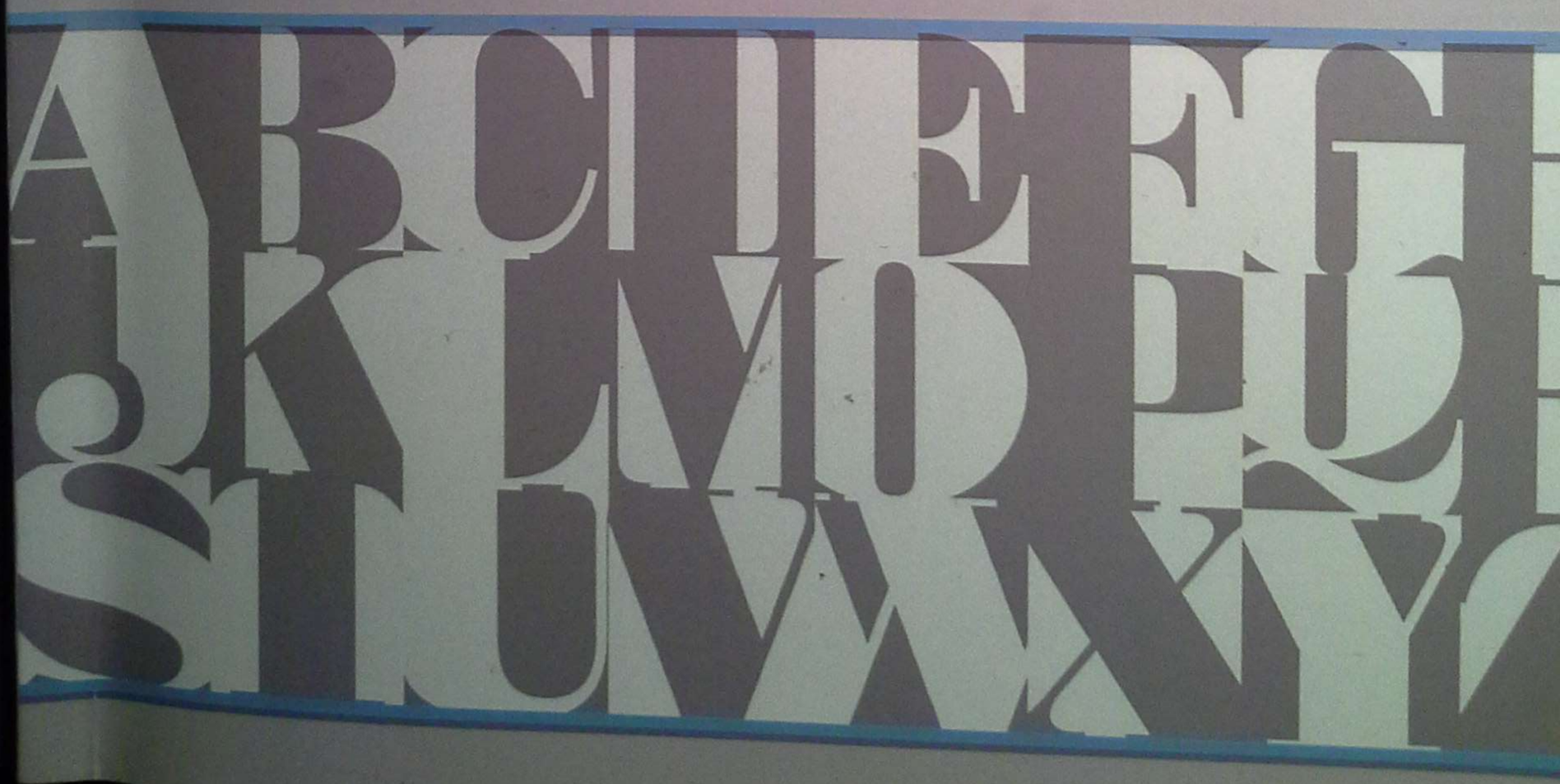
Л. Н. Беляева, О. А. Данилова, Т. Л. Джепа, О. Н. Камшилова,
Е. В. Карнуп, В. Р. Нымм, С. В. Чумилкин

ЛЕКСИКОГРАФИЧЕСКИЙ ПОТЕНЦИАЛ

современных лингвистических технологий

м о н о г р а ф и я



УДК 801
ББК 81.1 я 73
Л 43

Рецензенты:
д-р филологических наук, профессор **В. Е. Чернявская** (Санкт-Петербургский государственный политехнический университет);
д-р филологических наук, профессор **Н. Л. Шубина** (РГПУ им. А. И. Герцена)

Л 43 Лексикографический потенциал современных лингвистических технологий: монография / Беляева Л. Н., Данилова О. А., Джепа Т. Л., Камшилова О. Н., Карнуп Е. В., Нымм В. Р., Чумилкин С. В.; под ред. Л. Н. Беляевой. — СПб.: Изд-во ООО «Книжный Дом», 2014. — 168 с.

ISBN 978-5-94777-351-4

Постоянное изменение науки, техники и технологий, развитие новых направлений и новых отраслей знаний приводит сегодня к существенному устареванию специализированных переводных словарей, предназначенных для поддержки самостоятельной работы студентов и деятельности переводчиков. В монографии на примере терминологии филологии детально рассматриваются проблемы гармонизации терминологии, проблемы использования параллельных и псевдопараллельных корпусов текстов для решения задачи извлечения и выравнивания терминов. Оцениваются возможности извлечения терминов из одноязычных текстов на глобальном английском и русском языках, предлагаются ограничения процедуры лемматизации при анализе текстов на русском языке. Анализируются особенности критерия терминологичности.

ISBN 978-5-94777-351-4

СОДЕРЖАНИЕ

ПРЕДИСЛОВИЕ

Современная экономика знаний требует обеспечения полной многоязычности информации на всех этапах ее существования, что может быть обеспечено только на базе применения информационных технологий (ИТ) для создания систем поиска, генерации и поддержки многоязычной информации, для локализации данных и программного обеспечения. Эти задачи решаются путем создания и внедрения в деятельность специалистов разных направлений практических систем информационного поиска, автоматического (машинного) перевода, компьютерных словарных и обучающих систем и т. д. ИТ в области естественного языка (лингвистические технологии), реализующие алгоритмы автоматической переработки текста, являются необходимым условием решения множества актуальных задач.

Разработка методов извлечения знаний из источников разного состава, природы и качества [72] основывается на применении гуманитарных технологий в социальной сфере. Свободное и произвольное движение информации при решении конкретных проблем определяет необходимость не просто извлечь информацию, но и сделать ее активной, т. е. обеспечить максимальное использование информации на всех видах носителей, электронных в том числе, и содействовать распространению и получению знаний. В связи с этим прикладная лексикография становится сегодня важной отраслью прикладной лингвистики.

Автоматизированная лексикография является направлением прикладной лексикографии, характеризующимся особыми подходами не только к способам отображения информации, но и к содержанию словаря. Это связано с тем, что современные задачи словарной науки и практики заключаются:

1) в совершенствовании существующих типов словарей;
2) создании новых типов словарей на основе новых актуальных принципов и технологий;
3) соединении уже известных типов словарей в новые словарные комплексы;
4) создании компьютерной методологии лексикографической деятельности на основе проблемно-ориентированных корпусов текстов.

Соответственно, современные проблемы автоматизированной лексикографии можно рассматривать в нескольких аспектах:

- в аспекте создания электронных словарных баз, предназначенных для использования в задачах перевода и обучения языкам;
- в аспекте использования корпусов текстов для проведения лексикологических исследований, а также для создания и ведения различных лексикографических систем, в том числе и переводных словарей;
- в аспекте создания и ведения автоматических словарей различных информационных систем.

Умение пользоваться различными видами систем автоматической переработки информации, умение организовать свою исследовательскую и методическую работу с помощью компьютера является условием плодотворной работы и соответствия современному уровню знаний. Процедуры составления различных видов словарей, конкордансов, справочников, гипертекстов, являющиеся составной частью работы в гуманитарных областях знаний, без использования компьютера превращаются в рутинный трудоемкий процесс.

Предлагаемая читателю монография является результатом работы коллектива исследователей в рамках научно-исследовательского проекта «Исследование терминологических особенностей английского и русского языков филологии как языков для специальных целей», поддержанного грантом Министерства образования и науки в 2013 г.

Работы выполнялись на основе электронных массивов филологических текстов (корпусов) на русском и английском языках. Большое значение придается рассмотрению и разработке принципов создания псевдопараллельных (сопоставительных) корпусов для задач переводной лексикографии.

Другой важный аспект, рассматриваемый в работе, – проблемы и перспективы современной терминографии, в частности методы и метрики извлечения терминов и кандидатов в термины из корпу-

текстов.

На основе проведенного исследования установлены модели терминологических словосочетаний английского языка филологии в сравнении с моделями русского языка. Сопоставительный анализ структуры именных групп в русском и английском подъязыках предметной области «филология» позволил сделать вывод о различиях в принципах номинирования сложных объектов и степени отражения особенностей этих внеязыковых объектов при расчлененной (многокомпонентной) номинации.

Особое внимание в монографии уделено описанию сетевого инструментария анализа текста. Для оперативного извлечения сетевой информации в рамках работы был программно разработан и реализован алгоритм графематического анализа текстов, извлекаемых из системы Интернет.

Для извлечения и хранения терминологических словосочетаний была исследована возможность использования специализированного автоматического словаря, разработанного в лаборатории машинного перевода филологического факультета РГПУ им. А. И. Герцена на материалах системы машинного перевода WORD+. На этой основе создана база переводных терминологических словарей и разработана методика конвертации англо-русского словаря в словарь русско-английский.

Материалы монографии могут быть полезны филологам разных специальностей, студентам, аспирантам, преподавателям для учебной и исследовательской работы.

1. ЛЕКСИКОГРАФИРОВАНИЕ ЯЗЫКОВ ДЛЯ СПЕЦИАЛЬНЫХ ЦЕЛЕЙ

1.1. ПРОБЛЕМЫ ПЕРЕВОДА И ГАРМОНИЗАЦИИ ТЕРМИНОЛОГИИ В СФЕРЕ МЕЖЪЯЗЫКОВОЙ КОММУНИКАЦИИ

Межъязыковая коммуникация в сфере науки в современном глобализирующемся обществе является одним из важнейших направлений обмена информацией. Научная коммуникация, устная или письменная, реализуется в виде текста, предназначенного для того, чтобы сформулировать научную проблему, описать результат научного исследования, обобщить полученные данные и сделать выводы по результатам проведенного исследования. Главными причинами, затрудняющими коммуникацию в научной сфере, являются проблемы лингвистические – языковые и речевые, эти проблемы увеличиваются многократно, когда речь заходит о межъязыковой коммуникации, коммуникации на двух или нескольких языках. Язык научных текстов может рассматриваться как особый функциональный язык, разновидность общелитературного языка, которому придается большая автономность — язык для профессиональных (специальных) целей. В отличие от большинства искусственных систем переработки, хранения и передачи информации, язык представляет собой открытую динамическую неравновесную метасистему [128]. При этом текст можно рассматривать как результат решения задачи передачи информации и источник (отправную точку) ее извлечения. Соответственно, содержание научного текста, т. е. та часть его смысла, которая является универсальной и может быть извлечена при минимальном совпадении тезаурусов автора и получателя, определяется в основном информацией об объектах, лингвистически описываемых именами существительными и именными словосочетаниями. В си-

туации перевода возможность извлечения информации из научного текста определяется для читателя корректностью передачи терминов — имен объектов.

Основой межъязыковой коммуникации в сфере науки является адекватный перевод научного текста с одного или нескольких языков на другой или другие языки. Такой перевод (как научный, так и технический) требует точности и ясности изложения, максимально полного соответствия перевода оригиналу и корректности выбранной номинации понятий и объектов. При этом адекватность восприятия текста на лексическом уровне определяется насыщенностью текста именными единицами, степенью компрессии и/или развернутостью номинации объектов. Если рассматривать это в синергетическом аспекте, то можно предположить, что при «перенасыщенности» текста перевода слишком длинными или чрезмерно свернутыми лексическими комплексами, текст разрушается и его восприятие затруднено или даже невозможно. Учет сложности восприятия подобных лексических образований и их преобразование в более простые комплексы позволяют создавать адекватный текст перевода, т. е. от хаоса нагромождения терминологических единиц переходить к новому состоянию, отражающему структуру денотативного пространства текста и терминосистему конкретного языка в том числе.

Таким образом, на первый план выходит проблема достижения эквивалентности перевода научных текстов, затрудняет которую необходимость передачи содержания исходного текста с помощью треминосистемы языка перевода. Сегодня можно утверждать как некоторую аксиому, что объективное исследование терминологии конкретной предметной области требует предварительного моделирования этой предметной области для того, чтобы установить систему ее понятий и связей между ними. При этом принято различать **терминополе** как системное образование плана содержания, представляющее собой организованную совокупность специальных понятий и связей, и **терминосистему** как совокупность языковых средств, терминополе отражающих [17].

Продуктивность научного общения во многом зависит от того, насколько избранный метаязык (а терминосистема определенной области знания является ее метаязыком) понятен тем, кто в ней работает, насколько этот метаязык способствует установлению контакта между специалистами. В случае же несовместимости метаязыковых

систем / терминосистем и терминологий говорящего и слушающего (и соответственно, пишущего и читающего) обмен научной информацией существенно осложняется ввиду отсутствия единой системы понятий и терминов, являющейся основой плодотворного научного общения. Необходимо отметить также, что метаязык неизбежно отражает мировоззрение ученого, т. е. систему его представлений об окружающем мире [26].

