АВТОМАТИЧЕСКИЙ АНАЛИЗ МЕДИЙНЫХ ТЕКСТОВ

УДК 81’42

Тематическая организация текста инструкций как лингвистическая основа приобретения знаний интеллектуальным агентом*

Л. А. Каджая[1,2], *Ю. М. Кузнецова*[3], *В. А. Салимовский*[1], *М. И. Суворова*[3]

[1] Пермский государственный национальный исследовательский университет, Российская Федерация, 614068, Пермь, ул. Букирева, 15
[2] Шандуньский университет, Китайская Народная Республика, 264209, Вэйхай, Вэнхуа Си Лу, 180
[3] Федеральный исследовательский центр «Информатика и управление» РАН, Российская Федерация, 117312, Москва, пр. 60-летия Октября, 9

Для цитирования: Каджая, Л. А., Кузнецова, Ю. М., Салимовский, В. А., Суворова, М. И. (2021). Тематическая организация текста инструкций как лингвистическая основа приобретения знаний интеллектуальным агентом. *Медиалингвистика, 8* (1), 45–56.
https://doi.org/10.21638/spbu22.2021.104

Исследуется тематическая организация инструктивных текстов в аспекте проблематики, актуальной для работ по созданию когнитивного ассистента. Назначение ассистента — предоставить пользователю в соответствии с его поисковым запросом информацию, необходимую для следования правилам определенного сценария, способствующим успешному достижению поставленной цели. Уточняемый по мере решения задачи запрос, содержащий те или иные ключевые слова, ориентирован на развернутый набор тем, маркирующих предметные области, отраженные в сценарии. Предлагается обзор лингвистических работ, посвященных вопросам тема-рематического структурирования продуцируемого текста, а также его компрессии, пределом которой являются ключевые слова. Подчеркнуто значение описания тематических цепочек текста для получения детальной объективной информации о его тематической структуре. При сопоставлении списка ключевых слов, выявленных автоматической системой TextAppliance в коллекции инструктивных текстов, извлеченных из интернета, с результатами ручного анализа этих текстов, определяющего место различных номинативных единиц в тематической организации речевого произведения, были установлены наиболее значимые характеристики

* Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-00-00606 (18-00-00233).

ключевого слова, проявляющиеся у разных номинативных единиц в разной степени. Это высокий показатель идентификатора текста, содержательная емкость, а также коммуникативная значимость слова или субстантивного словосочетания как маркера важной для адресата информации. Определение ключевых слов в целых инструктивных текстах и в относительно автономных текстовых фрагментах (субтекстах), описывающих отдельные этапы достижения поставленной пользователем цели (например, этапов выбора автомобиля, его осмотра, оформления сделки, регистрации машины), должно способствовать улучшению качества идентификации сценария в Сети. Извлечение ключевых слов вместе с их контекстом позволит автоматически создавать базу рекомендаций пользователю. Раскрывается значимость анализа тема-рематической структуры текста как знака для его моделирования в знаковой картине мира.

Ключевые слова: интеллектуальный агент, текст, тема-рематическая организация текста, ключевое слово, тематическая цепочка.

Постановка проблемы. Традиционными для искусственного интеллекта методами сбора информации о предметной области являются опрос экспертов и создание онтологий. В последние годы высказывается мысль о возможности принципиально иного подхода к разработке инструктирующих интеллектуальных систем, который не требует предварительного сбора данных разработчиком, а основывается на психологическом и лингвистическом анализе технического задания как целевой установки, развертывающейся в текст. Предполагается, что интеллектуальный агент будет получать знания из интернета под конкретную задачу. В этом случае интеллектуальная система должна быть наделена картиной мира, а ее работа, как и работа человеческого сознания, опосредована языковыми знаками [Осипов и др. 2018; Салимовский и др. 2019]. Такая система была бы гибкой, поскольку выявляла те значения, свойства и модальности, которые важны именно для данного технического задания. Она обладала бы неограниченными познавательными возможностями, обеспечиваемыми выходом в интернет — к библиотекам, архивам, социальным сетям и другим источникам информации.

Для приобретения знаний интеллектуальному агенту потребовалась бы иерархически организованная система потенциальных запросов, базирующаяся на развернутом наборе тем — маркеров предметных областей, включенных в сценарий.

Особым типом интеллектуальных агентов является когнитивный ассистент, разрабатываемый в Институте проблем искусственного интеллекта ФИЦ «Информатика и управление» РАН. Базой способностей когнитивного ассистента служит придаваемая ему знаковая картина мира, в которой содержатся его назначение, цели, возможные действия и их сценарии, смыслы, способы и результаты достижения целей. Кроме своей собственной базы, ассистент в ходе самостоятельного когнитивного анализа коммуникации и поведения пользователя строит модель его картины мира, включающую в себя отраженные ассистентом сценарии и личностные смыслы, ценности, предпочтения и привычки человека. Общение когнитивного ассистента с пользователем строится с учетом этих двух картин мира [Смирнов и др. 2019].

