УДК 81’42
DOI 10.25205/1818-7935-2021-19-3-45-56

Структурные модели терминологических словосочетаний для разметки корпуса научно-технических текстов

Ю. И. Бутенко [1, 2], **Н. С. Николаева** [1], **Т. Д. Маргарян** [1]

[1] *Московский государственный технический университет им. Н. Э. Баумана*
Москва, Россия

[2] *Российский университет дружбы народов*
Москва, Россия

Аннотация

В статье представлены структурные модели терминологических словосочетаний из предметной области «Сварка» как основа для создания автоматизированных средств разметки корпусов научно-технических текстов. Обозначено место корпусов научно-технических текстов в корпусной лингвистике и перспективы дальнейших научных исследований на их основе. Актуальность исследования обусловлена необходимостью создания корпусов научно-технических текстов в целом и средств автоматической разметки терминов в частности. Обосновано, что в настоящее время основной проблемой при создании корпусов научно-технических текстов является автоматическая разметка терминологических словосочетаний. Проведен анализ современного состояния терминосистемы предметной области «Сварка». Представлены результаты анализа двух-, трех-, четырех- и пятикомпонентных терминологических словосочетаний предметной области «Сварка», а также созданы и проиллюстрированы примерами их структурные модели. Обоснована необходимость исчисления всех возможных структурных моделей терминологических сочетаний. Усложнение структуры терминологического словосочетания чаще всего связано с усложнением структуры постпозитивного определения в зависимости от выражаемых им видовых особенностей. Новизна исследования видится в обеспечении теоретического базиса для формирования базы данных структурных моделей терминологических словосочетаний как основы надкорпусной базы данных о структуре многокомпонентных терминов для повышения качества автоматической разметки корпусов научно-технических текстов; также предложен подход к автоматической разметке многокомпонентных терминов на основе структурных моделей терминологических словосочетаний. Результат полезен также для обработки терминов-кандидатов при проведении корпусных исследований при последующем использовании корпусов научно-технических текстов.

Ключевые слова

термин, терминологическое словосочетание, структурная модель, разметка, корпус научно-технических текстов

Для цитирования

Бутенко Ю. И., *Николаева Н. С.*, *Маргарян Т. Д.* Структурные модели терминологических словосочетаний для разметки корпуса научно-технических текстов // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2021. Т. 19, № 3. С. 45–56. DOI 10.25205/1818-7935-2021-19-3-45-56

Structural Models of Terminological Word Combinations for Marking up a Corpus of Scientific and Technical Texts

Iuliia I. Butenko [1, 2], **Natalia S. Nikolaeva** [1], **Tatiana D. Margaryan** [1]

[1] *Bauman Moscow State Technical University*
Moscow, Russian Federation

[2] *Peoples Friendship University of Russia*
Moscow, Russian Federation

Abstract

The article presents structural models of terminological phrases from the subject area "Welding" as the basis for creating automated tools to mark up the corpus of scientific and technical texts. The place of scientific and technical corpora in corpus linguistics and the prospects for their further research are outlined. The relevance of the research stems from the need to create corpora of scientific and technical texts in general and to provide tools for automatic detection of terms in particular. It is substantiated that the main problem in designing such corpora is the automatic markup of terminological phrases. The analysis of the current state of the term system of the subject area "Welding" has been carried out. The results of the analysis of two-, three-, four- and five-component terminological phrases of "Welding" and their structural models are presented and illustrated by examples. The necessity of listing all possible structural models of terminological combinations has been substantiated too. It has been established that the addition of a new component to the basic terminological combination most often occurs with introduction of one more postpositional attribute whose function is to add some specific feature to the basic meaning. The novelty of the study is seen in providing a theoretical approach for the formation of a database of structural models of terminological phrases which may be used as a core of a supersource database on the structure of the multicomponent scientific and technical terms. An approach to automatic markup of multicomponent terms is proposed too. It will be also helpful in future corpus research for identification of candidate word combinations as scientific and technical terms.

Keywords

term, terminological phrase, structural model, markup, corpus of scientific and technical texts

For citation

Butenko, Iuliia I., Nikolaeva, Natalia S., Margaryan, Tatiana D. Structural Models of Terminological Word Combinations for Marking up a Corpus of Scientific and Technical Texts. *Vestnik NSU. Series: Linguistics and Intercultural Communication*, 2021, vol. 19, no. 3, p. 45–56. (in Russ.) DOI 10.25205/1818-7935-2021-19-3-45-56

Введение

При проведении лингвистических исследований ручной сбор иллюстративного материала традиционно является трудоемким и дорогостоящим этапом работы. Использование электронных ресурсов, в частности электронных корпусов текстов, позволяет существенно сократить время на сбор информации или создание выборок для проведения исследования.

Сферы применения корпусов текстов включают автоматизированное извлечение информации, обучение на основе данных, текстовые поиски в крупномасштабных коллекциях текстов с использованием методов обработки естественного языка, автоматическую классификацию текстов, преподавание языков для специальных целей, проведение лингвистических исследований [Нагель, 2008. С. 53].