Наиболее традиционным подходом к изучению соотношения терминополей и отражающих их терминосистем является структурное моделирование на основе тезауруса. Материалом для тезауруса служит синхронный срез лексики, который определяется, как правило, современным состоянием моделируемого терминополя и отражающей его терминосистемы, которая извлекается из словарного и текстового материала на одном или нескольких языках. В то же время мы привыкли постулировать постоянное развитие различных научных и технических сфер деятельности человека, уделяя недостаточное внимание отражению этого развития в языке. Если вслед за Гийомом считать, что язык представляет собой систему систем или диахронию синхроний, то необходим именно диахронный подход к моделированию терминосистемы для того, чтобы установить границы возможных синхронических срезов, причины и особенности перехода от одного состояния к другому [17].

Такое различие терминосистем исходного языка и языка перевода диктует необходимость изучения треминосистем с целью выявления расхождений в системе понятий, выражаемых терминами исходного языка (ИП) и языка перевода (ПЯ), с последующим упорядочиванием, стандартизацией и унификацией терминологий разных языков, гармонизацией терминосистем этих языков, обеспечивающей решение проблем перевода терминов и эффективность межъязыковой коммуникации.

Определение и структура термина. Обращаясь к специфике понятия термина, необходимо отметить, что поиски точной дефиниции этого понятия не прекращаются до сих пор.

Термин в широком смысле определяется как специальное слово или словосочетание, принятое для обозначения чего-нибудь в той или иной среде, профессии, требующее дефиниции для установления своего значения в соответствующей системе понятий [34].

Термин должен соответствовать определенному ряду требований, к которым относят:

1) однозначность или стремление к однозначности;
2) точность семантики;
3) стилистическую нейтральность, отсутствие экспрессивности;
4) номинативность;
5) понятийность;
6) системность;
7) мотивированность [84; 76].

Однако вопрос о требованиях к термину также представляется довольно спорным, поскольку во многих работах лингвистов доказывается относительность данных требований и/или их неоднозначный характер. Поэтому целесообразно говорить не о самих требованиях, предъявляемых к термину, а скорее о лингвистических особенностях и признаках, которые выделяют термины из ряда других лексических единиц языка.

Выделяют четыре признака термина, которые находятся во взаимосвязи и влияют друг на друга:

- воспроизводимость — неизменность основных элементов содержательной и формальной структуры термина в разных условиях его существования;
- парадигматическая вариантность — возможность замены термина синонимической лексической единицей;
- относительная устойчивость — стабильной формальной структуры термина в процессе терминоупотребления и его общепринятость;
- внедренность — использование термина в течение определенного периода, достаточного для закрепления его в сознании членов социальной группы, использующих его [39; 53].

Термины могут иметь различную структуру. По числу компонентов выделяют:

- термины-слова, или однословные термины, реже именуемые монолексемными, к которым могут быть отнесены и сложные термины, образованные сложением основ и имеющие слитное или дефисное написание;
- термины-словосочетания, или составные, многокомпонентные термины, так называемые именные группы;
- термины-аббревиатуры;
- термины, выраженные символами, сочетаниями слов и букв-символов, сочетаниями слов и цифр-символов [28].

Вопрос о специфике терминов и проблемах их перевода при межъязыковой коммуникации всегда занимал особое место в сопоставительном языкознании. Важнейшей задачей перевода является обеспечение эквивалентности содержания текстов оригинала и перевода, особенно это актуально для перевода специальных научных текстов. При переводе специальных текстов научной направленности терминам следует уделять пристальное внимание, так как именно они определяют информационное содержание специального текста, являясь своеобразными ключами, организующими, структурирующими и кодирующими содержащуюся в тексте информацию. Поэтому именно применительно к терминам существенным становится вопрос о возможности достижения эквивалентности при существовании различия кодовых единиц, коими являются термины.

Спецификой перевода терминов является сохранение в переводе содержательной точности единиц ИЯ и обеспечение абсолютной идентичности понятий, выражаемых терминами ИЯ и ПЯ. Иными словами, если необходимым условием успешной коммуникации, как устной, так и письменной, является тождественность кодов отправителя и получателя информации соответствующей научной области, то именно обеспечение тождественности означаемых терминами понятий представляет собой главную задачу перевода специального текста. Важным шагом на пути к решению проблем перевода терминов при межъязыковой коммуникации является выявление расхождений в системе понятий, выражаемых терминами ИЯ и ПЯ.

Таким образом, важно еще раз отметить два основных свойства термина — это специфичность употребления (особая сфера употребления) и содержательная точность. Основной задачей терминологов является выявление системных связей между терминами на основе анализа их функционирования в специальном тексте, что позволяет установить парадигматические отношения, в которые вступает термин в терминосистеме, и точно определить объем выражаемого им понятия, его место в терминосистеме ИЯ и обнаружить, насколько точно передано значение терминологической единицы в ПЯ.

Терминология и терминосистема. Термин представляет собой член сложившейся совокупности единиц и входит в терминологию и терминосистему. Существуют следующие точки зрения на рассмотрения понятия «терминология», а именно:

- как лексика особого языка науки [34];
- как совокупность специальных наименований, объединенных в терминосистемы [78];
- как фрагмент лексико-семантической системы языка [31];
- как лексическая подсистема внутри лексической системы национального языка, обеспечивающая специальную коммуникацию [82];
- как синоним терминосистемы [30];
- как совокупность лексических единиц естественного языка, обозначающих понятия определенной специальной области знаний или деятельности, стихийно складывающаяся в процессе зарождения и развития этой области [51, с. 14].

Понятия «терминология» и «терминосистема» взаимосвязаны: терминология моделирует реальную предметную область, но системные свойства в ней выражены неявно, тогда как терминосистема представляет формализованную лингвистическую модель терминополя или логико-понятийной системы. Модель логико-понятийной системы может быть записана на любом естественном или искусственном языке. Термины могут быть выражены словами, словосочетаниями, аббревиатурами, символами, сочетаниями слов и букв-символов, сочетаниями слов и цифр-символов. Сама терминосистема является компонентом лексико-семантической системы того или иного национального языка [28].

Терминосистема развивается как реакция на терминологию; если терминология достаточно динамична, то терминосистема способствует ее упорядочению.

Таким образом, если терминологии присущи такие свойства, как способность отражать понятия определенной области [67], отсутствие четкой системности, стихийность образования, то терминологическая система может рассматриваться как знаковая модель специальной области знаний или деятельности, элементам которой служат лексические единицы конкретного языка для специальных целей (ЯСЦ), а структура в целом адекватна структуре системы понятий данной области знаний [51].

Терминология и терминосистема также различаются составом входящих в них единиц. Терминология характеризуется неоднородностью своего состава, помимо терминов в ней выделяются:

- номены — номенклатуры [80];
- терминоиды — специальные лексические единицы, используемые для номинации формирующихся и неоднозначно понимаемых понятий, как правило, не имеющие дефиниции [33];
- предтермины — единицы, используемые в качестве терминов при появлении новых понятий и имеющие временный характер, чаще всего описательные обороты;
- квазитермины — закрепившиеся в речи предтермины, не отвечающие требованиям, предъявляемым к терминам, чаще всего представляющие собой описательные обороты [33].

Все вышеперечисленные единицы являются составляющими терминологии; в процессе формирования терминосистем они вытесняются терминами, так как не соответствуют оптимальным моделям в своей содержательной и формальной структуре.

Сформировавшаяся терминосистема состоит из терминов, поскольку она сознательно конструируется из языковых единиц в процессе формирования теоретических основ предметной области в текстах, написанных на ЯСЦ, который соответствует этой области или деятельности [51].

Такой сознательный характер конструирования терминосистемы приводит к тому, что она обладает двумя важными особенностями: структурированностью и полнотой. Структурированность терминосистемы — это способ ее организации с ярко выраженными взаимосвязями между входящими в нее единицами. Полнота терминосистемы — это заполненность всех мест системы понятий по меньшей мере одним обозначением (термином). В противоположность этому терминология бывает неполной, в ней могут быть лакуны, как относящиеся к отдельным элементам, так и к целым участкам системы понятий.

Среди признаков, характеризующих как терминологию, так и терминосистему, можно выделить их целостность и относительную устойчивость, т. е. способность сохранять основную часть терминов в процессе функционирования системы понятий той или иной области [90].

Системность терминологии всегда рассматривалась как одна из ее важнейших характеристик. Отмечалось также, что системность терминологии имеет двойственную основу: системность понятийная, логическая, вытекающая из системности понятий самой науки, и системность лингвистическая, обусловленная системностью выражающих эти понятия языковых единиц.

Некоторые авторы выделяют словообразовательную системность терминологии [21], другие видят системность в соотнесенности терминов как по форме, так и по содержанию, рассматривая терминосистему как соотнесенную с определенной областью знания, проблемой, темой и т. д. совокупность терминов, связанных друг с другом на понятийном, лексико-семантическом, словообразовательном и грамматическом уровнях [6; 10; 18; 29; 83].

В основание системности терминологии некоторыми исследователями кладется принцип классификации понятий, отношения между которыми принято делить на логические (родовидовые) и онтологические (партитивные) [54]. Отношения между понятиями подразделяются также на иерархические (видовые и партитивные отношения, задающие иерархию понятий) и неиерархические (все остальные типы онтологических и логических отношений, не задающих иерархии понятий).

В настоящее время на первый план выходит социальная роль и коммуникативная функция лингвистического знака, изучение терминологии в реальных условиях профессиональной коммуникации. Изучение же терминосистем вне конкретного словесного окружения, ситуации речи и жанра высказывания порождает проблему определения границ терминологического поля, усугубляющуюся многозначностью терминов, функционирующих в различных узких научных областях в разных значениях.

Именно информационная функция термина, его коммуникативная активность позволяют рассматривать терминосистему как основу организации специального текста, в противоположность изучению совокупности терминов той или иной научной области, рассматриваемых в отрыве от особенностей функционирования их в языке.

Исходя из всего вышесказанного, следует разграничивать понятия терминологии и терминосистемы. Терминология как совокупность лексических единиц определенной области знания, или фрагмент

области знания, является объектом последующего упорядочения. Терминология может подвергаться систематизации, анализу, при которых возможно выявление недостатков, а также методов их устранения, с последующей нормализацией терминологии. Результатом такой работы по нормализации терминологии в одном языке научной области является терминосистема. Таким образом, под терминосистемой понимается соотнесенная с определенной областью знания совокупность терминов, связанных друг с другом на понятийном, лексико-семантическом, словообразовательном и грамматическом уровнях [64].

Терминополе и терминосистема как лексико-семантическое единство. Терминосистема непосредственно связана с планом выражения, в основе же плана содержания лежит терминополе. Их комбинация дает «двустороннее знаковое образование, именуемое терминологической лексико-семантической системой» [68].

Следовательно, терминополе — это экстралингвистическая область, с которой термин соотносится как член системы, внутри которой обычно наблюдается определенная лингвистическая упорядоченность элементов. Являя собой неразрывное единство, терминополе и терминосистема имеют следующие общие характеристики:

- терминополе и, соответственно терминосистема являются иерархическими образованиями [8; 27; 114]. Основные понятия в плане содержания обусловливают появление базовых терминов в плане выражения, из которых впоследствии образуются производные термины, соотносящиеся с периферийными понятиями в плане содержания;
- с развитием конкретной области происходит изменение терминополя и, реагируя на эти процессы, терминосистема, в свою очередь, создает новые лексические единцы и удаляет устаревшие;
- терминополе отражает определенную систему понятий, терминосистема также имеет определенные границы и отличается целостностью, не допуская пополнения терминами, не соответствующими лингвистическим характеристикам конкретной предметной области [74];
- в пределах одного языка, как правило, терминополе и терминосистема развиваются синхронно, тогда как в разных языках

такой синхронности порой не наблюдается, так как процессы развития той или иной области деятельности в разных странах идут по-разному. Такое положение ставит во главу угла вопрос синхронизации и гармонизации терминосистем [49; 52];

- термин невозможно проанализировать вне контекста терминосистемы.

Таким образом, терминосистема может рассматриваться как организованная иерархия терминов, обслуживающих терминополе. Терминополе можно определить как экстралингвистическую область, с которой соотносится термин, обладающую определенной лингвистической упорядоченностью и иерархической структурой. Терминополе можно рассматривать как модель предметной области [64].

Унификация и упорядочение терминологии. Для устранения недостатков в переводе терминологии при межъязыковой коммуникации необходима ее унификация. Однако для обозначения понятий, связанных с приведением терминологии в определенную упорядоченную совокупность, используются следующие термины: унификация, упорядочение, стандартизация и гармонизация.

Одним из основных направлений прикладного терминоведения является унификация терминов и терминосистем, что включает в себя стандартизацию, упорядочение и гармонизацию на национальном и международном уровнях [52].

Приведение терминологии определенной предметной области в терминосистему проходит три этапа: упорядочение, унификацию и стандартизацию. Термин «упорядочение» рассматривается как самое общее из этих понятий. Упорядочение определяется как основная составляющая практической работы по унификации терминологии, связанной с приведением терминов к единообразию, единой форме или системе, поэтому в задачу исследователя в процессе создания упорядоченной терминологии входит образование системы понятий.

Первым этапом для достижения состояния упорядоченности терминологии является ее унификация, сложная и многоаспектная работа по приведению отраслевой терминологии по возможности в систему на всех необходимых уровнях (содержательном, логическом и лингвистическом). Унифицированная терминология может стать объектом стандартизации [35]. Целью унификации является обеспечение однозначного соответствия между системой понятий

и терминосистемой. Работа по унификации проводится на всех уровнях — содержательном, логическом и лингвистическом. При этом осуществляется как лингвистический анализ терминов и учет общих норм и закономерностей языка, так и учет специфических моментов, характерных для нормативных критериев оценки терминологии.