При создании когнитивного ассистента важно иметь в виду, что по отношению к сценарию в целом и каждой его предметной области социальным сознанием выработаны правила целесообразного поведения субъекта. Однако отдельному пользователю они обычно неизвестны в достаточном объеме. Поэтому задача

когнитивного ассистента состоит в том, чтобы в соответствии с информационными запросами пользователя, уточняющими исходное задание, транслировать эти правила как рекомендации. Иными словами, агент, владея сформировавшимися в общественном сознании правилами решения задачи, направлял бы поисковую активность пользователя, указывая ему, какую именно информацию нужно получить для успешного достижения цели.

Конечно, любой социальный сценарий в определенном смысле открытый: он может развиваться, а его слоты могут рассматриваться в «укрупненном» виде как самостоятельные сценарии со своими слотами. Однако в каждый период своего существования сценарий обладает определенностью и может быть охарактеризован с необходимой полнотой.

Как известно, сжатой формой поисковых запросов, как и текстов, содержащихся в Сети, являются ключевые слова. В психолингвистике и теории деривации ключевые слова рассматриваются как результат компрессии речевого произведения, как маркеры его цельности (и субцельностей текстовых фрагментов) [Сахарный 1982; 1992; 1994; Мурзин 1982; 1984; Мурзин, Штерн 1991].

Логично думать, что система потенциальных запросов интеллектуального агента может быть построена на основе системы ключевых слов, в компрессированном виде представляющих содержание множества инструктивных текстов по решению определенной социально значимой задачи.

Цель статьи состоит в совершенствовании метода выделения ключевых слов из корпуса текстов, реализующих речевой жанр «план-инструкция» и воплощающих сценарий деятельности субъекта. Ключевые слова выявляются на основании анализа тематических цепочек [Матвеева 1990] с учетом тема-рематического структурирования текстов. Полученный набор ключевых слов сравнивается с их набором, устанавливаемым автоматической системой TextAppliance [Мбайкоджи, Драль, Соченков 2012; Ананьева и др. 2016]. Это сравнение позволяет проанализировать основные особенности выделяемых системой слов и словосочетаний и в дальнейшем учитывать их для совершенствования методов автоматического извлечения ключевых слов из текста.

Материалом исследования послужили собранные в интернете тексты инструкций по покупке автомобиля. Корпус насчитывает 100 инструкций, содержащих описание разных этапов покупки.

История вопроса. Характеризуя текстообразование как процесс тема-рематического структурирования цельности, Л. В. Сахарный рассматривает цельность как основную категорию текста. Под цельностью он понимает «психолингвистический феномен особого рода, который представляет собой возникающее в психике человека симультанное (одновременное), интегральное, полностью не осознаваемое динамическое представление о некотором объекте» [Сахарный 1994: 20]. Осмысление цельности говорящим (пишущим) предполагает выделение из нее «смысловых вех», или субцельностей, из которых, в свою очередь, выделяются более частные субцельности[1]. С использованием категориального аппарата функционального синтаксиса этот процесс описывается как тема-рематическое структурирование:

[1] Понятия цельности и субцельности относятся к организации конкретного речевого произведения. При продуцировании же нового текста субцельность может становиться цельностью. Примечательны случаи выделения субцельностей из текста и их последующего функционирования уже

цельность выступает темой, а осмысливаемая в ней субцельность — ремой. На каждом новом шаге продуцирования текста рема становится темой, в которой обнаруживается новая рема (или ремы). Выделяемые в тексте субцельности могут рассматриваться в качестве коррелята «тематических макроструктур», анализируемых Т. А. ван Дейком [Дейк 1989].

С опорой на категорию цельности Л. В. Сахарный исследует механизм не только развертывания текста, но и его компрессии (свертывания): сохраняя свое содержательное тождество, текст при переходе от одной ступени компрессии к другой, более глубокой, лишается только маргинальных элементов своего содержания (ср.: [Леонтьев 1976; Дридзе 2009]). Результатом компрессии становится набор ключевых слов [Сахарный 1992]. Они представляют тему целого текста и подтемы его основных смысловых фрагментов.

С близких позиций к изучению текстообразования подходит Л. Н. Мурзин [Мурзин 1982; 1984]. Согласно его концепции, в речемыслительном акте некоторый неопределенный объект получает определенность благодаря приписыванию ему некоторого признака. В результате возникает новый объект, менее неопределенный, чем предшествующий. Объекту соответствует понятие темы, а признаку — понятие ремы. На уровне лексико-грамматического воплощения глубинной тема-рематической структуры действуют механизмы контаминации и компрессии. Контаминация служит транспозиции предшествующего предложения в свободную позицию последующего. Тем самым она обеспечивает развертывание текста. Компрессия, напротив, обеспечивает устранение его избыточности, создаваемой включением каждого последующего предложения в предыдущее[2]. Пределом компрессии является «слово как наиболее компактная форма репрезентации текста» [Мурзин 1982: 27].

