

RU

Исследование синонимических отношений в терминологии компьютерной лингвистики

Польщикова О. Н.

Аннотация. Цель исследования - раскрыть характерные черты синонимических отношений в терминологии компьютерной лингвистики на основе проведения их комплексного анализа. В статье выявляются причины наличия терминов-синонимов профессиональной сферы компьютерной лингвистики, выполняется их исследование с точки зрения этимологии и морфологической природы. Научная новизна исследования заключается в составлении классификации синонимических отношений в анализируемой терминологии по наиболее значимым признакам. Полученные результаты показали использование в терминологии компьютерной лингвистики взаимозаменяемых словосочетаний, их усечённо-словесных форм, аббревиатур и синтаксических вариантов.

EN

Research of Synonymous Relations in the Terminology of Computational Linguistics

Polshchykova O. N.

Abstract. The aim of the research is to reveal the characteristic features of synonymous relations in the terminology of computational linguistics on the basis of their complex analysis. The article identifies the reasons for the presence of synonymous terms in the professional sphere of computational linguistics, and conducts their study from the point of view of etymology and morphological nature. The scientific originality of the research lies in the compilation of a classification of synonymous relations in the analysed terminology according to the most significant features. The results obtained have shown the use of interchangeable word combinations, their reduced verbal forms, abbreviations and syntactic variants in the terminology of computational linguistics.

Введение

Компьютерная лингвистика – сравнительно новая профессиональная междисциплинарная сфера, теоретические и прикладные наработки которой формируются на стыке языкознания, математики, вычислительных методов и технологий обработки данных на основе искусственного интеллекта (Большакова, Воронцов, Ефремова и др., 2017; Dowell, Nixon, Graesser, 2019; Ive, Viani, Kam, 2020). Сфера компьютерной лингвистики ориентирована на решение различных прикладных задач, связанных с автоматической обработкой естественного языка (Агузумцян, Великанова, Польщикова, 2021; Польщикова, Лазарев, Константинов, 2020; Кормалев, Куршев, Сулейманова и др., 2004; Польщикова, Польщикова, Игитян и др., 2019; Arts, Hou, Gomez, 2021; Dehouche, 2021; Ive, Viani, Kam, 2020; Polshchykov, Lazarev, Konstantinov, 2020; Polshchykov, Lazarev, Polshchykova et al., 2019; Qiu, Sun, Xu et al., 2020; Savin, Drews, Maestre-Andrés, 2020).

В терминологии компьютерной лингвистики, так же как и в других различных отраслевых терминологиях, распространена синонимия, что вызывает «определенные трудности в сфере профессиональной коммуникации» (Таранова, Бубырева, Таранов, 2016, с. 55). Термины-синонимы компьютерной лингвистики используются в устно-профессиональной речи, в учебных пособиях и учебниках, научных статьях, диссертациях и другой научно-исследовательской литературе. Учету синонимии в различных лингвистических ресурсах уделяется особое внимание. В лексической онтологии WordNet слова каждой части речи сгруппированы в синсеты (синонимические ряды), которые являются базовыми единицами словаря. Наличие синонимии вызывает «сложности выявления различных вхождений терминов в текст» (Большакова, Воронцов, Ефремова и др., 2017, с. 120). Это связано с тем, что «термины при употреблении достаточно часто видоизменяются – усекаются, сокращаются, заменяются синонимами, соединяются и т.п.: коммуникативная многозначность

запроса – коммуникативная многозначность, синтаксическое представление – СинП, вложенный файл – вложение. Подобные текстовые варианты представляют собой различные формы выражения одного и того же понятия и по возможности должны быть распознаны» (Большакова, Воронцов, Ефремова и др., 2017, с. 120).

Изучение классификации синонимических терминов и причин их появления ориентировано на решение остро востребованных в практической сфере важнейших задач, связанных с унификацией терминосистем, их упорядочением (Таранова, Бубырева, Таранов, 2016), что определяет актуальность темы изложенного в статье исследования.

Для достижения цели исследования в работе поставлены и решены следующие задачи: 1) выявить причины наличия синонимических отношений в терминологии компьютерной лингвистики; 2) сгруппировать термины-синонимы по различным классификационным признакам; 3) исследовать их с точки зрения этимологии и морфологической природы; 4) проанализировать формы вариантности и взаимозаменяемости терминов рассматриваемой профессиональной сферы.