При унификации терминологии производится кодификация терминосистемы, то есть оформление ее в виде нормативного словаря. Кодификация терминосистемы возможна в двух формах: 1) рекомендации наиболее правильных с точки зрения терминоведения терминов, результатом которой является сборник рекомендуемых терминов; 2) стандартизации, результатом которой является государственный или отраслевой стандарт на термины и определения (ГОСТ). В качестве самостоятельного явления рассматривается гармонизация разноязычных терминов [42, с. 207–213].

Понятие и задачи стандартизации терминологии. В результате упорядочивания и унификации терминологии производится кодификация терминосистемы, и в случае когда любые отступления от точного однозначного употребления термина являются недопустимыми, такая кодификация принимает форму стандартизации терминов. Целью стандартизации является создание стандартов на термины и определения, которые являются обязательными в документации определенной специальной области. Терминологический стандарт представляет собой правовой документ: законодательное закрепление в нем употребления терминов вызвано необходимостью их однозначного понимания в различных областях действительности.

Задачами, ставящимися при стандартизации научной терминологии, являются:

- фиксация в стандартах на термины и определения современного уровня научного знания и технического развития;
- гармонизация (обеспечение сопоставимости) научной терминологии национального и международного уровней;
- обеспечение взаимосвязанного и согласованного развития лексических средств, используемых в информационных системах;
- выявление и устранение недостатков терминологической лексики, используемой в документации и литературе.

В качестве основных этапов работы по стандартизации терминологии можно выделить:

- проведение полной систематизации всех названий, включая во
типы употребления терминов в текстах и в разговорной речи
все синонимы, как стандартные, так и жаргонные, профессио
нально-диалектные. На этом этапе необходимо подготовить ис
черпывающие терминологические словари самых разных жан
ров;
- разработка четкой логико-понятийной модели терминоси
стемы, на основе которой происходит оценка и унификация ре
ально существующей терминологии;
- собственно стандартизация терминологии. Анализ логико
грамматической организации, деривационной способности, си
стемности и других важных характеристик позволит выбрат
из общего массива терминов термин, рекомендуемый к офици
альному употреблению в изданиях разного рода [23].

Определение гармонизации терминологии. Методы унифика
ции терминов используются и в случае межъязыкового упорядоче
ния, т. е. обеспечения сопоставимости терминологии национальног
и международного уровней, или гармонизации.

Наблюдаемое в настоящее время усиление международного со
трудничества в области науки, культуры и экономики требует уско
рения работы по гармонизации терминологий наиболее развитых на
циональных языков.

Перевод терминологии во многом зависит от уровня организации
терминосистемы. При межъязыковой коммуникации возникает не
обходимость гармонизировать термины соответствующей предмет
ной области знаний и сделать полученные результаты доступными
и, что не менее важно, понятными для специалистов. Неточности
перевода научных документов, лексическая некомпетентность пере
водчика, а также отсутствие фоновых знаний могут привести к не
корректности перевода и к последующим за этим ошибкам и нару
шениям в работе.

Гармонизация терминологии представляет собой межъязыковое
упорядочение терминологии, т. е. обеспечение сопоставимости терми
нологии национального и международного уровней [42, с. 210–211].

Под гармонизацией понимается вид терминологической деятель
ности, заключающейся в согласовании терминов на национальном
и международных уровнях [77].

Международные и отечественные органы стандартизации разработали и ввели в действие ряд основополагающих нормативных документов, регламентирующих создание систем терминологических стандартов и толковых словарей. Базовым понятием этих документов является понятие «гармонизация». Гармонизация понятий определяется как целенаправленная деятельность, позволяющая устранить (или снизить до приемлемого уровня) различия, относящиеся к разным понятийным системам, описывающим один и тот же объект стандартизации. Гармонизация понятий осуществляется не только в рамках систем понятий, выраженных разными языками, но и в рамках одного языка.

Специфика гармонизации терминов заключается в том, что важнейшим условием достижения эквивалентности является сохранение в единице перевода содержательной точности единиц исходного языка, обеспечение абсолютной идентичности понятий, выражаемых терминами исходного языка и языка перевода. Иными словами, если термины ИЯ и ПЯ кодируют понятие соответствующей научной области, а тождественность кодов отправителя и получателя является элементарным условием успешной коммуникации, то именно обеспечение тождественности означаемых терминами понятий представляет собой важнейшую задачу перевода специального текста. Выявление расхождений в системе понятий, выражаемых терминами ИЯ и ПЯ, — важный шаг на пути межъязыковой гармонизации терминосистем, обеспечивающей решение проблем перевода терминов [32].

Этапы гармонизации терминологии. Гармонизация терминологии предполагает следующие этапы:

- системное сопоставление национальных терминологий и терминосистем;
- составление сводной классификационной схемы понятий с учетом всех понятий, отраженных в сопоставляемых национальных терминологиях;
- выработка соглашения об установлении однозначного понимания и использования эквивалентных национальных терминов;
- интернационализация, предусматривающая взаимное заимствование в национальных языках терминов для заполнения лакун в национальных терминосистемах;

- фиксация международных решений по упорядочению семантики терминов;
- разработка многоязычных электронных терминологических баз данных, в которых будет накапливаться и храниться информация о лингвистических и логических особенностях терминов, их употреблении, многоязычных (переводных) эквивалентах, а также степени их упорядочения.

Нормативными требованиями, применяемыми к терминам и необходимыми при гармонизации многоязычной терминологии, являются:

- системность терминологии;
- независимость от контекста;
- краткость;
- однозначность;
- простота и понятность;
- достаточная степень внедрения в профессиональный язык;
- соизмеримость дефиниций с определяемым понятием;
- возможность временной замены термина на краткую дефиницию из нескольких слов, при невозможности найти однословный термин [60].

Таким образом, для достижения адекватного перевода, максимально соответствующего оригиналу при межъязыковой коммуникации в сфере науки, необходимо систематизировать и сопоставить терминосистемы исходного языка и языка перевода на основе сводной системы понятий. Создание такой системы необходимо для взаимной гармонизации терминов и выражающих их понятий, результатом которой должно быть создание нормативных переводных словарей и терминологических стандартов, обеспечивающих точный и однозначный перевод в сфере научной и научно-технической коммуникации.

1.2. СОВРЕМЕННЫЕ ТЕРМИНОЛОГИЧЕСКИЕ СИСТЕМЫ: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

Постоянное изменение терминосистем различных областей знаний требует разработки методов, позволяющих достаточно быстро и объективно извлекать те лексические единицы, которые характе-

ризуют постоянно развивающиеся терминосистемы и необходимы для решения задач информационного поиска, машинного перевода, терминоведения, прикладной и общей лексикографии.

Использование информационных технологий в лексикографии, т. е. компьютеризация лексикографии и условия нового информационного пространства обусловливают необходимость разработки базовых принципов построения автоматизированных словарей в целом и лексикографических систем. Такими принципами являются [ср. 9; 11]:

- модульность [5; 13; 65] — этот принцип определяет необходимость формирования словарных статей как относительно независимых объектов, при этом статьи организуюися в системы, «которые в дальнейшем могут рассматриваться как подсистемы некой конкретной системы обработки данных» [5, с. 107];
- динамичность [16] — этот принцип подразумевает возможность оперативно и без существенных трудностей корректировать номенклатуру словаря, состав и объем информации в словарной статье, а также включать новые программы и файлы в уже созданную систему [55];
- гибкость [20] — определяет способность уже созданных баз данных и словарей удовлетворять новым требованиям без глобальной перестройки их организации, что подразумевает возможность информационного поиска и выборки данных, а также обратимость словаря [15; 16];
- сбалансированность — определяет системность комплектования баз данных и словарей, исключает произвольность и стихийность в отборе информации и подразумевает тщательный отбор материала, структурность базы, полноту представляемой информации [22, с. 13];
- дружественность — обеспечивает наиболее эффективный характер взаимодействия между системой и пользователем [36, с. 115]; она выражается в использовании графических изображений, разного вида меню, в звуковом сопровождении, т. е. в средствах, делающих работу пользователя максимально удобной. Такой принцип взаимодействия получил название «дружественного интерфейса», а также «интуитивно-понятного интерфейса» [70, с. 16].

Создание словарей на основе информационных технологий и автоматизированных лингвистических ресурсов может производиться различными способами [64]:

1) используя средства системы управления базой данных. Использование специальных программных средств позволяет создавать новые словари, вносить в них изменения, добавлять новую информацию и новые словарные статьи; таким же образом можно удалять или модифицировать словарные статьи в уже существующих словарях;
2) компилируя словник из исходного текста — использование специального пакета прикладных программ; в подобный пакет включается специальный компилятор, который на основе исходного словника создает базу данных; при обнаружении ошибок в исходном тексте компилятор выдает соответствующий отчет [61];
3) используя специальные приложения, в которых предусмотрены операции создания словарей, внесения описаний, добавления, удаления или модификации слов;
4) используя имеющиеся лексикографические ресурсы при пополнении системы новыми языками, что позволяет достичь существенной экономии трудозатрат на разработку словарных баз при сохранении поисковых возможностей системы [69].

Автоматизированный словарь представляет собой особый лексикографический объект, обладающий рядом специфических структурных особенностей [11; 73]. Под автоматизированными словарями будем понимать электронные словари, описывающие конкретную терминологию, энциклопедические, терминологические, толковые словари на машинных носителях, а также автоматические словари систем компьютерной переработки информации, включая автоматические словари систем машинного перевода [11; 14].

Автоматизированные и автоматические словари — словари систем переработки данных, ориентированы на использование программ автоматического анализа и переработки текстов на естественном языке. В современном варианте для таких словарей создается специализированная система их создания и ведения, обеспечивающая удобную работу лексикографа с информацией, которая может быть представлена в формальном и/или закодированном виде.

Так, принято различать:

- автоматизированные словари, предназначенные для конечного пользователя — человека [ср. 10; 27];
- автоматические словари – автоматизированные словари, предназначенные для программной обработки текста [ср. 27; 50].

Термин «автоматический словарь» относится именно к словарю, используемому системой автоматической переработки текста и не предназначенному для использования человеком [13, с. 67]. В такой формулировке под автоматическим словарем понимается автоматический переводной словарь (АПС), предназначенный в основном для обработки научно-технических текстов [58]. Автоматические словари, являясь упорядоченным массивом лингвистической информации, регистрируют и хранят лексические единицы (слова и словосочетания) с их морфологическими, синтаксическими и семантическими характеристиками, необходимыми для автоматического синтаксического анализа (парсинга), автоматического синтеза и перевода текста.

Автоматизированные и автоматические словари — словари систем переработки данных ориентированы на использование программ автоматического анализа и переработки текстов на естественном языке. В современном варианте для таких словарей создается специализированная система их создания и ведения, обеспечивающая удобную работу лексикографа с информацией, которая может быть представлена в формальном и/или закодированном виде.

Автоматизированные словари существенно отличаются от словарей традиционных, во-первых, способом хранения, во-вторых, строгой формализованностью записи информации [4, с. 46]. Фундаментальные «бумажные» словари — неизбежно словари устаревшие или устаревающие, так как они не обладают динамикой [69]. А для программных продуктов, таких как автоматические словари систем автоматической переработки текстов, характерны частая смена версий, наличие постоянной обратной связи с пользователями, динамичность существования, актуальность и открытость относительно каждого временного среза. Кроме того, автоматизированные словари, в отличие от традиционных, сочетают большой объем информации с удобством использования благодаря автоматизации механизмов поиска по соответствующему запросу [64].

Автоматизированный словарь рассматривается как база данных, в которой каждая статья представлена как текст, расписанный по зонам словарной статьи, при этом каждая зона имеет свое уникальное имя [57]. Такой словарь должен иметь жесткую организацию и высокий уровень формализации представления данных [5; 15]. Соответственно, автоматизированный словарь включает словарные файлы, индексы к ним, описание микро- и макроструктуры словаря и систему программ, обеспечивающую создание этой конструкции, ее поддержание в рабочем состоянии и обращение к ней за справками из других программ [5, с. 124].

Создание автоматизированного словаря представляет собой комплексный поэтапный процесс, включающий [ср. 11]:

1) создание массива данных, содержащего специальные тексты конкретной предметной области, подобный массив может быть организован как проблемно-ориентированный корпус текстов с соответствующей разметкой;
2) создание глоссария на основе статистического исследования массива;
3) установление степени терминологичности лексических единиц, зафиксированных в текстах предметной области;
4) описание терминополя и его формализацию в виде графической структуры (дерева), что требуется как для выявления структуры терминосистемы, так и для определения макроструктуры и состава словаря;
5) определение параметров оптимальной организации словарной статьи в зависимости от назначения словаря, его возможных функций и потенциальных пользователей;
6) выбор оптимальной структуры системы управления базой данных, в которую включается словарь;
7) непосредственное наполнение словаря.

Таким образом, автоматизированный словарь является оптимальным средством фиксации терминосистемы, находящейся в процессе становления, и основным блоком автоматизированного рабочего места (АРМ) специалиста. Еще раз подчеркнем, что создание автоматизированного словаря представляет собой поэтапный процесс, при этом реализация последних этапов невозможна без соответствующей технической поддержки. В качестве такой поддержки могут использоваться стандартные базы типа Microsoft Access или специально создаваемые средства типа систем ABBY Lingvo Content, Polyglossum и им подобные.

Как и любые лексикографические источники [10], автоматизированные и автоматические словари могут быть переводными (рассчитанными на два или более языка) или одноязычными (толковыми, словарями синонимов и т. д.), при этом принципы работы словарей различных типов практически не отличаются между собой.