Тема-рематическое структурирование цельности как глубинный процесс зеркально не отражается в композиционной и собственно речевой организации текста, которая определяется не только задачей раскрытия темы, но всем комплексом познавательно-коммуникативных установок, охватываемых авторским замыслом.

В предложенной Т. М. Дридзе концепции текста как иерархии коммуникативных программ [Дридзе 2009], развивающей идеи Н. И. Жинкина, в качестве предикации первого порядка рассматривается цель сообщения, предикации второго порядка — основной констатирующий тезис и аналитическая оценка ситуации, составляющие основные элементы общего содержания. Второстепенные же элементы образуют предикация третьего порядка — иллюстрации к основному тезису и предикация четвертого порядка — общий фон к цели сообщения. Автор подчеркивает, что ключевые слова «несут большую ценность с точки зрения информативности текста, если они входят в предикации высших порядков» [Дридзе 2009: 89].

Закономерности выражения темы в уже созданном тексте совокупностью тематических групп слов исследуются Т. В. Матвеевой. В ее работах эксплицируются, в частности, важные для нас понятия тематического поля текста и тематических цепо-

---

в качестве цельностей. (Примером могут служить фрагменты из пушкинского «Евгения Онегина», публикуемые как самостоятельные стихотворения о природе для детей.)

[2] Хрестоматийным примером семантической избыточности, возникающей в процессе текстопорождения, является известное стихотворение С. Маршака: «Вот дом, который построил Джек. А это пшеница, которая в темном чулане хранится в доме, который построил Джек. А это веселая птица-синица, которая часто ворует пшеницу, которая в темном чулане хранится, в доме, который построил Джек…».

чек [Матвеева 1990; 2019]. Тематическое поле образуют слова разных лексико-грамматических классов и номинативные словосочетания, обладающие общей семой. Наиболее значимы для тематического поля в семантическом и структурном отношениях непосредственные наименования предмета речи, т. е. предметные номинации.

Тема текста и его подтемы могут быть описаны в виде тематических цепочек. Основная цепочка, проходящая через весь текст, представляет его тему, а дополнительные цепочки определяют объем подтем [Матвеева 1990]. В составе тематических цепочек различаются основная номинация, наиболее точно и непосредственно обозначающая предмет речи, и дополнительные номинации, часто имеющие экспрессивно-эмоциональную окраску.

Т. В. Матвеевой охарактеризованы особенности реализации категории темы (наряду с другими текстовыми категориями) в речевых произведениях различной функционально-стилевой и жанровой принадлежности.

При решении задач в области информационного поиска и индексирования документов ключевое слово обычно определяется как «слово или словосочетание (термин) в тексте документа или запроса, несущий в нем существенную информационную нагрузку хотя бы по одной из тем, рассматриваемых в документе»[3]. Однако понятие «информационная нагрузка слова» трактуется исследователями по-разному. В большинстве случаев акцент делается на соотнесенности ключевых слов с основным содержанием текста, но нередко (в том числе в системе TextAppliance) — на их дифференцирующей функции при нахождении нужного документа.

Оценка информационной нагрузки в тексте тех или иных номинативных единиц и разработка более общей проблематики — изучение тематической организации речевого произведения, определение степени тематической близости различных произведений — это вопросы, от успешного решения которых во многом зависит совершенствование автоматического извлечения ключевых слов [Ванюшкин, Гращенко, Романишин 2019; Beliga 2015; Sterckx et al. 2019].

Для уточнения основных характеристик ключевых слов важно сравнить наборы этих единиц, установленные при разном понимании их информационной значимости.

Описание методики исследования. В системе TextAppliance вес ключевых слов определяется по формуле TF — IDF, где TF (term frequency) — частота употребления слова в анализируемом документе, а IDF (inverse document frequency) — отношение общего количества документов фоновой коллекции (т. е. текстов, содержащихся в TextAppliance) к количеству документов, в которых взвешиваемое ключевое слово встречается хотя бы один раз. Бóльшую значимость получают те слова, которые часто встречаются в анализируемом документе и относительно редко — в остальных документах коллекции. Тем самым система определяет, насколько то или иное слово специфично для рассматриваемого текста (или же изучаемого множества текстов, представляющих определенный сценарий, т. е. сверхтекста — «совокупности высказываний или текстов, объединенных содержательно и ситуативно» [Купина 2019: 374]). Иначе говоря, она устанавливает, насколько та или иная лексическая единица подходит на роль идентификатора (ключа), позволяющего обнаружить некоторый текст (или сверхтекст).

[3] ГОСТ Р 7.0.66-2010. ИСО 5963:1985. СИБИД (2010). Электронный ресурс http://docs.cntd.ru/document/1200084836.

Помимо использования указанной системы мы осуществляли ручной анализ текстов для определения места того или иного ключевого слова в их тематической организации. С этой целью анализировались тематические цепочки речевого произведения. Обращение к этим цепочкам позволяет, кроме того, выделить наиболее значимые в тематическом отношении слова, так как любая разрабатываемая автором тема (тема, актуальная для него) представлена именно цепочкой номинативных единиц — повторением одних и тех же слов, использованием синонимов, перифраз.