Корпус всегда проектируется для конкретной цели. На сегодняшний день существует множество корпусов русского языка, подробный обзор которых проведен в [Захаров, 2015. С. 20–65; Захаров, Богданова, 2020]. Распространенными являются корпуса, содержащие тексты художественной литературы, однако наличия корпусов, основанных исключительно на литературно-художественных произведениях, недостаточно для эффективного применения корпусной лингвистики в области лингвистических исследований.

В настоящее время в МГТУ им. Н. Э. Баумана ведутся исследования по разработке подходов, методов, алгоритмов и программных средств создания двунаправленного параллельного корпуса научно-технических текстов. Обработка текстов в параллельном корпусе осуществляется в два этапа: сначала размечают тексты на языке оригинала и на языке перевода каж-

дый отдельно, а затем устанавливают переводные соответствия между элементами этих текстов [Butenko, Garazha, 2021].

Тексты корпусов обычно размечаются для удобства пользования, т. е. текстам и содержащимся в них языковым единицам приписываются специальные метки. Размеченные корпуса обеспечивают специализированными поисковыми системами, реализующими грамматические и лексические виды поиска [Кружков, 2015. С. 141]. В зависимости от целей создания корпуса в него включают дополнительные виды разметки [Лесников, 2019. С. 28]. Для корпуса научно-технических текстов наибольшую значимость приобретает терминологическая разметка, так как именно термины выступают основным средством передачи информации.

Работа с корпусами научно-технических текстов требует особого инструментария для выявления устойчивых терминологических сочетаний. Среди наиболее распространенных методов выявления многокомпонентных терминов в текстах используют метод выявления устойчивых сочетаний, а также статистический и гибридный методы. В основе метода автоматического выявления устойчивых сочетаний лежит использование грамматики лексико-синтаксических шаблонов, представляющих собой структурные модели лингвистических конструкций. Статистический подход заключается в нахождении *n*-грамм по заданным частотным характеристикам. Гибридный подход для выделения терминологических сочетаний, объединяющий лингвистический и статистический методы, заключается в предварительном описании моделей, по которым могут быть построены термины, для последующего поиска их в корпусе. Внутри множеств однотипных синтаксических конструкций выполняется ранжирование в соответствии с той или иной статистической мерой [Захаров, Хохлова, 2012. С. 223].

Стоит отметить, что в настоящее время лексико-грамматические шаблоны многокомпонентных терминов создаются исследователями для каждого специализированного корпуса отдельно; существуют, например, шаблоны для корпусной лингвистики [Захаров, Хохлова, 2014. С. 187], информационной безопасности [Loukachevich, Dobrov, 2018. P. 185–187], нанотехнологии [Морев, 2012. С. 115–122] и др. В такой ситуации возникает необходимость в разработке специализированной технологии, позволяющей последовательно обрабатывать коллекции текстов разных предметных областей и фиксировать каждую модель многокомпонентного терминологического словосочетания в отдельной базе данных. М. Г. Кружков [2015. С. 152] предлагает использовать для обозначения таких специализированных баз данных термин «надкорпусные базы данных». Они позволяют, с одной стороны, использовать уже созданную базу структурных моделей терминологических словосочетаний, а с другой стороны, дополнять их новыми структурными моделями терминов с минимальными временными и человеческими затратами.

Целью статьи является описание русскоязычных структурных моделей терминологических сочетаний предметной области «Сварка» как основы надкорпусной базы данных о структуре терминов для автоматической разметки специализированных корпусов научно-технических текстов, а также разработка пробного подхода к автоматической разметке многокомпонентных терминов на ее основе.

Терминосистема предметной области «Сварка»

Зачатки современных терминологий зародились еще в древности. Являясь вербализованными результатами речемыслительной деятельности человека, термины проявляются как многофункциональные образования. Во-первых, с исторической точки зрения они представляют собой элементы когниции, т. е. отражают знания, зафиксированные в определенный момент развития той или иной отрасли. Во-вторых, термины могут способствовать развитию данной предметной области, иногда опережая ее состояние и вызывая теоретические и практические изыскания. В-третьих, выходя довольно часто за рамки языка для специальных це-

лей, термины представляют собой понятийные единицы для передачи знаний и опыта, накопленных в данной предметной сфере и вовне.

В условиях научно-технического прогресса отраслевые терминологии и терминосистемы представляют собой постоянно изменяющиеся информационные базы. Каждая вновь появляющаяся в настоящее время предметная область предполагает необходимость своей номинации. В связи с этим возникает существенный вопрос о том, являются ли принципы зарождения и развития терминов, формирования их структуры, а также их группировки в терминологии, хаотичными, т. е. специфичными для каждой терминологии, или же эти процессы подчиняются неким общим принципам, основополагающим законам [Николаева, 2011].

Для выявления устойчивых терминологических сочетаний и рассмотрения уже устоявшихся структурных моделей с последующей возможностью их разметки в корпусах научно-технических текстов была выбрана терминосистема предметной области «Сварка», которая является уже достаточно хорошо исследованной. Работы по упорядочению и унификации сварочной терминологии проводились почти на всем протяжении XX в. Первой попыткой ее систематизации стал «Бюллетень комиссии технической терминологии» под редакцией академика С. А. Чаплыгина и Д. С. Лотте, изданный в 1937 г. [Чаплыгин, Лотте, 1937]. Отвечая на вопрос о необходимости введения «основополагающего» закона, Д. С. Лотте ввел позднее понятие *правильноориентирующие термины*, т. е. те, семантика которых не вступает в противоречие с характером обозначаемого ими понятия [Лотте, 1961].