Так, например, терминологические базы данных (ТБД) представляют собой автоматизированные хранилища терминов, в которых термины снабжены дополнительной информацией не только собственно лингвистического (морфологические и семантические характеристики, сочетаемость, частотность, принадлежность к конкретному языку для специальных целей и т. п.), но и экстралингвистического (нормативность, стандартизованность и т. п.) характера. «В зависимости от цели создания ТБД их можно разделить на две группы: ориентированные на обеспечение работ по переводу научно-технической литературы и документации и предназначенные для обеспечения информацией о стандартизованной и рекомендованной терминологии» [52, с. 284].

Следует отметить, что системы, работающие с терминологией, существуют достаточно давно. Еще в 70-х гг. XX в. крупные компании и правительственные организации создавали машинные языковые фонды: параллельно с экономическим и техническим ростом постоянно появлялась новая терминология, и такие фонды предназначались для унификации терминов, используемых в определенных типах текстов и при переводе. Один из наиболее крупных фондов, работа над которым, к сожалению, завершена, – TEAM, разработанный компанией Siemens для работы с европейскими языками, в частности с русским, включает в себя около 700 000 лексических единиц (ЛЕ) из различных тематических областей (естественные науки, бизнес, техника и т. п.), соответственно сгруппированных [107, с. 291]. Материалы этого фонда используются и в настоящее время при создании специализированных словарей.

Современные системы создания и ведения ТБД предназначены не только для крупных компаний и столь масштабных проектов, как TEAM, но и для переводчика-профессионала (или группы перевод-

чиков), работающего с терминологией на персональном компьютере: объемы баз данных определяются пользователем и назначением системы.

Особый интерес представляют базы данных, создаваемые в университете Пенсильвании консорциумом LDC (Linguistic Data Consortium) на основе кооперации с другими университетами США и Европы. В настоящее время этой фирмой создано несколько сотен баз данных, на основе которых построены различные виды словарей.

Так, например, корпус CELEX [47] был создан в результате совместных исследований университета Неймегена, Института нидерландской лексикологии в Лейдене, Института психолингвистики им. Макса Планка в Неймегене и Института исследований восприятия в Эйндховене. Этот корпус содержит лексические базы данных для трех языков: английского (Версия 2.5), нидерландского (Версия 3.1) и немецкого (Версия 2.0), представленные в виде ASCII файлов. На основе обработки и форматирования этих материалов в консорциуме был создан специальный компакт-диск, распространяемый при условии предварительной оплаты.

Для ЛЕ каждого языка этот компакт-диск содержит детальную информацию относительно:

- орфографии (варианты правописания, расстановка переносов);
- фонологии (фонетические транскрипции, варианты произношения, слоговая структура, главное ударение);
- морфологии (деривационная и композиционная структура, парадигмы словоизменения);
- синтаксиса (лексический класс, субкатегории, определяемые лексическим классом слова, актантные структуры);
- частоты слова (накопленная частота словоформ и леммы, полученные на основе современных и репрезентативных корпусов текстов).

Второй выпуск CELEX содержит расширенный вариант немецкой лексической базы данных (2.5), включая приблизительно 1,000 новых словарных статей, пересмотренные характеристики морфологического анализа, актантные структуры глаголов, коды парадигм словоизменения и лексикона, представленный как корпус текстов. Для немецкого языка общее число включенных в базу данных лемм составляет 51 728, а количество флективных форм равно 365 530.

Для каждого из языков, т. е. английского, немецкого и нидерландского, последний вариант содержит детальную информацию относительно:

- орфографии (варианты правописания, расстановка переносов);
- фонологии (фонетические транскрипции, варианты произношения, структура слога, главное ударение);
- морфологии (деривационная и композиционная структура, парадигмы словоизменения);
- синтаксиса (лексический класс, специфическая субкатегоризация лексического класса, актантные структуры);
- частоты (накопленная частота словоупотребления и леммы, на основе современных и репрезентативных корпусов текстов) словоформ и лемм.

Уникальные идентификационные номера позволяют соединять информацию из различных файлов.

Все вышеперечисленные системы, несмотря на различия в интерфейсах, оснащены вполне типовым набором основных функций. Практически все из них поддерживают большой набор языков, т. е. дают возможность создания как одноязычной, двуязычной, так и мультиязыковой (для более двух языков) ТБД, в зависимости от требований пользователя и технических возможностей компьютера или компьютерной сети.

Кроме того, все подобные системы позволяют:

- вводить новые лексические единицы и составлять на них развернутые словарные статьи с указанием морфологических и грамматических характеристик, тематической области, примеров применения и т. д.;
- импортировать информацию из существующих словарей, глоссариев и т. д.;
- группировать терминологию по тематическим областям, рабочим проектам, другим критериям;
- создавать перекрестные ссылки при вводе терминологии (во многих системах эта опция автоматическая, например, в Termstar);
- осуществлять быстрый индексный и последовательный поиск терминов.

Кроме того, системы создания и ведения терминологических ба
не только позволяют устанавливать необходимый для конкретной
задачи терминологический стандарт и придерживаться его, но также
параллельно работать с несколькими базами данных и словарями
а для группы пользователей — одновременно пользоваться одной ба
зой данных.

Дополнительное преимущество систем создания и ведения ТБД
состоит в том, что они могут заменить используемые переводчиками
специализированные глоссарии (списки специфических слов, слово
сочетаний и оборотов, встречающихся в определенном тексте/кон
тексте, и их перевод), поскольку обычный глоссарий является про
стым списком и не обладает перечисленными выше возможностями
терминологических баз данных. К тому же, как уже отмечалось, со
временные системы создания и ведения ТБД оснащены утилитой,
позволяющей импортировать в базы данных ранее составленны
переводчиком глоссарии (таким образом, нет необходимости занов
создавать лексическую базу), или глоссарии, доступные через сет
Интернет и на компакт-дисках (например, глоссарии Microsoft, кото
рые представляют собой списки команд, традиционных сообщени
системы и другой терминологии поддержания коммуникации с опе
рационной системой на исходном языке и языке перевода).

Давно известные терминологические базы данных (например
TEAM и др.) являются прототипом современных лексикографиче
ских систем. Рассмотрение различных лексикографических систем
позволяет выявить систему параметров, по отношению к которой
могут описываться ЛЕ в базах данных. Эти параметры можно раз
делить:

- на служебные;
- лексико-грамматические;
- комбинаторные;
- семантические;
- энциклопедические.

К служебным параметрам описания относится информация:

- о времени введения ЛЕ в словарную статью базы данных;
- источнике, в котором слово было зафиксировано впервые;
- языке, на котором слово было зафиксировано (в случае много-
  язычной базы данных);

- времени внесения последних изменений в словарную статью;
- имени или идентификационном коде лексикографа, ответ-
  ственного за словарную статью или работавшего с ней послед-
  ним.

К лексико-грамматическим параметрам описания относится ин-
формация:

- о транскрипции ЛЕ, сопровождаемой записью стандартного
  (нормативного) варианта произношения;
- частеречной принадлежности лексической единицы; при кон-
  версионной омонимии слова этот параметр является источни-
  ком ветвления описания в соответствии с типом конверсион-
  ной омонимии. Количество и вид параметров зависят от части
  речи и типа языка;
- морфологической структуре слова, особенностям формообра-
  зования или типовой парадигме;
- валентностной структуре и/или управлении.

К семантическим параметрам описания относится информация:

- о предметной области в соответствии с принятой классифика-
  цией. В качестве основного классификатора сегодня активно
  используется универсальная классификация Леноха, включа-
  ющая 48 предметных областей, которые делятся на 785 групп
  и 2600 терминальных узлов [см. 123];
- значении лексической единицы, записываемом либо в виде
  текста – толкования или дефиниции, либо в формализованном
  виде. Формализованное представление может основываться
  на наборе лексических функций, семантических примитивов,
  системе отношений в соответствующей онтологии [111]. Выбор
  формализма для описания значения определяется потенциаль-
  ными возможностями его использования;
- значении ЛЕ, описываемом как узел семантической сети через
  связь с иерархически или ассоциативно связанными элемен-
  тами, синонимами и антонимами;
- значении ЛЕ, описываемом системой его вхождений в другие
  словарные статьи;
- переводе на другие языки, включенные в терминологическую
  базу или лексикографическое описание.

К энциклопедическим параметрам описания относится инфор
мация:

- о визуальном представлении соответствующего референта
если оно может быть получено либо его формат позволяет тако
включение. В зависимости от типа референта это может быт
статичный рисунок/фотография или видеозапись действия
движения или процесса;
- важных с точки зрения эксперта – специалиста в конкретно
области знаний — сведений об онтологической сущности рефе
рента, истории его открытия или создания и т. п. Эта задача от
носится к тем, что принято называть a lot-easier-said-that-done
task — задача, которую легче поставить, чем сделать [ср. 56].

Естественно, в предлагаемом наборе параметров можно выделит
обязательные и факультативные. Кроме того, сам набор может до
полняться новыми параметрами в зависимости от типа планируе
мого словаря и его предметной ориентации [11].

Базы данных терминологии должны включать иерархически ор
ганизованные словарные статьи, в которых каждая «ветвь» описани
ориентирована на принадлежность к определенной предметной об
ласти или подобласти. Соответственно, подобная классификаци
в виде структуры предметной области должна предшествовать тер
минологическому наполнению базы данных.

Дружественность и многофункциональность интерфейса для ра
боты лексикографа при создании/ведении автоматического словаря
определяет эффективность этой работы и возможность проверки
принимаемых решений.

Отбор терминов, фиксация и гармонизация терминологии пред
ставляют собой особую задачу, решение которой предшествует соз
данию словаря.

Как правило, гармонизация терминологии предполагает установление
не только соответствия терминов, но и понятий. Соответственно, процесс
гармонизации включает сравнение систем понятий, что предполагает
установление отношений между понятиями, фиксацию количества поня
тий, глубины структуры, устранение дублирования и т. п., что приводит
к построению новой гармонизированной системы понятий.

В основе решения этой задачи лежит структурирование термино
поля на базе выявления основных узлов семантической сети и связей

между ними: гиперо-гипонимических, меронимических, ассоциатив
ных и т. п. По отношению к этой структуре и происходит фиксация
терминосистемы конкретного языка для специальных целей или тер
миносистем разных языков. Наличие зафиксированной в виде графа
структуры терминосистемы позволяет соотносить терминологиче
ские единицы разных языков с одним и тем же узлом терминополя и,
соответственно между собой.

Задача выявления набора терминологических словосочетаний
также не имеет собственно формального решения, но для ее решения
могут привлекаться методы корпусной лингвистики.

Одним из сценариев терминологической работы является пе
реход от перевода и создания документов к лексикографическому
описанию единиц контрастируемых языков. В этом случае тер
минологическая работа включает опознание терминов в тексте
на входном языке, использование существующих глоссариев, ТБД
и, возможно, текстов соответствующей предметной области. Если
термин после применения всех этих источников не идентифици
рован, то он должен быть описан и введен в базу данных. Такая
терминологическая работа предполагает необходимость консуль
таций со специалистами в конкретной предметной области, тер
минологами и переводчиками. Кроме того, терминологическая ра
бота предполагает создание специальных средств, интегрирован
ных в среду перевода.

Проведение переводческой работы в специализированных цен
трах и с привлечением ТБД предполагает следующую последова
тельность операций [124, с. 19–20]:

1. Документ, предназначенный для перевода, передается для ана
лиза терминологу, который:
   - идентифицирует термины;
   - проверяет существование перевода терминов;
   - если термин уже введен в ТБД, вводит в его словарную статью
   идентификационный код документа;
   - при отсутствии эквивалентного перевода, создает его;
   - при высокой сложности текста он передает эксперту (резуль
   таты работы эксперта вводятся в ТБД).

2. Документ с отмеченной терминологией и информацией о кор
ректных переводах передается переводчику:

- Переводчик в процессе перевода может консультироваться с терминологом и вносить предложения по переводу терминов
- При необходимости терминолог модифицирует статьи в базе данных.

Подобная последовательность процедур позволяет проверять не только корректность выполненных переводов, но и полноту и корректность словарных статей в базе терминов. Важно отметить, что в современных центрах перевода первым с текстом работает не переводчик, а терминолог.

При создании терминологической базы данных постоянной частью работы является проверка ее корректности [ср. 124, с. 36–37], которая включает регулярный анализ соблюдения формальных требований и проверку правильности заполнения словарных статей.

Необходимость соблюдения формальных требований предполагает:

- проверку отсутствия дублирования словарных статей в базе;
- проверку целостности базы, т. е. введение информации в обязательные поля и отсутствие двойного заполнения;
- проверку полноты базы, что предполагает подтверждение функциональности перекрестных ссылок, правильность формы терминов, описание падежных систем, введение кодов языка в соответствии с выбранным стандартом;
- проверку орфографической правильности;
- проверку полноты и корректности грамматической информации, приписываемой каждому термину;
- проверку точности классификации, т. е. проверка отнесения словарной статьи к той предметной области, которая соответствует классификации областей.

Корректность заполнения словарных статей с точки зрения полноты и содержательности включает:

- проверку полноты системы понятий, выделенных в описываемом терминополе;
- проверку полноты фиксации терминосистемы, отражающей терминополе;
- проверку корректности формирования синонимических рядов;
- проверку точности дефиниций относительно контекста и их понятность для пользователя базы;

- проверку правильности помет об употреблении, стилистических помет;
- проверку правильности кодов, определяющих степень надежности термина.

Комплексное и регулярное выполнение терминологами и лексикологами этих проверок поддерживает базу данных в актуальном состоянии [11].