В тематической цепочке каждая из номинативных единиц выражает одно и то же ключевое понятие и в этом смысле является ключевым словом. Однако по отношению к сверхтексту ключевыми обычно оказываются лишь основные номинации цепочек (и некоторые их неоценочные синонимы), поскольку лишь они выражают соответствующее ключевое понятие во всех или многих текстах коллекции.

Ориентироваться в тематической структуре текста помогают подзаголовки, проспективные конструкции и вопросительные предложения, утвердительная часть которых называет развиваемую в дальнейшем тему.

При рассмотрении содержания текстовых фрагментов в аспекте деятельностной модели знака [Осипов и др. 2018] номинативные единицы характеризуются нами вместе с предицируемыми им признаками (ремами).

Анализ материала. Проанализируем один из текстов, хорошо иллюстрирующих рассматриваемые закономерности: «Какой купить автомобиль?»[4]. Его автор, поддерживая контакт с адресатом (инструктируемым лицом), активно использует диалогические речевые средства [Дускаева 2018], включая вопросительные предложения, которыми по ходу разговора маркируется каждый новый его предмет (тема): *Какую машину выбрать — отечественную или иномарку? Кто будет на ней ездить? Для чего мне этот автомобиль? Что я собираюсь на нем перевозить? Куда я собираюсь на нем ездить и с каким грузом? Каковы должны быть размеры вашего автомобиля? АКПП или МКПП? Какой объем двигателя выбрать? Выбрать задний или передний привод? Новая или подержанная?* и др. Некоторые вопросительные предложения используются в роли подзаголовков. Функцию выделения темы выполняют также проспективные конструкции: *Что касается систем безопасности автомобиля… Что касается конкретной марки автомобиля… По поводу «тюнинга»…*

Как уже отмечалось, автор текста определяет круг тем не произвольно, а в соответствии со сложившейся в общественном сознании моделью типичной ситуации — сценарием покупки автомобиля. Поэтому не только в рассматриваемом тексте, но и в других текстах, отражающих этот сценарий, представлен близкий состав тем: «цель покупки», «условия эксплуатации автомобиля», «цена автомобиля и сопутствующие покупке расходы», «возраст машины», «страна-производитель», «класс автомобиля», «марка автомобиля», «тип коробки передач», «характеристики силового агрегата», «вид привода», «система безопасности», «кузов», «салон» и др.

Тематические цепочки, маркирующие основную и каждую из частных тем рассматриваемого текста, дают объективное представление о его тематической организации, которая, в свою очередь, может быть соотнесена с содержанием сценария, воплощенного в тексте. При этом каждая тема-рематическая пара инструктивного текста, описывающего способы осуществления определенной деятельности, фик-

[4] http://вертикаль-оса.рф/publ/kakoj_kupit_avtomobil/1-1-0-4.

сирует уточнение более абстрактного содержания (*выбираем машину*) путем введения дополнительной информации (*выбираем марку/функционал/размеры* и т. п.), что позволяет адаптировать общий способ к конкретным условиям, в которых действует или планирует действовать человек. Поэтому с точки зрения психологии движение в тема-рематическом пространстве инструктивного текста соответствует структуре осуществления деятельности: ее общий мотив определяет содержание и последовательность ряда промежуточных действий, цели которых доопределяются в зависимости от конкретных обстоятельств реализации мотива.

Нужно иметь в виду, что любое синтаксически свободное словосочетание на глубинно-сематическом уровне — результат свертывания некоторой тема-рематической структуры. Так, предложение *Автомобиль* (предмет потребности будущего владельца, известное, тема) *выбирают* [или *страхуют, регистрируют*] (ремы) преобразуется в номинативную единицу *выбор автомобиля*, образующую в процессе развертывания текста новую тему, в которой выделяется тот или иной признак (рема). Психологическим коррелятом образования в инструктивном тексте тема-рематических цепочек выступает процесс последовательной операционализации планируемых целей, т. е. построения все более детальных схем действования с учетом все более конкретных условий. Преобразование ремы в тему отражает на речевом уровне ситуацию, когда сделанный на предыдущем этапе планирования выбор оказывается не конечным, а требующим дальнейшей детализации.

Исходная тема — *автомобиль* — маркируется проходящей через весь текст номинативной цепочкой: *автомобиль* (19 повторений), *машина* (16), *авто* (3), *легковушка*, *автомобильчик*, *машинка*, *пластмассовая игрушка*. Эта тема по разным основаниям связана родо-видовыми отношениями с вводимыми автором новыми темами, образующими свои номинативные цепочки («автомобиль отечественного производства» и «автомобиль иностранного производства», «новый автомобиль» и «подержанный автомобиль»): *отечественная машина*, *отечественное авто*, *произведение отечественного автопрома*, *произведение российского конструкторского гения* с видовыми номинативными цепочками и отдельными номинациями на более низких уровнях деления — *жигули* (2); *пятерка*, *восьмерка*, *девятка*, *десятка*; *иномарка* (3) и др. Отношениями целого и части исходная тема связана с темами, охватывающими различные агрегаты автомобиля: *АКПП, автоматическая КПП, автомат*, *МКПП, механика* (3), *механическая коробка, ручная коробка* и др.