В 1960-е годы начался следующий этап в работе по унификации сварочной терминологии: появляется «Англо-русский словарь по сварочному производству» В. Т. Золотых и «Словарь-справочник по сварке» Т.В. Кулика, а в 1984 г. была проведена унификация имевшихся на то время терминов, результатом чего стал ГОСТ 2601-84 «Сварка металлов. Термины и определения основных понятий». В этом и последующих стандартах закреплялись термины, наиболее приемлемые по семантической полноте и формальной структуре; в нем же отмечались недопустимые для употребления термины.

Формальная структура термина как фактор эффективности разметки корпуса

Подавляющее большинство терминов имеет формальную структуру лексических единиц того естественного языка, в сфере которого они функционируют и образуются [Лейчик, 1994. С. 169]. Ситуация осложняется, однако, тем, что сферы функционирования и образования не совпадают в случае заимствования термина или образования терминосочетаний на основе компонентов из разных языков. Термин может быть однокомпонентным и состоять из ключевого слова или представлять собой терминологическую группу, в состав которой входит ключевое слово или ядро группы, одно или несколько левых определений и одно или несколько правых или предложных определений, которые уточняют или модифицируют смысл терминологической единицы.

Наиболее сложное явление в процессе автоматической разметки терминов в корпусе научно-технических текстов представляют многокомпонентные термины – терминологические словосочетания, образованные лексическим и синтаксическим способами, то есть словосочетания, образованные по определенным моделям. Способ образование терминов в виде цепочки слов часто используется на практике [Лейчик, 1994. С. 169].

В основе анализа терминологических словосочетаний лежит вычленение исходного терминологического словосочетания и определение последовательности присоединения к нему остальных элементов. Исходным терминологическим словосочетанием, как правило, является двухкомпонентное субстантивное терминологическое словосочетание, которое в рамках трех-, четырехкомпонентного терминологического словосочетания характеризуется более тесными структурно-семантическими отношениями [Циткина, 1988. С. 84].

Таким образом, при создании системы автоматической разметки корпуса научно-технических текстов необходимо установить все возможные структурные модели русских многокомпонентных терминов на примере предметной области «Сварка».

Структурные модели русскоязычных терминологических единиц

При рассмотрении терминосистемы «Сварка» следует отметить присутствие **однокомпонентных** терминов, семантическое наполнение которых определяется наличием различных префиксов (заварка, наварка, подварка, приварка, проварка, сварка). Их малое количество в основном массиве свидетельствует о невысокой продуктивности данной модели терминообразования.

Более продуктивным способом номинации является образование составных терминов, состоящих из двух, трех и более компонентов. Скорее всего, данное явление связано с увеличением семантической дифференциации терминов и их мотивированности, что обусловлено развитием каждой отдельной предметной области сварки [Николаева, 2011].

Двухкомпонентные термины-словосочетания являются наиболее распространенным видом элементов исследуемой понятийной области в процентном отношении. Анализ их структуры показал наличие моделей, характерных для русского языка в целом и для русскоязычных терминологий и терминосистем в частности.

Чаще всего в данной понятийной области встречается атрибутивная модель, созданная по принципу «прилагательное (в функции препозитивного определения) + существительное». Преобладание такой модели обусловлено самой понятийной областью, в которой описывается способ защиты, материал или процесс, с помощью которого и происходит соединение материалов, т. е. сварка. Во всех схемах ядерным компонентом является термин «сварка» как родовой термин данной терминосистемы.

Характерной особенностью данной модели является различие в структуре определяющих элементов [Гринев-Гриневич, 2020]. Большая их часть является простыми прилагательными. Но в массиве можно выделить существенную часть сложных прилагательных, имеющих в составе усеченные основы суффиксальных прилагательных (газовый – *газопрессовая сварка, газоэлектрическая сварка;* вибрационный – *вибродуговая сварка;* высокий – *высокочастотная сварка;* электрический – *электромеханическая сварка, электрохимическая сварка, электрошлаковая сварка;* электронный – *электроннолучевая сварка*). Следует отметить наличие сложных прилагательных, имеющих в составе основу числительного (*двухдуговая сварка, двухэлектродная сварка; многодуговая сварка, многоимпульсная сварка, многоэлектродная сварка*). Словообразовательные модели определяющих элементов двухкомпонентных словосочетаний включают и осново-, и словосложение, которые могут осуществляться как при помощи интерфикса -о- (аргон + о + дуга – *аргонодуговая сварка*; ацетилен + о + кислород – *ацетиленокислородная сварка*; импульсный (окончание отбрасывается, так как не является словообразующим) + о + дуга – *импульсно-дуговая сварка*; магнитный (окончание отбрасывается, так как не является словообразующим) + о + импульс – *магнитно-импульсная сварка*), так и с нулевым интерфиксом (*сварка-пайка*).

Другая модель с определением в качестве определяющего элемента строится по схеме «причастие + существительное» (*защищенная сварка, механизированная сварка*). Она является гораздо менее продуктивной, чем предыдущая и, возможно, может быть рассмотрена как ее разновидность, являясь также атрибутивным словосочетанием.