Для анализа синонимических отношений в терминологии компьютерной лингвистики применены различные методы исследования. Систематизация рассмотренных терминов выполнена на основе использования описательного метода исследования. Происхождение некоторых терминов выявлено с помощью метода этимологического анализа. Для прослеживания семантических особенностей специальных названий послужил дефинитивный анализ фактического материала. В качестве вспомогательного использован метод количественного анализа.

Теоретической базой послужили публикации Г. Г. Бабаловой и С. Н. Широбокова (2018), Е. И. Большаковой, К. В. Воронцова и Н. Э. Ефремовой (2017), Е. Н. Тарановой, Ж. А. Бубыревой и А. О. Таранова (2016), D. Blei и J. Lafferty (2007; 2012), N. Dehouche (2021), R. Mitkov (2003) и других отечественных и зарубежных ученых. Источниками фактического материала послужили тексты по тематике компьютерной лингвистики, размещенные в справочниках, научных периодических журналах и сборниках, в переводных и толковых словарях, в частности в Русско-английском тезаурусе по компьютерной лингвистике (РАТКЛ).

Практическая значимость исследования состоит в том, что представленные в статье материалы могут быть использованы для преподавания терминоведческих дисциплин, лексикологии, учебных курсов по профилю компьютерной лингвистики.

Основная часть

Причины появления и использования синонимов в профессиональной лексике исследованы во многих работах (Бабалова, Широбоков, 2018; Вакулик, Сичкар, 2020; Дасовхаджиева, 2020; Таранова, Бубырева, Таранов, 2016; Pllana, Pllana, Pllana et al., 2020). В терминологии компьютерной лингвистики синонимия, по нашему мнению, обусловлена следующими факторами:

- 1) разнообразие способов формирования структуры терминов;
- 2) целесообразность подбора русскоязычных соответствий к терминам иноязычного происхождения;
- 3) наличие у определенного объекта номинации различных свойств и признаков, которые могут стать основой для названия;
- 4) появление новых понятий или свойств объектов в процессе постоянного развития наук и технологий, что заставляет искать соответствующие точные номинации и самые удачные названия;
- 5) имеющиеся общезыковые синонимы, из которых можно подобрать термины;
- 6) возможность параллельного употребления термина и его дефиниции (дефинитивная синонимия);
- 7) стремление избегать повторов одного и того же слова или словосочетания;
- 8) именованья одного и того же понятия различными исследовательскими школами, учеными;
- 9) необходимость языковой экономии, способствующая употреблению однословных сложносокращённых названий и аббревиатур.

По типу синонимического отношения различают абсолютные, относительные и комплексные синонимы. Первый тип составляют семантически тождественные термины, которые называются дублетами: *сегментация (токенизация)* – разбиение текста на сегменты «с более узким информационным содержанием» (Хобсон, Ханнес, Коул, 2020, с. 71); «процесс разбиения текста на лингвистически значимые единицы, т.е. слова (словоформы), знаки препинания, числа и буквенно-числовые выражения» (РАТКЛ, 2021); *стемматизация (нормализация, лемматизация)* – «приведение каждого слова в документе к его нормальной форме» (Большакова, Воронцов, Ефремова и др., 2017, с. 201); «группирование различных форм слова по кластерам» (Хобсон, Ханнес, Коул, 2020, с. 70); «процесс группировки различных флективных форм одного слова таким образом, чтобы при анализе они обрабатывались как одно слово» (РАТКЛ, 2021), *сирконстант (адъонкт)* – «единица, которая заполняет активную синтаксическую валентность, не соответствующую никакой семантической валентности» (Тестелец, 2001, с. 166); *слово (токен)* – «основная структурно-семантическая единица языка, служащая для именованья предметов и их свойств, явлений, отношений действительности, обладающая совокупностью семантических, фонетических и грамматических признаков» (РАТКЛ, 2021); «единица языка, служащая для называния отдельного понятия» (Кузнецов, 2000, с. 1210); «подстрока в тексте, находящаяся между знаками препинания» (Большакова, Воронцов, Ефремова и др., 2017, с. 32); *лингвистический (языковой) корпус* – «некоторый филологически-компетентный массив языковых данных, множество текстов, отобранных