Номинация *покупка автомобиля* предицируется признаками «цель покупки», «условия эксплуатации покупаемой машины», «цена», «опыт вождения» и др. Этим обусловлено появление в тексте рядов номинативных единиц, представленных функциональными эквивалентами: *удобство, безопасность, проходимость, статус* (*ради удобства, безопасности, проходимости; чтобы показать свой статус*), *семья, гонки* (*автомобиль для семьи, для гонок*); *трасса, поток машин; неадекватный сервис, плохое обслуживание* и др.

Анализ субтекстов, представленных ключевыми словами, позволяет анализировать текст как знак в аспекте картины мира. Так, показателями смысла (субъективной модальности) выступают а) волюнтативные высказывания — советы, рекомендации, предостережения и б) оценочные высказывания. Примеры: а) *При покупке автомобиля вы должны помнить о трех важнейших составляющих любой системы безопасности*; *Лучше взять машину с передним или задним приво-*

дом; б) *Трудно бывает смириться с бездушностью пластиковых салонов современных авто.*

Значение (опыт действования в сценарии) выражается предикатами, маркирующими последовательность рекомендуемых инструктором действий, а также детерминантами с семантикой последовательности (*сначала, затем, дальше*): *Естественно сначала ответить на вопрос: «Для чего мне этот автомобиль?» Дальше можно определяться с маркой и моделью автомобиля.*

Образ (воспроизведение свойств объекта) создается описательными высказываниями и текстовыми фрагментами: *Машина чистенькая, новенькая, все отлично работает, и нет ни единой царапинки*; *Японцы надежны, но в недорогих комплектациях зачастую страдают дешевым пластиком салона. Немцы дороги, удобны, но любят, когда за ними хорошо ухаживают и не прощают плохого обслуживания.* Разумеется, компоненты содержательной структуры знака могут совмещаться, выражаясь одними и теми же речевыми сегментами.

Результаты исследования. Описание тематических номинативных цепочек отдельного текста означает систематизацию номинативных единиц в соответствии с организацией субцельностей речевого произведения, маркерами которых данные единицы являются. Иными словами, это систематизация номинаций, отражающая на поверхностном уровне глубинное тема-рематическое структурирование речевого произведения.

В то же время система TextAppliance определяет бóльшую или меньшую информационную значимость слов и субстантивных словосочетаний для идентификации текста или множества текстов (сверхтекста). Номинативные единицы с повышенным индексом информационной значимости оцениваются как ключевые слова.

Такие номинации по их принадлежности к тем или иным субцельностям могут быть автоматически соотнесены с различными объектными областями сценария. Некоторые из этих номинативных единиц являются обозначениями данных областей: *класс автомобиля, объем двигателя, система безопасности* и др. Подобные обозначения, представляя соответствующие субцельности в свернутом виде, принадлежат к числу наиболее емких в содержательном отношении номинаций текста. Они, кроме того, обозначают основные подтемы текста, т. е. подтемы, осмысливаемые автором-инструктором как наиболее значимые для адресата.

Следовательно, появляется возможность установления группы номинаций, которые обладают всем комплексом основных характеристик ключевого слова. Во-первых, это слова и словосочетания, особенно значимые для идентификации текстов определенной тематики. Во-вторых, эти номинации в своей совокупности полно представляют предметное содержание текста. В-третьих, они выступают маркерами содержания, которое автор-инструктор считает наиболее важным и к которому он целенаправленно привлекает внимание адресата.

Каждая из указанных характеристик ключевого слова проявляется у отдельных номинативных единиц в большей или меньшей степени[5]. Например, номинация *снежные дороги* имеет один из самых высоких показателей идентификатора

[5] Не случайно исследователи, проводящие эксперименты по автоматическому аннотированию текстов, иногда намеренно не используют в задании респондентам номинацию «ключевое слово», а предлагают выбрать из текста слова и словосочетания, «которые описывают его содержание» [Ванюшкин и др. 2019: 211–212].

текста. Она относится к важной объектной области («проходимость автомобиля»), но сама не является названием одной из раскрываемых автором тем. Это же следует сказать о словосочетаниях *неопытные водители*, *полупустая машина*, *мягкая подвеска* и др. Напротив, номинация *покупка машины*, обозначая тему всего рассматриваемого текста, занимает в рейтинге его идентификаторов периферийное (98-е) место. Однако при включении в поисковый запрос в дополнение к ней других лексических единиц (например, *покупка подержанной машины*) возникает новая номинация со своими идентификационными характеристиками.