- Модель, образованная по схеме «существительное + существительное в творительном падеже» (*сварка взрывом, сварка трением, сварка лазером* (двухкомпонентная номинация, полученная при элиминации элемента «лучом» в трехкомпонентной – сварка лучом лазера), *сварка оплавлением*) встречается намного реже. Кроме того, эволюция терминологии пока-

зывает, что подобные термины преобразуются в синонимичные в соответствии с первой моделью. Так, например, *сварка лазером* в настоящее время заменена на *лазерную сварку*.

Трехкомпонентные терминологические словосочетания встречаются реже двухкомпонентных. Перечислим наиболее продуктивные структурные модели.

- «Прилагательное + прилагательное + существительное (ядерный элемент)»: *автоматическая точечная сварка*, *конденсаторная ударная сварка, точечная контактная сварка, лазерная гибридная сварка, электрическая контактная сварка*.

В качестве атрибутивного элемента в подобной конструкции может выступать, как и в двухкомпонентных образованиях, причастие (*комбинированная термитная сварка*).

- «Прилагательное + существительное (ядерный элемент) + существительное в творительном падеже»: *газовая сварка плавлением, холодная сварка давлением, термитная сварка давлением, термитная сварка плавлением, стыковая сварка оплавлением, стыковая сварка сопротивлением, точечная сварка пистолетом*, где последний элемент, хотя формально и соответствует данной схеме, но семантически отличается – вместо процесса обозначен инструмент.
- «Прилагательное + существительное (ядерный элемент) + наречие»: *кузнечная сварка ввилку, кузнечная сварка внахлест, кузнечная сварка встык, кузнечная сварка вразруб, кузнечная сварка врасщеп*.

Менее продуктивными моделями являются:

- «прилагательное + существительное (ядерный элемент) + существительное в винительном падеже»: *высокочастотная сварка пластмасс*; хотя, возможно, данная структурная модель станет более популярной при разработке технологий сварки трудносвариваемых металлов, разнородных металлов и неметаллических материалов;
- «прилагательное + существительное (ядерный элемент) + существительное в предложном падеже»: *диффузионная сварка в вакууме, кузнечная сварка в штампе*.

По материалам лексикографических источников **четырехкомпонентные** терминологические словосочетания встречаются чаще **трехкомпонентных**. Но практика применения стандартных рекомендованных терминов показывает, что в научных текстах по данной предметной области нередко элиминируется первый – родовой – компонент, что превращает словосочетания в **трехкомпонентные** варианты полных стандартных терминов: *дуговая сварка голой проволокой – сварка голой проволокой: дуговая сварка штучными электродами – сварка штучными электродами.*

Наиболее продуктивной структурной моделью является модель «прилагательное + существительное (ядерный элемент) + прилагательное + существительное в творительном падеже» (*аргонодуговая сварка вольфрамовым электродом, аргонодуговая сварка неплавящимся электродом, аргонодуговая сварка металлическим электродом, аргонодуговая сварка плавящимся электродом.*

Другие выявленные структуры являются гораздо менее продуктивными и их встречаемость в общем массиве немногочисленна.

- «Прилагательное + существительное (ядерный элемент) + прилагательное + существительное в творительном падеже (с предлогом «с»)»: *дуговая сварка с магнитным флюсом, газоэлектрическая сварка с магнитным флюсом, кузнечная сварка с контактным нагревом*.
- «Прилагательное + существительное (ядерный элемент) + прилагательное + существительное в предложном падеже»: *газопрессовая сварка в пластическом состоянии*.
- «Прилагательное + существительное (ядерный элемент) + существительное в предложном падеже + существительное в родительном падеже»: *дуговая сварка в среде гелия*.
- «Прилагательное + существительное (ядерный элемент) + существительное в творительном падеже (с предлогом) + существительное в родительном падеже»: *контактная сварка с накоплением энергии*.

- «Прилагательное + прилагательное + существительное (ядерный элемент) + существительное в творительном падеже (с предлогом)»: *автоматическая дуговая сварка под флюсом*.
- «Прилагательное + существительное (ядерный элемент) + существительное в творительном падеже + существительное в творительном падеже (с предлогом)»: *стыковая сварка оплавлением с осадкой*.
- «Прилагательное + существительное (ядерный элемент) + распространенное обособляемое определение, выраженное причастным оборотом»: *кузнечная сварка, выполняемая вручную*.

При анализе структуры выявленных **пятикомпонентных** терминологических словосочетаний прослеживаются следующие модели:

- «прилагательное + прилагательное + существительное (ядерный элемент) + (прилагательное + существительное в творительном падеже)»: *автоматическая дуговая сварка покрытым электродом*;
- «прилагательное + существительное (ядерный элемент) + (существительное в творительном падеже (с предлогом) + прилагательное + существительное в творительном падеже)»: *газопрессовая сварка с нагревом ацетиленокислородным пламенем*;
- «прилагательное + существительное (ядерный элемент) + (существительное в творительном падеже (с предлогом) + прилагательное + существительное в родительном падеже)»: *газовая сварка с расплавлением свариваемых кромок*;
- «прилагательное + существительное (ядерный элемент) + (прилагательное + прилагательное + существительное в творительном падеже)»: *дуговая сварка газообразующей электродной проволокой*;
- «прилагательное + существительное (ядерный элемент) + (прилагательное + существительное в творительном падеже (с предлогом) + существительное в родительном падеже)»: *дуговая сварка с поперечными колебаниями электрода, точечная сварка с интенсивным охлаждением электродов, многоточечная сварка с поочередным зажатием электродов*;
- «прилагательное + существительное (ядерный элемент) + (существительное в творительном падеже + существительное в родительном падеже + существительное в родительном падеже)»: *дуговая сварка методом опирания электрода*;
- «прилагательное + существительное (ядерный элемент) + (распространенное обособляемое определение, выраженное трехкомпонентным причастным оборотом)»: *кузнечная сварка, выполняемая под молотом или прессом*.