в соответствии с некоторой исследовательской задачей и специально подготовленных, размеченных, структурированных, представленных в унифицированном виде» (РАТКЛ, 2021); «представительный массив текстов, собранный по определённому принципу (по жанру, авторской принадлежности и т. п.) и обладающий лингвистической разметкой (морфологической, акцентной, синтаксической, дискурсивной)» (Большакова, Воронцов, Ефремова и др., 2017, с. 23); «примерная совокупность высказываний, отобранных для анализа и представленных в виде письменного текста, магнитофонной записи и т.п.» (Ахманова, 2004, с. 209); *шаблон (образец)* – «описание языковой конструкции, которая употребляется для отражения в тексте искомого факта или объекта и описывает различные атрибуты текста: морфологические признаки слов, их синтаксическое значение и соотношение, принадлежность отдельному фрагменту предложения; порядок слов, расстояние между словами, наличие ключевых слов, характеризующих ситуацию» (РАТКЛ, 2021); «формальное описание (образец) языковой конструкции, которую необходимо найти в тексте, чтобы извлечь нужную информацию» (Большакова, Воронцов, Ефремова и др., 2017, с. 99).

К относительным синонимам принадлежат семантически близкие термины, примерами которых являются: *компьютерная лингвистика* – «направление в прикладной лингвистике, ориентированное на использование компьютерных инструментов – программ, компьютерных технологий организации и обработки данных – для моделирования функционирования языка в тех или иных условиях, ситуациях, проблемных сферах и т.д., а также вся сфера применения компьютерных моделей языка в лингвистике и смежных дисциплинах» (РАТКЛ, 2021) и *автоматическая обработка естественного языка* – имеет более узкое значение: «направление исследований, занимающееся... моделированием на компьютере процессов анализа и генерации текстов на естественном языке, распознаванием и синтезом звучащей речи, а также решением прикладных задач, связанных с преобразованием текста и звучащей речи, например, разрешением многозначности слов в тексте, машинным переводом, автоматическим реферированием, определением диктора по голосу (идентификация диктора) и многих других» (РАТКЛ, 2021); *разрешение неоднозначности* – «удаление из представления лингвистического объекта (текста) конкурирующих вариантов с сохранением непротиворечивой интерпретации» (РАТКЛ, 2021) и *снятие омонимии* – имеет уточненную дефиницию: «этап анализа текста, на котором проводится выбор единственного варианта морфологического анализа для каждого токена» (Большакова, Воронцов, Ефремова и др., 2017, с. 42).

Комплексные синонимы объединяют термины первых двух типов, например, словосочетания *извлечение мнений* и *анализ тональности*, имеющие дефиницию «выявление эмоционально окрашенной лексики и эмоциональной оценки автора по отношению к объектам, речь о которых идет в тексте» (РАТКЛ, 2021), являются абсолютными синонимами, а термин *автоматическая обработка текста* («преобразование текста на искусственном или естественном языке с помощью ЭВМ» (РАТКЛ, 2021)) является относительным синонимом первых двух.

Термины-синонимы различаются по структуре. В терминологии компьютерной лингвистики обнаружено параллельное использование следующих терминов:

1) однословных терминов: *фонация* – *голособразование*; *теггирование* – *разметка*; *репрезентативность* – *сбалансированность*; *многозначность* – *полисемия*;

2) односложного термина и словосочетания: *ресинтез* – *вокодерный синтез*; *тема* – *логический субъект*; *тон* – *мелодика речи*; *форманта* – *полюс спектра*; *цель* – *конечная точка*; *чтение* – *репродуцированная речь*; *тезаурус* – *семантический словарь*; *парсер* – *синтаксический анализатор*; *оцифровка* – *цифровое кодирование*; *коллокация* – *устойчивое словосочетание*; *кластеризация* – *кластерный анализ*;

3) терминологических сочетаний: *извлечение информации* – *выделение концептов*; *группа числительного* – *количественная конструкция*; *референциальный анализ* – *разрешение анафоры*; *система управления терминологией* – *терминологический менеджер*.

Взаимозаменяемыми в контексте компьютерной лингвистики, кроме синонимов, являются различные формы выражения одних и тех же терминологических словосочетаний:

1) терминологических словосочетаний и их усечённо-словесных форм: *автоматическое реферирование* – *автореферирование*; *лингвистическая разметка* – *лингворазметка*; *восприятие речи* – *речевосприятие*; *корпусный менеджер* – *корпус-менеджер*; *перевод в режиме онлайн* – *онлайн-перевод*;

2) терминологического словосочетания и соответствующей аббревиатуры: *компьютерная лингвистика* – *КЛ*; *машинный перевод* – *МП*; *лингвистический процессор* – *ЛП*; *автоматическое распознавание речи* – *АРР*; *информационно-поисковая система* – *ИПС*; *искусственная нейронная сеть* – *ИНС*; *терминологический банк данных* – *ТБД*.