Одной из сложных проблем работы со знаниями является выбор адекватной модели их представления. В настоящее время разработано множество подходов к представлению знаний, в основе которых используются такие модели представления знаний, как формально-логические, продукционные, семантические сети, фреймовые модели, онтологии. Эти модели представления знаний имеют ограниченные области решаемых задач в силу присущих им свойств и ограничений [79; 71].

В распоряжении филологов подобный лингвистический энциклопедический словарь (ЛЭС) существует и, безусловно, является авторитетным изданием. Однако непосредственное использование его для получения списка терминов для переводного словаря представляется нецелесообразным на первом этапе выделения и анализа современной терминологии.

1.3. ОНТОЛОГИЯ КАК НОВАЯ ФОРМА СРАВНЕНИЯ ЛЕКСИКОГРАФИЧЕСКИХ СИСТЕМ

Современная прикладная лингвистика и многочисленные направления фундаментальных и прикладных исследований, использующих ее результаты, нуждаются в специальных методах и инструментальных средствах для уточнения и описания систем понятий различных предметных областей, отражаемых языками для специальных целей, для создания наборов метаданных или для разработки удобных и качественных поисковых устройств.

Основной тенденцией современных систем переработки информации (информационных систем) является интеграция знаний, извлекаемых из неформализованных текстов на естественном языке, и знаний формализованных (формальных), особым образом фик-

сируемых в банках и базах знаний. Знания, извлекаемые из текста более обширны по своей форме, и способы их представления вполн привычны любому пользователю информационных систем. Фор мальные знания представляют собой информацию, представленну в формализованной форме, и предназначены для использовани программами переработки текстов на естественном языке.

Следовательно, в основу различения автоматизированных и ав томатических словарей может быть положено использование ко кретного вида знаний: автоматизированные словари (электронны лексиконы) включают неформальные знания, а автоматические - знания формализованные.

Термин «онтология» для описания процесса концептуализаци прежде всего относится к проблеме создания открытого для общег и многократного использования информационного ресурса, форми руемого как словарь понятий (термин в этом значении был введе Томом Грубером в 1991 г. [104]). При современном подходе онтоло гия представляет собой базу знаний, хранящую информацию о поня тиях, существующих в мире или предметной области, их свойства и о том, как они связаны друг с другом.

Онтология отличается от тезауруса тем, что содержит только не зависящую от языка информацию и множество семантических от ношений, кроме того, она содержит таксономические отношения Тем самым задача построения онтологии представляет собой задачу создания некоторой модели мира, необходимой для смысловой пере работки текста. Онтология должна задавать понятия для представле ния значений слова в лексиконе. Соответственно, можно считать, что тезаурус представляет собой плоскостную двумерную модель терми носистемы, а онтология — трехмерную модель предметной области, ее содержательного знания.

Предполагается, что существуют два основных пути получения содержательного знания о предметной области:

- извлечение знаний экспертов соответствующей предметной об ласти;
- извлечение онтологических знаний из текстов естественного языка, описывающих эту предметную область [25, с. 25].

В основе автоматизации работ по созданию онтологий может ле жать использование энциклопедических словарей как выверенного

Рис. 1. Некоторые понятия области структурирования знаний

способа отражения информации, при этом заглавия словарных статей подобного словаря могут использоваться в качестве ключевых слов. Специальные поднаборы этих терминов могут использоваться как названия подобластей соответствующей предметной области [25, с. 25].

Рассмотрим особенности представления информации и знаний в форме онтологии на примере графического представления связи понятий из области структурирования знания (рис. 1).

Онтологии и классификации являются способами структурирования знаний. Согласно онтологии, приведенной на рисунке 1, модели и системы классификации можно различать следующим образом: цель модели состоит в том, чтобы дать упрощенное представление знаний о явлениях, в то время как целью классификационной системы является последовательное деление явлений на классы, которые формируют основу для упорядочивания «вещей» [115, с. 280].

Особый интерес для лингвистического исследования представляют терминологические онтологии, т. е. онтологии, которые основываются на методе терминологического анализа, используя характеристики и основания деления. Терминологическая онтология является проблемно-ориентированной. Как правило, термин «терминологическая онтология» используется как синоним термина «система

понятий», который обычно используется в работе с терминологией. Грубер [104] описывает онтологию следующим способом: «Онтология уточняет словарь, с помощью которого делаются утверждения, которые могут быть входом или выходом для агентов знания (типа программного продукта)».

Лингвистические процессоры, активно разрабатываемые для решения задач автоматической обработки текста и извлечения знаний и ориентированные на использование онтологий, должны учитывать:

- многовариантность результатов автоматического синтаксического анализа предложения, связанную с лексической и синтаксической омонимией, снятие которой часто вызывает затруднения даже при «ручном» анализе;
- синтаксическую и семантическую многозначность структур предложения в целом и структур именных и глагольных групп, составляющих функциональные компоненты предложения;
- особенности реализации процедур трансфера с учетом сопоставительного анализа структурных характеристик исходного языка и языка перевода (ср. [45, с. 120]).

Сегодня системы управления терминологией (менеджмента терминологии) поддерживают функции обнаружения, оценки и использования терминологических единиц в рамках конкретной предметной области или (у́же) конкретной организации. Эти системы поддерживают деятельность авторов, экспертов, редакторов и переводчиков (см, например, систему Acrolinx IQ terminology manager — http://www.acrolinx.com/acrolinx-iq-terminology_en.html).

При использовании подобных систем предполагается [11], что в тексте описывается новый продукт, процесс или устройство, для которых в исходном языке может не быть принятых номинаций. В этом случае с помощью глоссариев системы автор текста имеет возможность:

- проверить статус выбираемого им термина, который может быть предлагаемым, не рекомендуемым, принятым, предпочтительным, в некоторых случаях слово имеет статус «не-термин»;
- создать новый термин на основе имеющихся терминов в базе онтологии или тезауруса.

В задачу эксперта в конкретной предметной области при использовании такой системы входит оценка правильности выбранной

номинации относительно предметной области и языка для специальных целей, а также связанной с номинацией информации. Технический редактор проверяет последовательность использования терминологии во всем проекте в целом и оценивает лингвистическую правильность используемых терминов.

В задачу переводчика при использовании такой системы входит:

- установление существующей в исходном языке терминологии;
- исследование значения выделенных терминов в языке перевода;
- создание новых терминов в языке перевода.

Редактор перевода также оценивает последовательность использования переводной терминологии во всем проекте в целом.

2. ПРИНЦИПЫ И МЕТОДЫ ПРОЕКТИРОВАНИЯ ЛЕКСИКОГРАФИЧЕСКОЙ БАЗЫ ДАННЫХ НА ОСНОВЕ АНАЛИЗА ПРЕДМЕТНО-ОРИЕНТИРОВАННОГО КОРПУСА ТЕКСТОВ

2.1. ИССЛЕДОВАТЕЛЬСКИЙ КОРПУС ТЕКСТОВ ДЛЯ ЗАДАЧ ПЕРЕВОДНОЙ ЛЕКСИКОГРАФИИ

Развитие науки свидетельствует о возникновении новых направлений на стыках традиционных отраслей знаний и, следовательно, настоятельную потребность создания узкоспециализированных словарей, как автоматизированных, так и «бумажных».

Современный подход к созданию словарей предполагает формирование и использование параллельного корпуса реальных текстов, который может рассматриваться как база данных для решения не только исследовательских задач, но и практических лексикографических задач. Корпусы письменных текстов, как правило, включают сами тексты, а также разметку текстов с точки зрения формата и предложений по результатам парсинга, позволяющую установить принадлежность лексических единиц к конкретным частям речи. Эти тексты могут служить для создания конкордансов, словарей слов и словосочетаний в случае одноязычного корпуса, а также для создания многоязычных лексиконов и многоязычных конкордансов в случае корпуса параллельных массивов.

При создании предметно-ориентированного переводного словаря на основе корпуса текстов необходимо:

- определить принципы формирования выборочной совокупности для создания исследовательского параллельного корпуса текстов и ее необходимый и достаточный объем;

- установить требования к разметке текстов в корпусе и получить базовую лингвистическую информацию;
- определить необходимые для работы средства информационных технологий, включая системы машинного перевода, переводческой памяти, средства выравнивания параллельных текстов, редакторы разметки (тэггирования), средства формального извлечения терминов из текста и т. д.

Средства информационных технологий могут быть ориентированы на выбор и обработку терминов и/или понятий, на работу с конкретной языковой парой, многоязычными или одноязычными ресурсами.

Задачи консолидации, слияния и обобщения национальных терминологических ресурсов, гармонизации и стандартизации терминологии рассматриваются сегодня как часть процесса глобализации и, в частности, создания единой Европы.

Создание исследовательского корпуса текстов для решения лексикографических задач основывается на следующих принципах формирования:

1) установление подобластей предметной области на основе традиционной классификации знаний. В качестве такой классификации могут использоваться библиотечная десятичная классификация, предметные классификации информационно-поисковых систем, универсальная предметная классификация Леноха, тезаурусы; классификация Леноха включает 48 предметных областей, которые делятся на 785 групп и 2600 терминальных узлов [см. 123];
2) отбор текстов разных жанров и сопоставимого объема;
3) разметка текстов, включающая фиксацию словосочетаний — именных групп и глаголов с послелогами.

Одним из представительных корпусов данных, предназначенных для решений как исследовательских, так и практических задач, является канадский корпус Hansard. Этот корпус текстов состоит из параллельных текстов на английском языке и канадском варианте французского языка, извлеченных из официальных документов заседаний Канадского Парламента. Этот корпус является проблемно-ориентированным, поскольку содержание материалов ограничено проблемами законодательства, он охватывает широкий спектр тем,

а диапазон речевых жанров включает оцифрованные и аудиотексты парламентских дебатов, официальную переписку, наряду с законодательными предложениями и подготовленными выступлениями.

Коллекция текстов охватывает промежуток времени от середины 1970-х до 1988 г. и состоит из двух частей – массивов текстов, представленных фирмами IBM и Bell, соответственно. Каждому английскому предложению поставлено в соответствие предложение французское, минимальная разметка включает разбивку по предложениям и абзацам. При этом тексты, переданные фирмой IBM, представлены как последовательность параллельных предложений (почти 2.87 миллионов параллельных пар предложений), данные фирмы Bellcore представлены как последовательность параграфов. Корпус Hansard является примером параллельного корпуса текстов, в котором принцип параллельности выдержан полностью, однако создание подобного лингвистического продукта возможно только при государственной поддержке.

Создание базы полнотекстовых данных предполагает обеспечение хранения, модификации и поиска текстов произведений художественной и научной литературы на разных языках с формированием массивов параллельных и псевдопараллельных текстов. Эта база может непосредственно использоваться в учебном процессе для анализа конкретных лингвистических и литературоведческих фактов, проведения сравнительного стилистического анализа, изучения особенностей авторского стиля и т. д. Кроме того, подобная база является важным источником сведений для создания словарей разного состава и назначения.

Использование корпуса параллельных текстов в двуязычной лексикографии позволяет не только максимально автоматизировать отбор терминологических словосочетаний, но также служит:

- для обогащения набора словарных статей за счет выбора свободных словосочетаний, используемых в исходных текстах, что чрезвычайно важно для тех, кто переводит на язык, не являющийся родным;
- уточнения употребительности конкретных словосочетаний в текстах определенной предметной области;
- верификации значений лексических единиц, зафиксированных в двуязычных словарях, особенно в том, что касается идиом и терминологических выражений;

- выделения устойчивых словосочетаний и идиом, которые целесообразно вводить в словарь конкретной отрасли знаний.

На основе полнотекстовых баз параллельных выровненных текстов возможно выделение устойчивых пар слов типа «исходное слово-перевод», однако применение статистических процедур, как правило, допускает соответствие слов, но не словосочетаний. Это достаточно жесткое ограничение для выбора потенциальных компонентов словаря может быть ослаблено, если параллельные тексты предварительно проходят процедуру парсинга или подвергаются ручной разметке, что позволяет соотносить не отдельные слова, а фрагменты предложения.

Создание корпуса текстов для решения задач извлечения и исследования терминологии определяет требования к процедурам и условиям отбора материала [86, с. 149]. Дело в том, что особую лексикографическую проблему представляют словарные системы, предназначенные для перевода текстов с глобального английского языка. Этот язык является сегодня не только языком международного общения, но и явно испытывает активное влияние родных языков тех, кто на нем говорит и пишет [41]. Соответственно, можно в глобальном английском языке различать его «варианты»: германский, славянский, романский, китайский, японский и т. д. Эти варианты в терминологическом аспекте в рамках одного языка для профессиональных целей различаются структурой номинации референтов, особенно в случае расчлененной номинации.

Важным источником для исследования особенностей терминологии в глобальном английском языке являются материалы международных конференций, поскольку они:

- жестко ориентированы на конкретную предметную область или проблему в рамках подобласти;
- отражают современные тенденции развития науки и техники и включают новые терминологические единицы, номинирующие эти тенденции, новые объекты и методы исследования;
- не подвергаются редактированию носителями английского языка и отражают весь спектр особенностей разных национальных вариантов глобального английского языка;
- единообразны по структуре и объему отдельных текстов.

Структурно материалы международных конференций включают следующие компоненты:

- название;
- информация об авторе/авторах;
- ключевые слова;
- аннотация;
- введение*;
- собственно текст;
- заключение*;
- список использованной литературы*.

Компоненты, помеченные *, являются факультативными, обязательность их формального выделения определяется принятой для конкретных материалов структурой или традицией регулярно проводимых конференций.

Решение задачи лексикографирования терминологии предметной области должно опираться на исследование всего текста, а не его частей.

В то же время следует учитывать, что ключевые слова, выделяемые авторами каждой статьи или доклада, не дают репрезентативной базы терминов, поскольку их выбор опирается на интуитивное представление авторов текстов, а оно может не совпадать у разных специалистов и носителей разных языков.