Важно отметить, что определение набора ключевых слов в текстах или субтекстах, посвященных отдельным этапам развертывания сценария (в нашем случае — выбору автомобиля, его осмотру, заключению сделки, регистрации транспортного средства и другим более частным), может способствовать улучшению качества автоматического распознавания сценариев в электронных массивах текстов. По-видимому, для выделения номинативных единиц, обладающих комплексом указанных выше характеристик ключевого слова, целесообразно с помощью системы TextAppliance анализировать тексты, из которых предварительно устранены фрагменты, реализующие периферийный предикации, т. е. предикации, которые, согласно концепции Т. М. Дридзе, не входят в число основных элементов содержания речевого произведения.

Существенно также, что извлечение ключевых слов вместе с контекстами, в которых они используются, позволяет автоматически формировать базу рекомендаций, значимых для инструктируемого лица. Примеры таких контекстов: *Какой ОБЪЕМ ДВИГАТЕЛЯ выбрать? Если вы ездите по городу, лучше взять небольшой: 1,2–1,6 литра… Если частенько приходится ездить по загородным дорогам, можно брать 1,8–2,5 литра. ДЖИПЫ ИЛИ ВНЕДОРОЖНИКИ… Подумайте, нужен ли вам расход 15–17 литров на сотню; Выбрать задний или передний ПРИВОД? С небольшим опытом вождения лучше взять машину с передним или полным приводом.*

Таким образом, предложенный подход к анализу текста, учитывающий закономерности его тема-рематического структурирования и компрессии, а также организации номинативных единиц на текстовой плоскости, может стать лингвистической базой для дальнейшего совершенствования систем автоматического извлечения ключевых слов.

Выводы. Создание когнитивных ассистентов, инструктирующих пользователя при реализации им определенной цели, требующей следования некоторому сценарию, делает актуальной задачу разработки иерархически организованной системы потенциальных запросов, представленных ключевыми словами. Сформировавшийся в общественном сознании сценарий содержит правила целесообразного поведения субъекта. Эти правила должны транслироваться ассистентом пользователю в соответствии с его информационными запросами.

Решению указанной задачи могут способствовать лингвистические исследования по теории текста, прежде всего работы, посвященные механизмам тема-рематического структурирования порождаемого речевого произведения, а также его компрессии, пределом которой является набор ключевых слов (Л. В. Сахарный, Л. Н. Мурзин), а также иерархии коммуникативных программ текста как семантико-смысловой структуры особого рода (Т. М. Дридзе), его тематической организации (Т. В. Матвеева и др.).

Выделение ключевых слов из текста с помощью автоматизированной системы TextAppliance дополнялось нами определением их места в тематической организации речевого произведения, для чего описывались тематические цепочки последнего. Рассмотрение образующих эти цепочки номинативных единиц в единстве с предицируемыми им признаками (ремами) позволяет изучать содержательную организацию текста как знака, создаваемую смыслами, значениями и образами.

Сделаны первые шаги в разработке технологии выделения из текста слов и субстантивных словосочетаний, обладающих комплексом основных характеристик ключевого слова — высоким показателем идентификатора документа, содержательной емкостью, а также коммуникативной значимостью номинативной единицы как маркера важной для адресата информации.

Литература

Ананьева, М. И., Девяткин, Д. А., Зубарев, Д. В., Осипов, Г. С., Смирнов, И. В., Соченков, И. В., Тихомиров, И. А., Швец, А. В., Шелманов, А. О. (2016). TextAppliance: поиск и анализ больших массивов текстов. В *Национальная конференция по искусственному интеллекту с международным участием. Т. 3* (с. 220–228). Смоленск: Универсум.

Ванюшкин, А. С., Гращенко, Л. А., Романишин, Г. В. (2019). Разметка коллекции текстов ключевыми словами: практические аспекты автоматизации. *Новые информационные технологии в автоматизированных системах, 22,* 210–216.

Дейк ван, Т. А. (1989). *Язык. Познание. Коммуникация.* Москва: Прогресс.

Дридзе, Т. М. (2009). *Язык и социальная психология.* Москва: URSS.

Дускаева, Л. Р. (2018). Диалогичность. В *Медиалингвистика в терминах и понятиях: словарь-справочник* (с. 32–38). Москва: Флинта.

Купина, Н. А. (2019). Сверхтекст. В *Стилистический энциклопедический словарь русского языка* (с. 374–376). Москва: Флинта.

Леонтьев, А. А. (1976). Признаки связности и цельности текста. В *Смысловое восприятие речевого сообщения (в условиях массовой коммуникации)* (с. 46–48). Москва: Наука.

Матвеева, Т. В. (1990). *Функциональный стили в аспекте текстовых категорий.* Свердловск: Изд-во Урал. ун-та.

Матвеева, Т. В. (2019). Тема текста. В *Стилистический энциклопедический словарь русского языка* (с. 252–254). Москва: Флинта.

Мбайкоджи, Э., Драль, А. А., Соченков, И. В. (2012). Метод автоматической классификации коротких текстовых сообщений. *Информационные технологии и вычислительные системы 3,* 93–102.