По морфологической природе наиболее распространенными в исследуемой лексике являются синонимы-существительные (75,2%), например, *анализ* – *разбор*; *артикуляция* – *произнесение*; *определение* – *толкование*. Реже встречаются прилагательные-синонимы (18,9%), например, *речевой* – *голосовой*; *сконструированный* – *искусственный*; *автоматический* – *машинный*. Выявлено незначительное количество глагольных синонимов (5,9%), например, *тестировать* – *испытывать*; *корректировать* – *исправлять*.

К синонимическому ряду в лексике компьютерной лингвистики может принадлежать разное количество компонентов:

1) два компонента: *многоязычный* – *мультиязыковой*; *именной* – *субстантивный*; *просодическая* – *суперсегментная*;

2) три компонента: *толкование* – *дефиниция* – *определение*; *онтологический* – *энциклопедический* – *внелингвистический*; *тональный* – *интонационный* – *мелодический*;

3) четыре компонента: *рамка – структура – схема – формат*;

4) пять компонентов: *распознавание – выделение – извлечение – обнаружение – выявление; накопитель – память – архив – банк – база*.

Термины-синонимы компьютерной лингвистики целесообразно различать и с точки зрения этимологии. Выявлено параллельное употребление терминов иноязычного происхождения: *актант – аргумент; синтаксический анализ – парсинг; морфологический анализ – морфологическая интерпретация*. Довольно часто можно наблюдать синонимическую пару, которую составляют термин иноязычного происхождения и автохтонный термин: *акцент – ударение; интерлингва – язык-посредник; вокабула – заголовок; информация – сведения; факультативная – необязательная; пассивная – внешняя*. К тому же, в исследуемой профессиональной лексике параллельно употребляются русскоязычные термины: *словарный вход – заглавное слово; прямой – пословный; поле – ряд; смысл – значение*.

Кроме синонимии, многим отраслевым терминологиям свойственна вариантность, то есть формальное видоизменение одного и того же словосочетания без нарушения тождества значений. В терминологии компьютерной лингвистики выявлено использование синтаксических вариантов – словосочетаний с различными возможными видами грамматической связи: *корпус речи – речевой корпус; корпус текстов – текстовый корпус; интерфейс пользователя – пользовательский интерфейс; мера ассоциации – ассоциативная мера; ресурсы лингвистики – лингвистические ресурсы; фрейм целей – целевой фрейм*.

Заключение

Представленное исследование позволило выявить причины наличия синонимических отношений в терминологии компьютерной лингвистики, которые, в первую очередь, связаны с разнообразием способов формирования структуры терминов, необходимостью подбора русскоязычных соответствий к терминам иноязычного происхождения, интенсивным появлением новых понятий вследствие стремительного развития профессиональной сферы автоматической обработки естественного языка. Кроме того, терминологическая синонимия обусловлена образованием специальных названий из имеющихся языковых синонимов, стремлением избегать повторов и экономить языковые ресурсы, а также параллельным функционированием терминологических баз различных научных школ.

Термины-синонимы компьютерной лингвистики классифицируются по типу синонимического отношения (абсолютные, относительные и комплексные), структуре (однословные термины, однословные термины и словосочетания, терминологические сочетания), морфологической природе (существительные, прилагательные, глагольные синонимы), числу компонентов синонимического ряда (двух-, трех-, четырех-, пятикомпонентные), этимологическому признаку (термины-синонимы иноязычного происхождения, термин иноязычного происхождения и синонимический автохтонный термин, русскоязычные термины-синонимы).

В терминологии компьютерной лингвистики выявлены взаимозаменяемые словосочетания, их усечённо-словесные формы и аббревиатуры, а также синтаксические варианты в виде словосочетаний с различными видами грамматической связи.

Перспективы дальнейшего исследования мы видим в проведении глубокого научного анализа терминологии компьютерной лингвистики, в частности в выявлении источников и установлении закономерностей формирования терминов этой специальной сферы, изучении их системной организации, способов создания и особенностей функционирования.