Современная электронная лексикография предполагает использование предметно-ориентированных корпусов текстов, с помощью которых значительно расширяются возможности сбора и обработки данных о существующем составе и функционировании терминов заданной предметной области. Преимущества использования корпусов текстов в лексикографии обеспечиваются объемом корпуса, гарантирующим типичность данных; возможностью выявления нежесткой лексической сочетаемости исследуемых терминологических единиц и многократностью использования корпуса.

Следует учитывать, что лексикографическая работа даже с использованием возможностей информационных технологий остается работой творческой и не может быть полностью автоматизирована. В то же время существуют возможности подготовки массивов текстов для автоматизации лексикографического анализа. Идеальным источником материала являются корпусы параллельных текстов,

построенные на основе материалов узкой предметной области (статей, монографий, материалов конференций и их переводов на другой язык). Такой корпус должен быть выровнен по предложениям, что позволяет выявлять и анализировать термины и их переводы, оценивать стандартизованность и единство переводов, распространенность конкретных вариантов. Одним из вариантов создания материала для последующего лексикографического анализа является формирование особых корпусов текстов, включающих параллельное представление исходных текстов, их машинных переводов и отредактированных переводов, согласованных с экспертами в конкретной области знаний [12].

Важно отметить, что качество и потенциал такого корпуса в большой степени зависит от сотрудничества с экспертами при отборе исходного материала и редактировании переводов. Однако этот тип корпуса имеет свои ограничения: не по всем предметным областям существует достаточное количество доступных для использования параллельных текстов. Особенно их недостаток характерен для новых областей знания и стремительно развивающихся технологий, а также для тех отраслей, в которых обмен технической информацией с переводом на разные языки блокируется политикой безопасности. Кроме того, многие исследователи предпочитают работать с сопоставительными или псевдопараллельными корпусами текстов (comparable corpora), дающими возможность наблюдать живую речь, а не с переводами, отражающими влияние исходного текста [110, с. 18; 96, с. 211]. Эти ограничения были выявлены и в работе по созданию корпуса текстов по предметной области «авиационное газотурбинное двигателестроение». Недостаточное количество параллельных текстов в исследуемой области объясняется слабой вовлеченностью России в процесс обмена научно-технической информацией в области двигателестроения и другими факторами. В отрасли существует несколько международных проектов в области НИОКР, однако они не предполагают регулярной публикации документов на всех используемых в проектах языках. «Ненатуральность» переводных текстов свойственна большой части доступных документов.

2.2. ПРОЦЕДУРЫ И СРЕДСТВА ЛЕКСИКОГРАФИЧЕСКОГО АНАЛИЗА КОРПУСА ТЕКСТОВ

Средства работы с терминологией предполагают средства извлечения терминов, предназначенные для автоматизации обработки корпусов текстов, и средства менеджмента терминологии, обеспечивающие ее запись, обработку, сохранение и использование в области создания документов, словарей, процесса перевода и терминологической работы. Эти средства могут быть ориентированы на обработку терминов и/или понятий, работу с конкретной языковой парой, многоязычными или одноязычными ресурсами. Многие подобные средства включаются в системы поддержки автоматизации перевода и включают системы переводческой памяти, средства выравнивания параллельных текстов, редактора тэггирования, в некоторых случаях сюда же включаются средства извлечения терминов из текста.

Рассмотрение полнотекстового корпуса параллельных текстов в качестве лексикографической базы предполагает необходимость ее дополнения корпусом машинных переводов текстов, что позволяет явным образом выделить те лексические единицы, которые должны быть введены в словарь или модифицированы с точки зрения набора переводных эквивалентов. Работа терминолога и лексикографа при создании переводных словарей непосредственно связана с осуществлением перевода. Постредактирование результатов МП и получение окончательного варианта перевода текста требует обращения к словарным и энциклопедическим базам данных, выбранным переводчиком, а также к заранее выбранным корпусам текстов. При решении вопроса о выборе перевода конкретной терминологической единицы необходимо привлечение миниконкорданса. В результате работы на этапе собственно перевода должен формироваться пользовательский словарь, характеризующий терминологические особенности конкретного текста. Этот словарь на этапе поддержки автоматизированного рабочего места специалиста добавляется в его лингвистические ресурсы.

Для проведения лексикографического исследования корпуса необходимо формирование отрицательного словаря — словаря лексических единиц, не являющихся терминологическими в любой предметной области. Подобный словарь может быть сформирован

на основе базовых словарей систем машинного перевода или в результате сопоставления словарей, полученных по представительным выборкам из текстов различных предметных областей. В его состав должны входить лексические единицы служебных частей речи, вспомогательные глаголы и глаголы широкой семантики, устойчивые словосочетания со значением времени, места, последовательности действий или событий, условий, цели и т. п.

Подобный отрицательный словарь может использоваться при выявлении терминологических единиц в предметно-ориентированном корпусе текстов. На основе размеченного корпуса текстов можно получить частотный словарь слов и словосочетаний, который должен быть сопоставлен с отрицательным словарем.

В результате удаления из полученного частотного списка единиц отрицательного словаря получаем список слов и словосочетаний, которые могут рассматриваться как терминологические. При фиксации именных групп появляется возможность объединять в гнезда лексические единицы с одним и тем же ядерным элементом.

Рассмотрим эту процедуру на примере полученного из корпуса текстов Болонского процесса анализа ядерных элементов system и systems и словосочетаний с ними, извлеченных из частотного словаря слов и именных словосочетаний. В список, приведенный ниже, вошли 119 именных словосочетаний:

academic system	banner system
accreditation and quality assurance systems	binary system(s)
accreditation system	coherent system
accumulation and transfer system	common credit system
accumulation system	common education system
adequate peer review system	computer system
appeals system	computer'based systems
appropriate systems	convergent system
assessment system(s)	countries' qualifications systems

credit accumulation and transfer system(s)
credit-based systems
credit system(s)
credit transfer and accumulation system(s)
credit transfer system
currently rigid national system
current education systems
current guidance system
current quality assurance systems
current system
dual system
ECTS grading system
ECTS system
educational system(s)
education and training systems
education system(s)
effective systems
European credit transfer and accumulation system
European credit transfer system
European education system
European higher education system
European system
Europe's education systems
EU engineering educational systems
existing European systems
explicit systems
Finnish university system
formal system
general system
German system
given education system
global higher education system
global university system
higher educational system
higher education and research systems
higher education system(s)
Hungarian accreditation system
independent and credible appeals system
individual education systems
information system(s)
Italian higher education system
italian university system
large software system
level systems
long-established and powerful higher education systems
long cycle system
mass higher education system
mobility system
modular system
national and institutional quality assurance systems
national credit system(s)
national diploma systems

national education system(s)
national higher education system(s)
national or other relevant system
national or related systems
national qualifications systems
national quality assurance system(s)
national system(s)
new computer systems
newly launched system
new post-Bologna system
new system
own higher education system
own quality assurance systems
peer review system
political systems
potential systems
present system
proposed peer review system
qualifications systems
quality assurance system(s)
quality-related information systems
recognition and credit systems
recognition system
respective higher education systems
school system(s)
separate credit system
separate systems
single system
slightly different system
spanish university system
specific national educational systems
still very liberal system
student support systems
student-centred system
suitable common credit system
traditional systems
truly international system
two-degree system
two-prong system
two-tier system
unitary system
university information system
university system(s)
user-friendly education systems
VET credit system
vocational and education training systems
vocational education and training systems
well-established system
widely accepted system
wider pan-European credit

Первым этапом анализа этого списка является исключение из него словосочетаний с прилагательными, характеризующими национальную принадлежность или оценку системы, например словосочетания German system, current system из списка исключаются, а словосочетание wider pan-European credit accumulation and transfer system приводится к форме credit accumulation and transfer system. В результате такой процедуры число словосочетаний уменьшается до 70, т. е. сокращается на 41 %.

Анализ этого списка позволяет выделить единицы для включения в словарь соответствующего языка для профессиональных целей. Кроме того, он позволяет установить различные типы отношений между терминами (отношения относительной синонимии, в частности) и может использоваться для их классификации.

academic system
accreditation and quality assurance systems
accreditation system
accumulation and transfer system
accumulation system
appeals system
assessment system(s)
banner system
binary system(s)
common credit system
computer system(s)
computer-based systems
convergent system
credit accumulation and transfer system(s)
credit-based systems
credit system(s)
credit transfer and accumulation system(s)
credit transfer system
dual system
ECTS grading system
ECTS system
education and training systems
education system(s)
educational system(s)
engineering educational systems
guidance system
higher education and research systems
higher education system(s)
higher educational system
independent and credible appeals system
individual education systems

information system(s)
level systems
long cycle system
mobility system
modular system
national and institutional quality assurance systems
national credit system(s)
national diploma systems
national education system(s)
national educational systems
national higher education system(s)
national or other relevant system
national or related systems
national qualifications systems
national quality assurance system(s)
national system(s)
peer review system
political systems
post-Bologna system
qualifications systems
quality assurance system(s)
quality-related information systems
recognition and credit systems
recognition system
school system(s)
separate credit system
software system
student support systems
student-centred system
two-degree system
two-prong system
two-tier system
unitary system
university information system
university system(s)
user-friendly education systems
VET credit system
vocational and education training systems
vocational education and training systems

В следующем списке отражены отношения синонимии и гипонимии, жирным шрифтом выделены словосочетания, являющиеся ядерными для построения структуры соответствующего терминополя, курсивом выделено терминологическое словосочетание, установленное на материале анализа списка, но не зафиксированное в корпусе как самостоятельное.

На основе подобной процедуры лексические единицы, входящие в гнездо с одним и тем же ядерным элементом, могут быть отсортированы так, что устойчивые комплексы выделяются и оцениваются в соответствии с их суммарной частотой использования. Более того, на следующем этапе лексические единицы, составляющие повторяющиеся части выделенных из текста словосочетаний могут быть установлены в структуре гнезд с другими ядерными элементами, что позволяет установить связи между отдельными гнездами.

Лексическая структура, получаемая на уровне выделенных гнезд, может рассматриваться как база для установления терминополя. Рассмотрим это на примере словосочетаний с элементом higher education, являющемся самым частотным в корпусе текстов Болонского процесса. Для установления структуры лексикографического описания — терминосистемы поля 'higher education', рассмотрим все лексические единицы, непосредственно относящиеся к данному полю (табл.1).

Таблица 1

Алфавитный словарь словосочетаний с компонентом higher education в корпусе текстов Болонского процесса

№	Словосочетание	Длина в слово-формах	Частота в массиве	Модель
1	accredited vocational higher education	4	1	107
2	accredited vocational higher education programs	5	2	412
3	Austrian higher educational law	4	1	395
4	autonomous and effective higher education institutions	6	1	245
5	autonomous higher education institutions	4	3	43
6	classic higher education programs	4	1	43
7	coherent European higher education area	5	1	134
8	department of science and higher education	6	3	459

№	Словосочетание	Длина в слово-формах	Частота в массиве	Модель
9	different higher education communities	4	1	43
10	different higher education qualifications	4	1	43
11	different higher education qualifications and titles	6	3	460
12	effective European higher education area	5	1	134
13	EHEA's higher education offering	4	1	81
14	engineering higher education	3	1	115
15	environment Serbian higher education	4	1	476
16	European area of higher education	5	1	477
17	European higher education	3	9	52
18	European higher education area	4	93	54
19	European higher education degrees	4	1	54
20	European higher education institutions	4	2	54
21	European higher education political agenda	5	1	480
22	European higher education qualifications	4	3	54
23	European higher education sector	4	1	54
24	European higher education system	4	4	54
25	European ministers in charge of higher education	7	1	481
26	European ministers of higher education	5	1	477
27	European ministers responsible for higher education	6	1	482
28	European national higher education frameworks of qualifications	7	2	483
29	existing higher education institutions	4	1	135
30	first and second higher education degrees	6	1	499

№	Словосочетание	Длина в слово-формах	Частота в массиве	Модель
31	flexible higher education frameworks of qualifications	6	1	502
32	French community ministry for higher education and research	8	1	490
33	further and higher education	4	1	14
34	further and higher education and training	6	1	494
35	global higher education system	4	1	43
36	good quality higher education	4	1	78
37	higher education	2	291	1
38	higher education accreditation	3	1	10
39	higher education act	3	2	10
40	higher education activity	3	1	10
41	higher education and research	4	12	140
42	higher education and research and industry	6	1	500
43	higher education and research council	5	1	114
44	higher education and research sectors	5	1	114
45	higher education and research systems	5	1	114
46	higher education and science	4	1	140
47	higher education and training awards council	6	1	501
48	higher education area	3	1	10
49	higher education assessments	3	1	10
50	higher education authority	3	1	10
51	higher education awarding bodies	4	1	103
52	higher education degree	3	1	10
53	higher education degrees	3	1	10

№	Словосочетание	Длина в слово-формах	Частота в массиве	Модель
54	higher education diploma	3	1	10
55	higher education engineering institutions	4	1	103
56	higher education entry and exit points	6	1	502
57	higher education frameworks of qualifications	5	4	503
58	higher education institution	3	5	10
59	higher education institutions	3	86	10
60	higher education institutions recognition	4	1	18
61	higher education level	3	3	10
62	higher education opportunities	3	1	10
63	higher education part	3	1	10
64	higher education policies	3	2	10
65	higher education policy	3	1	10
66	higher education policy makers	4	1	18
67	higher education professors/teachers	4	1	400
68	higher education programme of study	5	1	503
69	higher education programmes of learning	5	1	503
70	higher education qualification	3	1	10
71	higher education qualification subject area	5	1	36
72	higher education qualifications	3	36	10
73	higher education quality assurance	4	1	18
74	higher education reform	3	1	10
75	higher education reform project	4	1	18
76	higher education reforms	3	1	10
77	higher education schemes	3	1	10