Мурзин, Л. Н. (1982). О деривационных механизмах текстообразования. В *Теоретические аспекты деривации* (с. 20–29). Пермь: Перм. гос. ун-т.

Мурзин, Л. Н. (1984). *Основы дериватологии*. Пермь: Перм. гос. ун-т.

Мурзин, Л. Н., Штерн, А. С. (1991). *Текст и его восприятие.* Свердловск: Издательство Уральского университета.

Осипов, Г. С., Чудова, Н. В., Панов, А. И., Кузнецова, Ю. М. (2018). *Знаковая картина мира субъекта поведения.* Москва: Физматлит.

Салимовский, В. А., Осипов, Г. С., Кузнецова, Ю. М., Суворова, М. И., Чудова, Н. В. (2019). Лингвистические аспекты целеполагания в когнитивном моделировании (на материале речевого жанра «план-инструкция»). *Искусственный интеллект и принятие решений, 4,* 10–22.

Сахарный, Л. В. (1982). Актуальное членение и компрессия текста (к использованию методов информатики в психолингвистике). В *Теоретические аспекты деривации* (с. 29–38). Пермь: Перм. гос. ун-т.

Сахарный, Л. В. (1992). Тексты-примитивы и закономерности их порождения. В *Человеческий фактор в языке: Язык и порождение речи* (с. 221–236). Москва: Наука.

Сахарный, Л. В. (1994). Человек и текст: две грамматики. В *Человек. Текст. Культура* (с. 7–59). Екатеринбург: Институт развития регион. обр.

Смирнов, И. В., Панов, А. И., Скрынник, А. А., Чистова, Е. В. (2019). Персональный когнитивный ассистент: концепция и принципы работы. *Информатика и ее применения*, *13* (3), 105–113.

Beliga, S. (2015). Keyword extraction: a review of methods and approaches. *Journal of Information and Organizational Sciences. 39* (1), 1–20.
Sterckx, L., Demeester, T., Deleu, J., Develder, C. (2019). Creation and evaluation of large keyphrase extraction collections with multiple opinions. *Language Resources and Evaluation, 52* (2), 503–532.

Статья поступила в редакцию 2 июля 2020 г.;
рекомендована в печать 4 ноября 2020 г.

Контактная информация:

Каджая Людмила Алексеевна — канд. филол. наук; kadzhaya1@icloud.com
Кузнецова Юлия Михайловна — канд. психол. наук; kuzjum@yandex.ru
Салимовский Владимир Александрович — д-р филол. наук; salimovsky@rambler.ru
Суворова Маргарита Игоревна — главный специалист; suvorova@isa.ru

Thematic organization of instructional texts as a linguistic basis for the acquisition of knowledge by an intelligent agent*

L. A. Kadzhaya[1,2], *Iu. M. Kuznetsova*[3], *V. A. Salimovskii*[1], *M. I. Suvorova*[3]

[1] Perm State University,
15, ul. Bukireva, Perm, 614068, Russian Federation
[2] Shandong University,
180, Wenhua Xilu, Weihai, 264209, China
[3] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences,
9, pr. 60-letiia Oktiabria, Moscow, 117321, Russian Federation

For citation: Kadzhaya, L. A., Kuznetsova, Iu. M., Salimovskii, V. A., Suvorova, M. I. (2021). Thematic organization of instructional texts as a linguistic basis for the acquisition of knowledge by an intelligent agent. *Media Linguistics, 8* (1), 45–56. https://doi.org/10.21638/spbu22.2021.104 (In Russian)

The article examines the thematic organization of instructional texts in the aspect of problems relevant to work on the creation of a cognitive assistant. The purpose of the assistant is to provide a user with the necessary information to follow the rules of a particular scenario to successfully achieve a goal according to the search query. The query containing certain keywords, further specified as the task being solved, is focused on a detailed set of topics which mark the subject areas reflected in the scenario. The authors of the article provide a review of some linguistic works devoted to the issues of theme-rhematic structuring of a produced text and its compression within the limits of keywords. The importance of the description of the text's thematic chains, to obtain the detailed objective information on its thematic structure, is emphasized. When comparing the list of keywords identified by the automatic system TextAppliance in a collection of Internet-extracted instructional texts retrieved from the Internet with the results of hand-held analysis of these texts, to determine the place of various nominative units in the text's thematic organization, the authors consider the most significant characteristics of a keyword shown in different nominative units to varying degrees. This is a high indicator of a text identifier, content capacity, and communicative significance of a word or a substantive phrase as a marker of important information for a recipient. Defining keywords in whole instructional texts and in relatively independent text fragments (subtexts) that describe individual stages of the user's goal achievement (for example, the stages of selecting a car, its inspection, making a transaction, car registration) makes it possible to improve the quality of scenario identification in the Network. Extracting keywords along with their context allows for the creation of a recommendations'

* The research was supported by Russian Foundation for Basic Research, project no. 18-00-00606 (18-00-00233).

database for users automatically. The significance of the theme-rhematic text structure analysis, as a sign for its modeling in the sign picture of the world, is revealed.