Подход к автоматическому выделению терминов-кандидатов на основе структурных моделей

Предлагаемый подход к автоматической разметке терминов на основе надкорпусной базы структурных моделей терминологических словосочетаний состоит из пяти основных этапов. В качестве примера рассмотрим терминологическую разметку для следующего фрагмента текста:

Дуговая сварка неплавящимся электродом в защитной атмосфере инертного газа – метод дуговой сварки, который используется для сварки алюминия, магния и их сплавов, нержавеющей стали, никеля, меди, бронзы, титана, циркония и других неферромагнитных металлов.

На **первом этапе** производится автоматический морфологический анализ текста, то есть каждому слову приписывается часть речи, род, число, падеж и др. В качестве примера ниже представлен морфологический анализ, выполненный при помощи инструментов сайта www.textovod.ru. Результат анализа имеет следующий вид:

Дуговая $^{\text{ПРИЛ жр,ед,им}}$ сварка $^{\text{СУЩ,неод,жр ед,им}}$ неплавящимся $^{\text{ПРИЛ мр,ед,тв}}$ электродом $^{\text{СУЩ,неод,мр ед,тв}}$ в $^{\text{ПР}}$ защитной $^{\text{ПРИЛ,кач жр,ед,тв}}$ атмосфере $^{\text{СУЩ,неод,жр ед,пр}}$ инертного $^{\text{ПРИЛ,кач мр,ед,рд}}$ газа $^{\text{СУЩ,неод,жр,sg,geo}}$

ед,им _ ЗПР метод СУЩ,неод,мр ед,им дуговой ПРИЛ мр,ед,им сварки СУЩ,неод,жр ед,рд , ЗПР который ПРИЛ,мест-п,субст?,Анаф мр,ед,им используется ГЛ,несов,неперех ед,3л,наст,изъяв для ПР сварки СУЩ,неод,жр ед,рд алюминия СУЩ,неод,мр ед,рд , ЗПР магния СУЩ,неод,мр ед,рд и СОЮЗ их МС,3л,Анаф мн,рд сплавов СУЩ,неод,мр мн,рд , ЗПР нержавеющей ПРИЛ жр,ед,рд стали ГЛ,сов,неперех мн,прош,изъяв , ЗПР никеля СУЩ,неод,мр ед,рд , ЗПР меди СУЩ,неод,жр ед,рд , ЗПР бронзы СУЩ,неод,жр ед,рд , ЗПР титана СУЩ,неод,мр ед,рд, ЗПР циркония СУЩ,неод,мр ед,рд и СОЮЗ других ПРИЛ,мест-п,субст? мн,рд неферромагнитных ПРИЛ мн,рд металлов СУЩ,неод,мр мн,рд . ЗПР

На **втором этапе** из размеченного текста необходимо убрать части речи (см. зачеркнутые единицы), которые не входят в состав терминологических словосочетаний, такие как глаголы, местоимения, междометия и т. д. (в том числе знаки препинания):

Дуговая ПРИЛ жр,ед,им сварка СУЩ,неод,жр ед,им неплавящимся ПРИЛ мр,ед,тв электродом СУЩ,неод,мр ед,тв в ПР защитной ПРИЛ,кач жр,ед,тв атмосфере СУЩ,неод,жр ед,пр инертного ПРИЛ,кач мр,ед,рд газа СУЩ,неод,жр,sg,гео ед,им ~~__ ЗПР~~ метод СУЩ,неод,мр ед,им дуговой ПРИЛ мр,ед,им сварки СУЩ,неод,жр ед,рд ~~, ЗПР~~ ~~который ПРИЛ,мест-п,субст?,Анаф мр,ед,им~~ ~~используется ГЛ,несов,неперех ед,3л,наст,изъяв~~ для ПР сварки СУЩ,неод,жр ед,рд алюминия СУЩ,неод,мр ед,рд ~~, ЗПР~~ магния СУЩ,неод,мр ед,рд ~~и СОЮЗ их МС,3л,Анаф мн,рд~~ сплавов СУЩ,неод,мр мн,рд ~~, ЗПР~~ нержавеющей ПРИЛ жр,ед,рд ~~стали ГЛ,сов,неперех мн,прош,изъяв~~ ~~, ЗПР~~ никеля СУЩ,неод,мр ед,рд , ~~ЗПР~~ меди СУЩ,неод,жр ед,рд ~~, ЗПР~~ бронзы СУЩ,неод,жр ед,рд ~~, ЗПР~~ титана СУЩ,неод,мр ед,рд ~~, ЗПР~~ циркония СУЩ,неод,мр ед,рд ~~и СОЮЗ других ПРИЛ,мест-п,субст? мн,рд~~ неферромагнитных ПРИЛ мн,рд металлов СУЩ,неод,мр мн,рд ~~. ЗПР~~