№	Словосочетание	Длина в слово-формах	Частота в массиве	Модель
78	higher education short cycle	4	1	78
79	higher education specialists	3	1	10
80	higher education staff	3	1	10
81	higher education stakeholders	3	1	10
82	higher education standards	3	1	10
83	higher education standards and guidelines	5	1	461
84	higher education structures	3	1	10
85	higher education studies	3	1	10
86	higher education system	3	10	10
87	higher education systems	3	15	10
88	higher educational system	3	2	9
89	higher educations frameworks	3	2	10
90	individual higher education institution	4	1	54
91	individual higher education institutions	4	1	54
92	institute of higher education	4	1	429
93	institutes of higher education	4	4	429
94	interdisciplinary higher education study programs	5	1	64
95	Irish higher education provision	4	1	43
96	Italian higher education system	4	1	54
97	long-established and powerful higher education systems	7	1	520
98	mass higher education system	4	2	54
99	minister for higher education	4	1	429
100	most important higher education reform process	6	1	531

№	Словосочетание	Длина в слово-формах	Частота в массиве	Модель
101	national higher education frameworks of qualifications	6	3	535
102	national higher education qualifications	4	1	54
103	national higher education system	4	2	54
104	national higher education systems	4	3	54
105	new European quality assurance network for higher education	8	1	543
106	new higher education act	4	1	43
107	new style higher education	4	1	78
108	OECD higher education review	4	1	81
109	other higher education institutions	4	1	43
110	other higher education qualifications	4	1	43
111	own higher education system	4	1	43
112	quality assurance agency for higher education	6	1	572
113	recognised higher education programme of study	6	1	575
114	respective higher education systems	4	1	43
115	schools of higher education	4	1	24
116	short cycle higher education	4	2	78
117	short cycle higher education qualifications	5	1	87
118	short higher education	3	3	9
119	short higher education qualifications	4	2	43
120	shorter higher education	3	4	9
121	shorter higher education programs	4	1	43

№	Словосочетание	Длина в словоформах	Частота в массиве	Модель
122	state minister of higher education and science	7	1	592
123	three higher education laws	4	1	119
124	typical higher education qualifications	4	1	18
125	university higher education	3	1	23
126	virtual higher education	3	1	9

В исследуемом массиве зафиксировано 126 словосочетаний с элементом *higher education*, являющимся самым частотным в массивах текстов Болонского процесса.

Максимальная длина для рассмотренных словосочетаний составляет 8 элементов (отметим, хотя максимальная длина словосочетания, зафиксированная в массиве текстов Болонского процесса, составляет 11 элементов, это модель типа N1+Pr+N2+A+N3+N4+pr+N5+C+N6+N7 – Council of Europe Parliamentary Assembly Committee on Science and Technology Conference с частотой 1).

Для выявления структуры лексикографического описания – терминосистемы поля 'higher education' получим частотный словарь лексем, составляющих эти словосочетания, исключив из него служебную лексику (табл. 2).

Таблица 2

Частотный словник лексических единиц из словосочетаний с блоком higher education (Фрагмент)

Частота	ЛЕ	Частота	ЛЕ
106	higher	1	authority
104	education	1	awarding
16	European	1	awards
15	qualification	1	bodies

Частота	ЛЕ	Частота	ЛЕ
13	system	1	charge
11	institution	1	classic
7	program(mes)	1	coherent
6	research	1	department
6	area	1	diploma
4	degree	1	EHEA
5	frameworks	1	entry
5	Minister	1	environment
5	national	1	established

В результате получаем словник ключевых единиц, который можно использовать как базу структурирования терминосистемы (табл. 3).

Таблица 3

Словник частотных ключевых единиц для словосочетаний с компонентом higher education в текстах Болонского процесса (Фрагмент)

Ключевая единица	Частота	Синоним из списка	Частота	Суммарная частота
qualification	15	diploma	1	16
system	13			13
institution	11	institute	2	15
		schools	1	
		university	1	
program(mes)	7	cycle	3	12
		guidelines	1	
		schemes	1	
research	6	science	3	8
area	6	sector	1	7

Ключевая единица	Частота	Синоним из списка	Частота	Суммарная частота
minister	5	ministry	1	16
		agency	1	
		authority	1	
		bodies	1	
		community	2	
		council	2	
		department	1	
		OECD	1	
frameworks	5	networks	1	6
study	4	training	2	7
		learning	1	
reform	4			4

Используя выделенные ключевые единицы как основу для объединения, получаем следующую структуру терминосистемы (табл. 4):

Таблица 4

Структура терминосистемы higher education

Словосочетание	Длина	Частота	Модель
higher education			
Программы высшего образования			
higher education program	3	0	10
higher education programme of study	5	1	503
higher education programmes of learning	5	1	503
recognised higher education programme of study	6	1	575
accredited vocational higher education programs	5	2	412
classic higher education programs	4	1	43
shorter higher education programs	4	1	43
interdisciplinary higher education study programs	5	1	64
Системы высшего образования			
higher education system	3	10	10
higher education systems	3	15	10
higher educational system	3	2	9
European higher education system	4	4	54
global higher education system	4	1	43
Italian higher education system	4	1	54
long-established and powerful higher education systems	7	1	520
mass higher education system	4	2	54
national higher education system	4	2	54
national higher education systems	4	3	54
own higher education system	4	1	43
respective higher education systems	4	1	43
Учреждения высшего образования			
higher education institution	3	5	10
higher education institutions	3	86	10
institute of higher education	4	1	429
institutes of higher education	4	4	429
schools of higher education	4	1	24
higher education institutions recognition	4	1	18
autonomous and effective higher education institutions	6	1	245
autonomous higher education institutions	4	3	43

Словосочетание	Длина	Частота	Модель
different higher education communities	4	1	43
European higher education institutions	4	2	54
existing higher education institutions	4	1	135
individual higher education institution	4	1	54
individual higher education institutions	4	1	54
higher education engineering institutions	4	1	103
higher education structures	3	1	10
other higher education institutions	4	1	43
Квалификации высшего образования			
higher education qualification	3	1	10
higher education qualification subject area	5	1	36
higher education qualifications	3	36	10
different higher education qualifications	4	1	43
different higher education qualifications and titles	6	3	460
European higher education qualifications	4	3	54
national higher education qualifications	4	1	54
other higher education qualifications	4	1	43
short cycle higher education qualifications			
short higher education qualifications	4	2	43
typical higher education qualifications	4	1	18
Структуры высшего образования			
higher educations frameworks	3	2	10
higher education frameworks of qualifications	5	4	503
European national higher education frameworks of qualifications	7	2	483
flexible higher education frameworks of qualifications	6	1	502

Словосочетание	Длина	Частота	Модель
national higher education frameworks of qualifications	6	3	535
new European quality assurance network for higher education	8	1	543
Законодательная база высшего образования			
Austrian higher educational law	4	1	395
European higher education political agenda	5	1	480
higher education accreditation	3	1	10
higher education act	3	2	10
new higher education act	4	1	43
higher education policies	3	2	10
higher education policy	3	1	10
higher education policy makers	4	1	18
higher education standards	3	1	10
higher education standards and guidelines	5	1	461
three higher education laws	4	1	119
Типы высшего образования			
accredited vocational higher education	4	1	107
EHEA's higher education offering	4	1	81
engineering higher education	3	1	115
environment Serbian higher education	4	1	476
European higher education	3	9	52
further and higher education	4	1	14
further and higher education and training	6	1	494
good quality higher education	4	1	78
higher education short cycle	4	1	78
Irish higher education provision	4	1	43

Словосочетание	Длина	Частота	Модель
new style higher education	4	1	78
short cycle higher education	4	2	78
short higher education	3	3	9
shorter higher education	3	4	9
university higher education	3	1	23
virtual higher education	3	1	9
Управление высшим образованием			
department of science and higher education	6	3	459
European ministers in charge of higher education	7	1	481
European ministers of higher education	5	1	477
European ministers responsible for higher education	6	1	482
French community ministry for higher education and research	8	1	490
higher education and research council	5	1	114
higher education and training awards council	6	1	501
higher education authority	3	1	10
higher education awarding bodies	4	1	103
minister for higher education	4	1	429
state minister of higher education and science	7	1	592
Степени высшего образования			
higher education degree	3	1	10
higher education degrees	3	1	10
European higher education degrees	4	1	54
first and second higher education degrees	6	1	499

Словосочетание	Длина	Частота	Модель
Аудит качества высшего образования			
higher education assessments	3	1	10
higher education level	3	3	10
higher education quality assurance	4	1	18
quality assurance agency for higher education	6	1	572

Исследование словосочетаний, выявленных и организованных на основе предложенной процедуры, позволяет установить те из них, которые следует включать в словари и глоссарии, а также определить потенциально возможные словосочетания (см., например, *higher education law, higher education program*).

Еще раз отметим, что лексикографическая работа даже с использованием возможностей информационных технологий остается работой творческой и не может быть полностью автоматизирована. Одним из вариантов создания материала для последующего лексикографического анализа является формирование особых корпусов текстов, включающих параллельное представление исходных текстов, их машинных переводов и отредактированных переводов, согласованных с экспертами в конкретной области знаний. Важно отметить, что качество и потенциал такого корпуса в большой степени зависит от сотрудничества с экспертами при отборе исходного материала и редактировании переводов. Размер корпуса для создания терминологического словаря по узкой предметной области зависит от терминонасыщенности текстов, в среднем составляя по оценкам исследователей 200 тысяч словоупотреблений [48, с. 319].

2.3. МЕТОДЫ И ИНСТРУМЕНТЫ ОБРАБОТКИ ТЕКСТОВЫХ ДАННЫХ

Последующая работа с корпусом предполагает сопоставление потенциалов существующих на сегодняшний день методов обработки текстовых данных: методов создания конкордансов (и последующего семантического анализа выявленных терминологических единиц), методов автоматического выделения терминов в специальных текстах, методов поиска текстовых единиц на основе лексико-синтаксических шаблонов и т. д.

Конкорданс. Конкорданс является «первой производной корпуса», дающей все употребления данного слова в контексте со ссылками на источник [38]. Если корпус представителен, то конкорданс стремится к тезаурусной полноте описания, т. е. ориентирован на исчерпывающую фиксацию всех вариантов употребления слова и описание его контекстов. Таким образом, конкорданс является базовым ресурсом, характеризующим совместную встречаемость лексических единиц. При этом в конкордансе принципиально важна и морфологическая информация, так как она помогает выявить особенности употребления форм слова, которые при необходимости можно отразить в словаре.

Одним из важнейших вопросов, возникающих при анализе терминов и терминологических сочетаний, полученных из конкорданса либо с применением других методов обработки текста, является оценка степени терминологичности рассматриваемых единиц. В решении этого вопроса большой вес имеет мнение экспертов в данной предметной области. Для фиксации полученных терминов необходимо также определить способ описания значения, т. е. инвариантного представления слова в лексикографической базе с учетом определимости одних терминов через другие [86; 103].

Автоматическое извлечение терминов. Методики выделения терминов при вероятностно-статистических подходах опираются на представление о том, что термины, как правило, являются наиболее частотными словами и словосочетаниями, встречающимися в специальных текстах и выражающими понятия предметной области [59].

Возможность использования статистических оценок для определения степени терминологичности и устойчивости словосочетаний

была осознана очень давно, более полувека тому назад (см., например, работы Р. Г. Пиотровского, Д. Н. Андреева, В. И. Перебейнос, М. В. Арапова, А. Я. Шайкевича и их коллег и учеников). Однако применение тогда же предложенных оценок было затруднено отсутствием больших массивов данных на электронных носителях. Создание электронных ресурсов и, в частности, корпусов данных позволило заново обратиться к статистическому анализу лингвистического материала, объем которого гарантировал достоверность и репрезентативность получаемых результатов.

Поскольку большинство терминологических единиц представляют собой не универбы, а многословные словосочетания, то предполагается, что установление границ подобных словосочетаний — коллокаций может основываться на применении частичного синтаксического анализа, т. е. анализа на уровне групп. Синтагматическая сочетаемость лексических единиц позволяет определить подход к описанию и выделению коллокаций как вида словосочетаний [98]. Необходимость такого рассмотрения связана с тем, что при решении задач, относящихся к сфере лингвистических технологий, выделение словосочетаний, являющихся необходимым элементом автоматических словарей, как правило, осуществляется разработчиками интуитивно.

Элементы коллокаций характеризуются определенной семантической взаимообусловленностью, а сами коллокации — воспроизводимостью [98, с. 38] и предсказуемостью [98, с. 43], что может устанавливаться статистическими методами. При этом терминологичность определяется как та «степень, до которой устойчивая лексическая единица связана с некоторым количеством понятий, зависящих от предметной области», а синтагматичность как та «степень, до которой последовательность слов способна формировать устойчивую лексическую единицу» [128].

Соответственно, на основе понятий коллокаций, терминологичности и синтагматичности устанавливаются методы статистического анализа и метрики, используемые для выделения коллокаций из текста. Эти меры, активно используемые сегодня при выделении терминологических словосочетаний из корпусов текстов, оценивают:

- информацию о сочетаемостных предпочтениях слов (unithood) — коэффициент синтагматической близости MI;

- информацию о степени терминологичности коллокации (termhood) — коэффициент терминологичности T (T-score), устанавливающий меру ассоциации [98, с. 22];
- информацию о значимости (salience) коллокации для конкретного корпуса текстов или терминологии языка для специальных целей.