Источники | References

1. Агузумян Р. В., Великанова А. С., Польщиков К. А. О применении интеллектуальных технологий обработки естественного языка и средств виртуальной реальности для поддержки принятия решений при подборе исполнителей проектов // Экономика. Информатика. 2021. Т. 48. № 2.
2. Ахманова О. С. Словарь лингвистических терминов. М.: Едиториал УРСС, 2004.
3. Бабалова Г. Г., Ширококов С. Н. Синонимия компьютерной терминологии // Наука о человеке: гуманитарные исследования. 2018. № 4.
4. Большакова Е. И., Воронцов К. В., Ефремова Н. Э. и др. Автоматическая обработка текстов на естественном языке и анализ данных. М.: Изд-во НИУ ВШЭ, 2017.
5. Вакулик И. И., Сичкарь И. Ю. Медицинская терминология и эквивалентная синонимия: особенности образования // Гуманитарные научные исследования. 2020. № 11.
6. Дасовхаджиева А. А. К вопросу о синонимии в терминологии спорта // Lingua-Universum. 2020. № 3.
7. Кормалев Д. А., Куршев Е. П., Сулейманова Е. А. и др. Архитектура инструментальных средств систем извлечения информации из текстов // Программные системы: теория и приложения. 2004. Т. 2.
8. Кузнецов С. А. Большой толковый словарь русского языка. СПб.: Норинт, 2000.
9. Польщиков К. А., Лазарев С. А., Константинов И. С. Модель для оценки эффективности выполнения роботехнической системой коммуникативных функций // СТИН. 2020. № 6.

10. Польщиков К. А., Польщикова О. Н., Игитян Е. В. и др. Алгоритм поддержки принятия решений по выбору средств обработки больших массивов естественно-языковых данных // Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. 2019. Т. 46. № 3.
11. Русско-английский тезаурус по компьютерной лингвистике (ПАТКЛ). 2021. URL: <https://uniserv.iis.nsk.su/thes/search.php>
12. Таранова Е. Н., Бубырева Ж. А., Таранов А. О. Проблема синонимии в специальной терминологии // Вестник Томского государственного педагогического университета. 2016. № 2.
13. Тестелец Я. Г. Введение в общий синтаксис. М.: РГГУ, 2001.
14. Хобсон Л., Ханнес Х., Коул Х. Обработка естественного языка в действии. СПб.: Питер, 2020.
15. Arts S., Hou J., Gomez J. C. Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures // Research Policy. 2021. Vol. 50.
16. Blei D. M. Probabilistic topic models // Communications of the ACM. 2012. Vol. 55. № 4.
17. Blei D., Lafferty J. A correlated topic model of science // Annals of Applied Statistics. 2007. Vol. 1.
18. Dehouche N. Plagiarism in the age of massive Generative Pre-trained Transformers (GPT-3) // Ethics in Science and Environmental Politics. 2021. Vol. 21.
19. Dowell N. M. M., Nixon T. M., Graesser A. Group communication analysis: A computational linguistics approach for detecting sociocognitive roles in multiparty interactions // Behavior Research Methods. 2019. Vol. 51.
20. Ive J., Viani N., Kam J. Generation and evaluation of artificial mental health records for Natural Language Processing // NPJ Digital Medicine. 2020. Vol. 3.
21. Mitkov R. The Oxford handbook of computational linguistics. N. Y.: Oxford University Press, 2003.
22. Pllana S., Pllana G., Pllana E., Pllana Z. Synonymy and terminological doublet in economic terminology // Russian Linguistic Bulletin. 2020. № 3.
23. Polshchikov K. A., Lazarev S. A., Konstantinov I. S. Assessing the Efficiency of Robot Communication // Russian Engineering Research. 2020. Vol. 40. No. 11.
24. Polshchikov K., Lazarev S., Polshchikova O. et al. The Algorithm for Decision-Making Supporting on the Selection of Processing Means for Big Arrays of Natural Language Data // Lobachevskii Journal of Mathematics. 2019. Vol. 40. No. 11.
25. Qiu X. P., Sun T. X., Xu Y. G. et al. Pre-trained models for natural language processing: A survey // Science China. Technological Sciences. 2020. Vol. 63.
26. Savin I., Drews S., Maestre-Andrés S. Public views on carbon taxation and its fairness: A computational-linguistics analysis // Climatic Change. 2020. Vol. 162.

Информация об авторах | Author information



Польщикова Ольга Николаевна¹, к. филол. н.

¹ Белгородский государственный национальный исследовательский университет



Polshchikova Olga Nikolaevna¹, PhD

¹ Belgorod National Research University

¹ polshchikova@bsu.edu.ru

Информация о статье | About this article

Дата поступления рукописи (received): 08.10.2021; опубликовано (published): 31.01.2022.

Ключевые слова (keywords): компьютерная лингвистика; автоматическая обработка естественного языка; синонимы; классификация; вариантность; computational linguistics; automatic natural language processing; synonyms; classification; variance.