Keywords: intelligent agent, text, theme-rhematic text organization, keyword, thematic chain.

References

Anan'eva, M. I., Deviatkin, D. A., Zubarev, D. V., Osipov, G. S., Smirnov, I. V., Sochenkov, I. V., Tikhomirov, I. A., Shvets, A. V., Shelmanov, A. O. (2016). TextAppliance: search and analysis of large volumes of texts. In *National conference on artificial intelligence with international participation. Vol. 3* (pp. 220–228). Smolensk, Universum Publ. (In Russian)

Beliga, S. (2015). Keyword extraction: a review of methods and approaches. *Journal of Information and Organizational Sciences, 39* (1), 1–20.

Deik van, T. A. (1989). *Language. Knowledge. Communication*. Moscow, Progress Publ. (In Russian)

Dridze, T. M. (2009). *Language and social psychology.* Moscow, URSS Publ. (In Russian)

Duskaeva, L. R. (2018). Dialogism. In *Metalinguistics in terms and concepts: a dictionary reference* (pp. 32–38). Moscow, Flinta Publ. (In Russian)

Kupina, N. A. (2019). Hypertext. In *Stylistic encyclopedia of the Russian language* (pp. 374–376). Moscow, Flinta Publ. (In Russian)

Leont'ev, A. A. (1976). Features of coherence and integrity of a text. In *Semantic perception of a speech message (in terms of mass communication)* (pp. 46–48). Moscow, Nauka Publ. (In Russian)

Matveeva, T. V. (1990). *Functional styles in the aspect of text categories.* Sverdlovsk, Ural University Publ. (In Russian)

Matveeva, T. V. (2019). Text subject. In *Stylistic encyclopedia of the Russian language* (pp. 252–254). Moscow, Flinta Publ. (In Russian)

Mbaikodzhi, E., Dral', A. A., Sochenkov, I. V. (2012). Method for automatic classification of short text messages. *Informatsionnye tekhnologii i vychislitel'nye sistemy 3,* 93–102. (In Russian)

Murzin, L. N. (1982). On derivation mechanisms of text formation. In *Theoretical aspects of derivation* (pp. 20–29). Perm', Perm State University Publ. (In Russian)

Murzin, L. N. (1984). *Fundamentals of derivatology.* Perm', Perm State University Publ. (In Russian)

Murzin, L. N., Shtern, A. S. (1991). *Text and its perception.* Sverdlovsk, Ural University Publ. (In Russian)

Osipov, G. S., Chudova, N. V., Panov, A. I., Kuznetsova, Iu. M. (2018). *Symbolic world picture of a behavior subject.* Moscow, Fizmatlit Publ. (In Russian)

Sakharnyi, L. V. (1982). Actual text division and text compression (on the use of computer science methods in psycholinguistics). In *Theoretical aspects of derivation* (pp. 29–38). Perm', Perm State University Publ. (In Russian)

Sakharnyi, L. V. (1992). Texts-primitives and laws of their generation. In *Human factor in language: Language and speech generation* (pp. 221–236). Moscow, Nauka Publ. (In Russian)

Sakharnyi, L. V. (1994). Individual and Text: two grammars. In *Individual. Text. Culture* (pp. 7–59). Ekaterinburg, Institute of Regional Education Development Publ. (In Russian)

Salimovskii, V. A., Osipov, G. S., Kuznetsova, Iu. M., Suvorova, M. I., Chudova, N. V. (2019). Linguistic aspects of goal setting in cognitive modeling (on the material of the speech genre "plan-instruction"). *Iskusstvennyi intellekt i priniatie reshenii, 4,* 10–22. (In Russian)

Smirnov, I. V., Panov, A. I., Skrynnik, A. A., Chistova, E. V. (2019). Personal cognitive assistant: concept and principles of work. *Informatika i ee primeneniia, 13* (3), 105–113. (In Russian)

Sterckx, L., Demeester, T., Deleu, J., Develder, C. (2019). Creation and evaluation of large keyphrase extraction collections with multiple opinions. *Language Resources and Evaluation, 52* (2), 503–532.

Vaniushkin, A. S., Grashchenko, L. A., Romanishin, G. V. (2019). Marking up a collection of texts with keywords: practical aspects of automation. *Novye informatsionnye tekhnologii v avtomatizirovannykh sistemakh, 22,* 210–216. (In Russian)

Received: July 2, 2020
Accepted: November 4, 2020

Authors' information:

Liudmila A. Kadzhaya — PhD in Philology; kadzhaya1@icloud.com
Iuliia M. Kuznetsova — PhD on Psychology; kuzjum@yandex.ru
Vladimir A. Salimovskii — Dr. Sci. in Philology; salimovsky@rambler.ru
Margarita I. Suvorova — Chief Specialist; suvorova@isa.ru