После удаления любого из элементов, не входящих в состав терминологического словосочетания, следующая цепочка слов начинается с новой строки. В результате обработки рассматриваемого фрагмента текста на втором этапе получаем следующий список терминов-кандидатов:

Дуговая ПРИЛ жр,ед,им сварка СУЩ,неод,жр ед,им неплавящимся ПРИЛ мр,ед,тв электродом СУЩ,неод,мр ед,тв в ПР защитной ПРИЛ,кач жр,ед,тв атмосфере СУЩ,неод,жр ед,пр инертного ПРИЛ,кач мр,ед,рд газа СУЩ,неод,жр,sg,гео ед,им

метод СУЩ,неод,мр ед,им дуговой ПРИЛ мр,ед,им сварки СУЩ,неод,жр ед,рд
для ПР сварки СУЩ,неод,жр ед,рд алюминия СУЩ,неод,мр ед,рд
магния СУЩ,неод,мр ед,рд
сплавов СУЩ,неод,мр мн,рд
нержавеющей ПРИЛ жр,ед,рд
никеля СУЩ,неод,мр ед,рд
меди СУЩ,неод,жр ед,рд
бронзы СУЩ,неод,жр ед,рд
титана СУЩ,неод,мр ед,рд
циркония СУЩ,неод,мр ед,рд
неферромагнитных ПРИЛ мн,рд металлов СУЩ,неод,мр мн,рд

На **третьем этапе** однокомпонентные термины-кандидаты автоматически переходят сразу на пятый этап, а для словосочетаний еще необходимо выполнить проверку на соответствие ряду дополнительных формальных и семантических условий. Так, например, терминологическое словосочетание в своей исходной форме не может начинаться с предлога, и программа должна уметь отделять его: *для* ПР *сварки* СУЩ,неод,жр ед,рд *алюминия* СУЩ,неод,мр ед,рд; или же после прилагательного может следовать семантически связанное существительное, *нержавеющей* ПРИЛ жр,ед,рд. В первом случае предлог подлежит автоматическому исключению из терминологического словосочетания, а во втором, наоборот, слово, следующее после анализируемого, нужно добавить, чтобы сформировать полную словарную внеконтекстуальную структуру терминологического словосочетания.

На **четвертом этапе** происходит сравнение полученных таким образом словосочетаний со структурными (образцовыми, эталонными) моделями терминологических словосочетаний.

Дуговая ПРИЛ жр,ед,им сварка СУЩ,неод,жр ед,им неплавящимся ПРИЛ мр,ед,тв электродом СУЩ,неод,мр ед,тв в ПР защитной ПРИЛ,кач жр,ед,тв атмосфере СУЩ,неод,жр ед,пр инертного ПРИЛ,кач мр,ед,рд газа СУЩ,неод,жр,sg,гео ед,им

метод СУЩ,неод,мр ед,им дуговой ПРИЛ мр,ед,им сварки СУЩ,неод,жр ед,рд
сварки СУЩ,неод,жр ед,рд алюминия СУЩ,неод,мр ед,рд
нержавеющей ПРИЛ жр,ед,рд стали ГЛ,сов,неперех мн,прош,изъяв
неферромагнитных ПРИЛ мн,рд металлов СУЩ,неод,мр мн,рд

На **пятом этапе** каждое словосочетание проверяется по терминологическому словарю, а оставшиеся нераспознанными словосочетания передаются на ручную разметку:

Дуговая ПРИЛ жр,ед,им сварка СУЩ,неод,жр ед,им неплавящимся ПРИЛ мр,ед,тв электродом СУЩ,неод,мр ед,тв в ПР защитной ПРИЛ,кач жр,ед,тв атмосфере СУЩ,неод,жр ед,пр инертного ПРИЛ,кач мр,ед,рд газа СУЩ,неод,жр,sg,гео ед,им

сварки СУЩ,неод,жр ед,рд алюминия СУЩ,неод,мр ед,рд
нержавеющей ПРИЛ жр,ед,рд стали ГЛ,сов,неперех мн,прош,изъяв

Таким образом, в рассмотренном примере из 12 терминов-кандидатов только 4 передаются на ручную разметку. Это значит, что эффективность автоматической разметки равна здесь около 66,5 %, что позволяет надеяться на значительное уменьшение доли ручного труда при масштабной разметке корпуса научно-технических текстов. При этом мы считаем, что каждый автоматически выявленный термин, соответствующий структурным моделям, но не содержащийся в словаре терминов корпуса, должен проходить дополнительную ручную разметку.

