Н., Кашкин Е. В. Типы информации о лексических конструкциях в системе ФреймБанк // Труды Института русского языка им. В. В. Виноградова. 2015. № 6. С. 464–555.

Мангалова Е. С., Агафонов Е. Д. О проблеме выделения информативных признаков в задаче классификации текстовых документов // Вестник Томского государственного университета. Управление, вычислительная техника и информатика. 2013. № 1 (22). С. 96–103.

Плунгян В. А. Залог и актантная деривация // Плунгян В. А. Общая морфология: Введение в проблематику: учеб. пособ. 2-е изд,, испр. М.: Едиториал УРСС, 2003. С. 191–224.

Русская грамматика: в 2 т. М.: Наука, 1980.

Тестелец Я. Г. Диатеза, залог, актантная деривация // Тестелец Я. Г. Введение в общий синтаксис. М.: РГГУ, 2001. С. 411–436.

Храковский В. С. Диатеза // Лингвистический энциклопедический словарь. М., 1990.

Byjlova N. Amateur Prose on the Web: Verb Construction as a Feature of Genre Classification // ARANEA 2018: Web Corpora as a Language Training Tool: Proceedings / A. Butašová, V. Benko, Z. Puchovská, eds. Bratislava: Univerzita Komenského v Bratislave, 2018.

Ho T. K. Random Decision Forests // Proceedings of the 3rd International Conference on Document Analysis and Recognition. Los Alamitos (Calif.), 1995. Vol. 1. P. 278–282.

Kessler B., Nunberg G., Schutze H. Automatic Detection of Text Genre // 35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics: Proceedings. Morristown (N.J.), 1997. P. 32–38.

Langacker R. W. Grammar and conceptualization. Berlin; New York: Walter de Gruyter, 2010. (Cognitive Linguistics Research; 14.)

Nivre J. et. al. Universal dependencies v1: A multilingual treebank collection // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Aachen, 2016. P. 1659–1666.

Straka M., Hajic J., Strakova J. UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Aachen, 2016. P. 4290–4297.

Vestnik Pravoslavnogo Sviato-Tikhonovskogo gumanitarnogo universiteta.
Seriia III: Filologiia.
2021. Vol. 66. P. 11–23
DOI: 10.15382/sturIII202166.11-23

Nadezhda Builova,
Teacher,
National Research University
"Higher School of Economics",
20 Miasnitskaya, Moscow,
101000, Russian Federation
byjlovs@yandex.ru
ORCID: 0000-0002-0604-6603

Olga Lyashevskaya,
Candidate of Sciences in Philology,
Professor,
V. V. Vinogradov Institute of the Russian Language
of the Russian Academy of Sciences;
National Research University
"Higher School of Economics"
20 Miasnitskaya, Moscow,
101000, Russian Federation
olesar@yandex.ru
ORCID: 0000-0001-8374-423X

Lexical and Syntactic Markers of Small Literary Genres

N. Builova, O. Lyashevskaya

Abstract: This article presents a case study of surface syntax constructions of Russian verbs with respect to their use in various genres of fiction. The article primarily deals with verb complementation and constructions with adjuncts, as well as with lexical constraints on the construction slots. Five types of dependency relations are taken into consideration, i.e. subject, direct and oblique object, infinitive clause, and adverbial clause. The prototypical constructions of the verb complementation are mapped to reduced and extended constructions, other changes such as an ellipse in a subordinate group, and to actant derivations motivated by the specialisation of the lexical sense of the verb in a particular genre. In order to identify constructions for the analysis, we used the Random Forest classification method and calculated the metric of importance for the surface syntax verb constructions. The article also outlines the constructional portraits of four small genres, i.e. detective stories, romance novels, fantasy, and science fiction. It is shown that the topics, the structure of the narrative, the authors' use of cliches, the way how the text follows the literary trends affect not only the choice

of vocabulary, but also the constructional potential of the verb. In general, the article provides a starting point for studying the interaction of lexicon, semantics, and syntax at the level of microgenres in fiction. The data obtained can be used in models of automatic text classification for electronic libraries.

Keywords: construction grammar, verb constructions, surface syntax, lexical-syntactic markers, automatic genre classification, pop fiction, Russian language.

References

Bakhtin M. (1986) *Estetika slovesnogo tvorchestva* [Aesthetics of Verbal Art]. Moscow (in Russian).

Byilova N. (2018) "Amateur prose on the Web: Verb construction as a feature of genre classification", in A. Butašová, V. Benko, Z. Puchovská (eds) *ARANEA 2018: Web corpora as a language training tool: Proceedings*, Bratislava: Univerzita Komenskeho v Bratislave.

Ho T. K. (1995) "Random Decision Forests", in *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Los Alamitos (Calif.)*, vol. 1, pp. 278–282.

Kessler B., Nunberg G., Schutze H. (1997) "Automatic Detection of Text Genre", in *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics: Proceedings*, Morristown (N.J.), pp. 32–38.

Khrakovskii V. (1990) "Diateza" [Diathesis], in *Lingvisticheskii entsiklopedicheskii slovar'* [Encyclopaedic dictionary of linguistics], Moscow (in Russian).

Langacker R. W. (2010) *Grammar and conceptualisation*. Berlin; New York: Walter de Gruyter. (Cognitive Linguistics Research; 14).

Lyashevskaya O., Kashkin E. (2015) "Tipy informatsii o leksicheskikh konstruktsiiakh v sisteme FrameBank" [Types of information on lexical patterns in the system FrameBank]. *Trudy Instituta russkogo iazyka im. V. V. Vinogradova*, 2015, no. 6, pp. 464–556 (in Russian).

Mangalova E., Agafonov E. (2013) "O probleme vydeleniia informativnykh priznakov v zadache klassifikatsii tekstovykh dokumentov" [Identifying informative features in the classification of textual documents]. *Vestnik Tomskogo gosudarstvennogo universiteta. Upravlenie, vychislitel'naia tekhnika i informatika*, 2013, no. 1 (22), pp. 96–103 (in Russian).

Nivre, J. et al. (2016) "Universal dependencies v1: A multilingual treebank collection", in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, Aachen, pp. 1659–1666.

Plungian V. (2003) *Obshchaia morfologiia* [Introduction to morphology]. Moscow (in Russian).

Russkaia grammatika [Russian grammar] (1980). Moscow (in Russian).

Straka M., Hajic J., Strakova J. (2016) "UDPipe: trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging, and parsing", in *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*, Aachen, 2016, pp. 4290–4297.

Testelets Ia. (2001) *Vvedenie v obshchii sintaksis* [Introduction to general syntax]. Moscow (in Russian).

Vainshtein O. (1997) "Rozovyi roman kak mashina zhelanii" [Pink novel as a machine of desires]. *Novoe literaturnoe obozrenie*, 1997, no. 22, pp. 303–331 (in Russian).

Veretennikov I., Kartashev E., Tsaregorodtsev A. (2017) "Otsenka kachestva klassifikatsii tekstovykh materialov s ispol'zovaniem algoritma mashinnogo obucheniia 'Sluchainyi les'" [Evaluation of quality of classification of textual materials when using the algorithm of machine learning 'Random Forest']. *Izvestiia Altaiskogo gosudarstvennogo universiteta,* 2017, no. 4 (96), pp. 78–83 (in Russian).