Заключение

Итак, для реализации проекта по автоматической разметке корпусов научно-технических текстов появилась необходимость в выборе достаточно устоявшегося массива терминов, который мог бы продемонстрировать характерные структурные образования, присущие развитым терминологиям и особенно терминосистемам. В качестве такого устоявшегося массива была выбрана терминосистема предметной области «Сварка», которая, с одной стороны, имеет давнюю историю формирования, лексикографического описания и изучения, а с другой стороны, является относительно молодой, поскольку основной этап ее наполнения пришелся на XX в., что упрощает процедуры подбора и анализа терминов. В связи с тем, что терминосистема этой предметной области достаточно хорошо изучена, а сама она стандартизирована, она может стать надежной основой для создания в близкосрочной перспективе особой *надкорпусной* базы данных, описывающей структурные модели терминологических словосочетаний предметной области «Сварка» и требующей инструментов их разметки.

Рассмотрев формальную структуру элементов терминосистемы «Сварка», мы выявили, что ее наиболее частотной и, вероятно, продуктивной струкутруно-семантической моделью является сочетание опорного ядерного существительного с именем прилагательным в позиции препозитивного определения с функцией передачи вариативного предметного расширения семантики опорного предметного существительного. Наиболее явно и устойчиво данная синтаксическая модель функционирует в случае двухкомпонентных словосочетаний, но анализ более сложных образований показывает, что модель «с левым определением, присоединенным к предметно-семантическому – родовому – ядру термина» присутствует и в них, демонстрируя связанные родовые признаки. Однако дальнейшее семантическое развитие терминологического словосочетания чаще всего происходит справа от опорного компонента по мере усложнения семантики постпозитивного определения (или определений), несущего в себе видовые особенности. На основе этих структурных и семантических особенностей, свойственных описанным моделям терминологических словосочетаний, в статье предложен подход к автоматической разметке многокомпонентных терминов. В частности, надкорпус-

ная база, идея создания которой защищается авторами на примере терминосистемы «сварка», может быть использована для автоматической разметки научно-технических текстов самых разных предметных областей. Перспективой дальнейшего исследования является разработка методов автоматического определения ядерного элемента терминологических словосочетаний, разбиения терминов, состоящих из пяти и более компонентов, на сочетания двух и более терминов, а также способы отсеивания общеупотребительных слов с оценочной или внепредметной семантикой (*современный, передовой, новый* и т. п.).

Список литературы

Захаров В. П. Корпуса русского языка // Труды института русского языка им. В. В. Виноградова. 2015. Т. 6. С. 20–65.

Захаров В. П., Богданова С. Ю. Корпусная лингвистика: учебник. 3-е изд., перераб. СПб.: Изд-во С.-Петерб. ун-та, 2020. 234 с.

Захаров В. П., Хохлова М. В. Автоматическое извлечение терминов из специальных текстов с использованием дистрибутивно-статистического метода как инструмент создания тезаурусов // Структурная и прикладная лингвистика. 2012. № 9. С. 222–233.

Захаров В. П., Хохлова М. В. Автоматическое выявление терминологических словосочетаний // Структурная и прикладная лингвистика. 2014. № 10. С. 182–200.

Гринев-Гриневич С. В., Сорокина Э. А. Опыт описания формальной структуры термина (на материале английской терминологии лексикологии // Вестник Московского государственного областного университета. Серия: Лингвистика, 2020. № 5. С. 74–85. DOI 10.18384/2310-712X-2020-5-74-85

Кружков М. Г. Информационные ресурсы контрастивных лингвистических исследований: электронные корпуса текстов // Системы и средства информатики. 2015. Т. 25, № 2. С. 140–159.

Лейчик В. М. Исходные понятия, основные положения, определения современного терминоведения и терминографии // Вестник Харьковского политехнического университета. 1994. № 1. С. 147–180.

Лесников В. С. Виды разметок текстовых корпусов русского языка // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2019. № 9. С. 27–30.

Лотте Д. С. Основы построения научно-технической терминологии. М.: Изд-во АН СССР, 1961. 158 с.

Морев Н. А. К проблеме лингвистического анализа терминологии в области нанотехнологий (о необходимости разработки исследовательского корпуса терминологических единиц) // Вестник МГЛУ. 2012. № 13 (646). С. 115–124.

Нагель О. В. Корпусная лингвистика и ее использование в компьютеризированном языковом обучении // Язык и культура, 2008. № 4. С. 53–59.

Николаева Н. С. Особенности происхождения основных терминов терминосистемы «Сварка» (на материале английской и русской терминологий) // Вестник Московского государственного областного университета. Серия: Лингвистика. 2011. № 1. С. 132–138.

Циткина Ф. А. Терминология и перевод. Львов: Высшая школа, 1988. 157 с.

Терминология сварки металлов / Под ред. С. А. Чаплыгина, Д. С. Лотте. М.: Изд-во Акад. наук СССР, 1937. 31 с.

Butenko Iu. I., Garazha V. V. BMSTU Corpus of Scientific and Technical Texts: Conceptual Framework. *Applied Linguistics Research Journal*, 2021, vol. 5(3), p. 76–81. DOI 10.14744/alrj.2021.15579

Loukachevitch N., Dobrov B. Ontological Resources for Representing Security Domain in Information-Analytical System. *Open Semantic Technologies for Intelligent Systems Design*, 2018. № 8. P. 185–191.

References

Butenko Iu. I., Garazha V. V. BMSTU Corpus of Scientific and Technical Texts: Conceptual Framework. *Applied Linguistics Research Journal*, 2021, vol. 5 (3), p. 76–81. DOI 10.14744/alrj.2021.15579.

Citkina F. A. Terminologija i perevod [Terminology and translation]. Lvov, Vysshaja shkola, 1988, 157 p.

Grinev-Grinevich S. V., Sorokina Je. A. Describing the formal structure of a term (based on the english terminology of lexicology). *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Serija: Lingvistika*, 2020, no. 5, p. 74–85. (in Russ).

Kruzhkov M. G. Information resources for contrastive studies: electronic text corpora. *Systems and means of informatics*, 2015, no. 25 (2), p. 140–159. (in Russ.)

Lejchik V. M. Iskhodnye ponyatiya, osnovnye polozheniya, opredeleniya sovremennogo terminovedeniya i terminografii [Basic concepts, basic provisions, definitions of modern terminology and terminography]. *Vestnik Har'kovskogo politekhnicheskogo universiteta*, 1994, vol. 1, p. 147–180. (in Russ.)

Lesnikov S. V. The types of marking of text corpora of the Russian language. *Nauchno-tekhnicheskaya informaciya. Seriya 2. Informacionnye processy i sistemy*, 2019, vol. 9, p. 27–30. (in Russ.) DOI 10.36535/0548-0027-2019-09-4

Lotte D. S. *Osnovy postroeniya nauchno-tekhnicheskoj terminologii* [Fundamentals of building scientific and technical terminology]. Moscow Izdatel'stvo AN SSSR, 1961, 158 p. (in Russ.)

Loukachevitch N., Dobrov B. Ontological Resources for Representing Security Domain in Information-Analytical System. *Open Semantic Technologies for Intelligent Systems Design*, 2018, vol. 8, p. 185–191.

Morev N. A. K probleme lingvisticheskogo analiza terminologii v oblasti nanotehnologij (o neobhodimosti razrabotki issledovatel'skogo korpusa terminologicheskih edinic) [On the Problem of Linguistic Analysis of Nanotechnology Terminology (On the Need to Develop a Research Corpus of Terminological Units)]. *Vestnik MGLU*, 2012, no. 13 (646), p. 115–124.

Nagel O. V. Corpus linguistics and its use in computer-based language teaching. *Language and culture*, 2008, no. 4, p. 53–59 (in Russ.)

Nikolaeva N. S. Osobennosti proiskhozhdeniya osnovnyh terminov terminosistemy "Svarka" (na materiale anglijskoj i russkoj terminologij) [Peculiarities of origin of the basic terms of the "Welding" system (on the material of English and Russian terminology)]. *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Lingvistika*, 2011, no. 1, p. 132–138. (in Russ.)

Terminologija svarki metallov [Terminology of metal welding]. Eds. S. A. Chaplygina, D. S. Lotte. Moscow, AS USSR, 1937, 31 p. (in Russ.)

Zaharov V. P. Russian corpora. *Trudy instituta russkogo jazyka imeni V. V. Vinogradova*, 2015, vol. 6, p. 20–65. (in Russ.)

Zakharov, V. P., Khokhlova, M. V. Automatic extracting of terminological phrases. *Strukturnaya i prikladnaya lingvistika*, 2014, vol. 10, p. 182–200. (in Russ.)

Zakharov, V. P., Bogdanova, S. Yu. Methods in Corpus Linguistics: Course book. 3rd revised ed. St. Petersburg, SPb University Press, 2020, 234 p. (in Russ.)

Zakharov, V. P., Khokhlova, M. V. Automatic term extraction and statistical analysis in a special text corpus as a tool for thesaurus construction. *Strukturnaya i prikladnaya lingvistika*, 2019, no. 9, p. 222–233. (in Russ.)

Материал поступил в редколлегию
Date of submission
11.01.2021

Сведения об авторах / Information about the Authors

Бутенко Юлия Ивановна, канд. тех. наук, доцент кафедры «Романо-германские языки» Московского государственного технического университета им. Н. Э. Баумана (Москва, Россия); ассистент кафедры иностранных языков факультета гуманитарных и социальных наук, Российский университет дружбы народов (Москва, Россия)

Iuliia I. Butenko, candidate of technical sciences, associate professor, department of Romance-Germanic languages, Bauman Moscow State Technical University (Moscow, Russian Federation); assistant professor of Foreign Languages Department, Peoples Friendship University of Russia (Moscow, Russian Federation)

iubutenko@bmstu.ru
ORCID 0000-0002-9776-5709

Николаева Наталия Сергеевна, канд. филол. наук, доцент, доцент кафедры «Английский для машиностроительных специальностей», Московский государственный технический университет им. Н. Э. Баумана (Москва, Россия)

Natalia S. Nikolaeva, candidate of philological sciences, associate professor, department of English for machine building specialties, Bauman Moscow State Technical University (Moscow, Russian Federation)

natalynic@yandex.ru
ORCID 0000-0001-6230-8340

Маргарян Татьяна Дмитриевна, канд. ист. наук, доцент, заведующая кафедрой «Английский для машиностроительных специальностей», Московский государственный технический университет им. Н. Э. Баумана (Москва, Россия)

Tatiana D. Margaryan, candidate of historical sciences, associate professor, head of English for machine building specialties department, Bauman Moscow State Technical University (Moscow, Russian Federation)

sunnymood77@hotmail.com
ORCID 0000-0002-1645-1215