Методы, используемые для оценки синтагматичности, т. е. устойчивости фиксации элементов словосочетания, основываются на оценках частоты встречаемости словосочетания в текстах конкретной предметной области. Для оценки терминологичности словосочетания требуется сравнение частотных характеристик словосочетания в разных предметных областях, т. е. требуется подход, который принято называть контрастивным.

Контрастивные подходы основываются на зависимости терминов от конкретной предметной области, что может иметь количественное выражение. Как следствие, в этой области они появляются более часто, чем в других областях. Так, например, метод анализа устойчивых лексических маркеров [98] принимает во внимание два свойства лексических единиц, специфических для предметной области. Как и другие контрастивные подходы, этот метод позволяет извлекать слова, которые имеют частоту выше средней в специализированном корпусе текстов, но дополнительно к этому метод обеспечивает, что у этих же слов высокая дисперсия в специализированном корпусе текстов. В этом состоит преимущество отфильтровывания любого смещения частоты, которое могло быть вызвано просто использованием в части корпуса текстов. Это явление может быть связано со смещением в теме текста, например встречается в только одном тексте, в котором активно обсуждается тема, в других обстоятельствах не имеющая отношения к исследуемой предметной области.

Следует отметить, что в принципе все используемые сегодня статистические оценки терминологичности и синтагматической устойчивости по сути являются эвристиками, поскольку в предлагаемые в них формулы вводятся коэффициенты, которые позволяют получить корректные результаты для различных предметных областей. Необходимость таких эвристик связана с тем, что (как показывают исследования) в различных языках для специальных целей структуры терминологических сочетаний различаются кардинально.

Лексико-синтаксические шаблоны. Формализация повторяющихся, шаблонных, элементов текста успешно используется для решения некоторых задач автоматической переработки текста. Например, шаблоны жестко структурированных текстов (патентов, ведомостей, спецификаций и т. п.) используются для определения их жанровой и тематической принадлежности [66, с. 117]. Концепция лексико-синтаксических шаблонов сформировалась в ходе изучения дискурсивных особенностей научно-технической прозы. Характерные для нее конструкции, типа «далее докажем P», «допустим, что S», были формализованы, т. е. множество входящих в подобные конструкции лексем, их возможных грамматических форм и синтаксических условий были зафиксированы в некоторой декларативной структуре — лексико-синтаксическом шаблоне языковой конструкции.

Язык записи лексико-синтаксических шаблонов LSPL (Lexical-Syntactic Pattern Language) был детально проработан с учетом специфики русского языка и применен в первую очередь для создания шаблонов, регулярно используемых в научно-технических текстах фраз-определений новых (авторских) терминов [19]. Во многих исследованиях и созданных на их основе программных инструментах практикуется комбинированный подход, заключающийся в (полу) автоматической обработке специальных корпусов текстов.

2.4. ЛИНГВИСТИЧЕСКИЙ СЕТЕВОЙ ИНСТРУМЕНТАРИЙ АНАЛИЗА ТЕКСТА

На сегодняшний день отечественными и зарубежными разработчиками создано множество программ, конкретные методы обработки текстовых данных. В качестве базовых критериев выбора программ обработки текстовых данных можно считать следующие:

- основное назначение (парсинг, исследование поведения слов в текстах, автоматическое аннотирование и т. д.);
- функциональность (возможность использования разных приложений программы для решения нескольких исследовательских задач);
- удобство использования (удобство настроек и интерфейса, возможность работы с файлами в требуемом формате);

- условия приобретения (платные/бесплатные, возможность тестирования демоверсии).

В соответствии с первым критерием — основное назначение — для решения задач данного проекта рассматривались программы, предназначенные для извлечения терминов из неразмеченного корпуса текстовых файлов, получения статистической информации по выбранным терминам и построения к ним конкордансов. С точки зрения функциональности предпочтение отдавалось программам, совмещающим несколько функций. Платные программы рассматривались при условии наличия демоверсии и подробного описания. На соответствие данным критериям были изучены и протестированы несколько программных средств, условно разделенных на три группы: инструменты извлечения терминов, инструменты получения статистической информации и инструменты построения конкордансов.

Инструменты извлечения терминов. Удобным и открытым для свободного доступа инструментом извлечения терминов является программа Terminology Extraction исследовательского центра T-Labs. Извлечение осуществляется на основе распределения Пуассона, вычисления коэффициента подобия и частоты встречаемости термина по сравнению с 100-миллионными корпусами текстов на английском, итальянском и французском языках. Единицы, часто встречающиеся в документе, но редко в языке, признаются программой вероятными терминами, вероятность их терминологичности обозначается баллами. Результат извлечения выводится в виде таблицы. Исходный текст помещается под таблицей, найденные в нем терминологические единицы маркируются, таким образом можно увидеть их в контекстном окружении. Недостатком программы является то, что с ее помощью каждый документ корпуса должен обрабатываться отдельно, поэтому ее нельзя признать полностью соответствующей критериям назначения и удобства.

Одновременную обработку нескольких текстов поддерживает программа WordTabulator v.2.2.3. Инструмент предназначен «для построения упорядоченного индекса символьных элементов в заданном множестве текстов». Обрабатываемые элементы задаются в настройках, типы элементов определяются разработчиками как словоформы, словосочетания и синтагмы. В индексе (таблице результатов) указывается частота каждого элемента и документы,

в которых он обнаружен. Программа поддерживает русский и английский языки, формат исходных файлов –.html и.txt. Возможность настройки стоп-листа отсутствует, меры ассоциации не вычисляются, поэтому программу нельзя назвать очень удобной для извлечения терминологии.

Как следует из описаний инструментов LogiTermPro (Terminotix) и SDL MultiTerm Extract 2009, они в полной мере соответствуют необходимым критериям, но в связи с высокой стоимостью и отсутствием бесплатных демоверсий не тестировались.

Инструменты получения статистической информации. Наиболее простые и доступные программы для получения статистической информации об элементах текста — приложение для создания частотных списков словаря Мультитран, SimWordSorter, Content Analyser v0.52. Данные инструменты, за исключением Content Analyser, вычисляют только частоту встречаемости отдельных слов в отдельных текстах. Content Analyser обрабатывает слова и словосочетания, но только в файлах в формате.html, так как его основное назначение — анализ содержания тематических web-страниц. Таким образом, указанные инструменты только частично соответствуют критериям назначения и функциональности и не соответствуют критерию удобства использования.

Инструменты построения конкордансов. Инструмент Simple Concordance Program 4.09, несмотря на название, следует причислить к программам извлечения элементов текста, так как с его помощью можно получить список словосочетаний с заданным количеством слов, но контекстный просмотр найденных элементов отсутствует.

Издательство Athelstan предлагает набор программных средств построения конкордансов. MonoConc Pro — инструмент для загрузки и поиска терминов по неразмеченному корпусу текстов на английском и нескольких других языках. Тексты должны быть в формате. txt. ParaConc Pro — инструмент для работы одновременно с двумя, тремя или четырьмя параллельными текстами, обеспечивает выравнивание текстов. Еще один инструмент — Collocate — предназначен для поиска в корпусе окружения для заданного слова (от 2 до 6 слов) и вычисления мер ассоциаций для извлеченных элементов. Используются меры Log Likelihood, Mutual Information, t-score. К сожалению, ни одна из этих программ не сопровождается демоверсией, под-

робное руководство пользователя предоставляется только при заказе программы.

Пакет программ WordSmith Tools позволяет получать список слов отдельного англоязычного документа и совокупности документов, сравнивать полученный список с частотным списком слов Британского национального корпуса, вычислять ключевые слова текста и корпуса текстов, получать конкорданс к отдельному слову и нескольким словам, вычислять кластеры слов. Используемые меры ассоциаций: Specific Mutual Information, MI3, Z Score, Log Likelihood, t-score; меняя настройки приложений, их можно использовать совместно или избирательно. Программа позволяет соотнести полученный список ключевых слов с имеющейся базой терминов, однако подключаемая база должна быть сформирована также в формате .txt и состоять из однословных наименований. Программа платная, но снабжена подробным, доступным руководством и удобной демоверсией. Поддерживаются другие языки, но для них нет частотных списков национальных корпусов.

Достоинством программы AntConc, помимо удобного интерфейса и бесплатного распространения, является возможность установки на разные операционные системы, включая Windows, MacOS и Linux. Программа представляет собой многофункциональный инструмент для наблюдения за поведением ключевых слов в избранной совокупности текстов (как в единичном тексте, так и в корпусе). Доступны контекстный просмотр терминов, просмотр исходных файлов, построение n-грамм различной длины для заданных элементов, формирование частотного списка слов всего корпуса, а также выявление ключевых слов одного корпуса в результате сравнения с более крупным корпусом (используются меры Log-Likelihood и Chi Squared). Разработчик не предъявляет никаких требований к формату загружаемых файлов корпуса, однако при обработке файлов в формате .txt не всегда распознается кодировка текста.

Индексирование документа и модели. Термины используются как ключевые элементы в задачах индексирования фрагментов текста, что позволяет индексировать не документ в целом, а каждый его фрагмент отдельно. Точно также происходит индексация модели, при которой строятся классы элементов.

Генерация гиперссылок. Выявление терминов в документе и модели является основой для автоматического формирования гиперссылок, что позволяет (кроме всего прочего) пользователю редактировать сформированные гиперссылки: дополнять или исключать их.

Генерация модели. Общеизвестно, что иерархия понятий отражает иерархию терминов, выявляемых из документов. Это свойство может использоваться для создания остова модели терминосистемы, которая может дополняться и конкретизироваться вручную.

2.5. МЕТОДЫ И МЕТРИКИ ИЗВЛЕЧЕНИЯ ТЕРМИНОЛОГИИ ИЗ КОРПУСОВ ТЕКСТОВ

Как уже говорилось выше, идея создания автоматизированных систем извлечения терминов из корпусов параллельных текстов насчитывает уже более 20 лет и в той или иной степени реализуется в различных проектах. Очень важно понимать, что даже самая изощренная система извлечения терминов не дает окончательного результата для включения их в переводной словарь, а предоставляет лишь удобно организованный и оперативно получаемый ресурс для работы терминолога или лексикографа.

Существуют два основных подхода к проблеме автоматического извлечения из корпусов текстов информации о переводных соответствиях двух языков и построения на ее основе лексических конкордансов. Решение задачи в рамках первого подхода начинается с выравнивания параллельных текстов.

Для выравнивания параллельных текстов обычно используются определенные эвристические соображения (или просто эвристики), которые помогают выбрать точки соответствия (точнее, кандидаты на то, чтобы ими быть). В роли последних могут выступать, например, числа, аббревиатуры, даты, имена собственные, акронимы, устойчивые словосочетания, имеющие однозначные переводные эквиваленты на втором языке. Для достаточно близких языков, как, например, испанский и португальский в роли пар кандидатов могут рассматриваться слова, имеющие одинаковые корни и интернационализмы в целом.

Если в более ранних работах эвристики использовались напрямую, без какой-либо поддержки статистическими методами, то в настоящее время использование статистических методов играет ключевую роль в содержании большинства публикаций по проблеме автоматического выравнивания параллельных текстов. Наиболее общий метод статистической поддержки, инвариантный относительно того, какие эвристики были использованы при создании исходного списка пар-кандидатов, основан на использовании корреляционной зависимости между позициями точек (лексических единиц), составляющих пару.

Исходным материалом для выполнения алгоритма является предварительный список таких пар-кандидатов (w^1, w^2), что число вхождений первого компонента пары в текст на языке 1 равно числу вхождений второго компонента в текст на языке 2. Слова текста на языке нумеруются числами от 1 до длины текста в словах. Аналогичным образом слова второго текста нумеруются числами от 1 до длины текста на языке 2. Это позволяет каждому i-му вхождению лексической единицы w^1 (как первого элемента некоторой билингвистической пары) и i-му вхождению лексической единицы w^2 (как второго элемента билингвистической пары) сопоставить пару чисел (x^1_i, x^2_i), задающих номера их позиций в текстах на языке 1 и языке 2 соответственно. В некоторых работах [например, 62] расстояние от начала текста до текущего слова измеряется не в словах, а в литерах.

Во избежание недоразумений следует заметить, что кроме пар лексических единиц в состав исходного списка могут включаться перечисленные выше пары чисел, аббревиатур, дат, имен собственных и т. п. На первой стадии выполнения алгоритма сформированные пары следует рассматривать в качестве кандидатов на роль разделителей параллельных текстов на соответствующие друг другу сегменты.

Множество полученных пар $\{(x_i, y_i)\}_{i=1}^{n}$ служит основой для построения простейшей корреляционной связи между координатами вхождения соответствующих друг другу точек в каждый из двух текстов — функции линейной регрессии (точнее, функции линейной регрессии второго рода [112]) переменной y относительно переменной x, представленной в виде $y=a+bx$, где a и b подобраны так, чтобы выполнялось условие

$$f(a,b) = \sum_{i=1}^{n} (y_i - (a + bx_i)) = min$$

Другими словами, функция $f(a, b)$ определяется так, чтобы она давала в среднем по возможности лучшее приближение к наблюдаемым значениям случайной переменной y, принятой за зависимую.

Коэффициенты a и b находятся стандартным путем. Приравнивание нулю частных производных функции $f(a, b)$ по a и b приводит к системе линейных уравнений

$$\begin{cases} (\sum_{i=1}^{n} x_i)^2 \times a + n \times b = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} (x)^2 \times a + (\sum_{i=1}^{n} x_i) \times b = \sum_{i=1}^{n} (x_i y_i) \end{cases},$$

решение которой дает искомые значения коэффициентов

$$\begin{cases} a = \dfrac{n \times \sum_{i=1}^{n}(x_i y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{(\sum_{i=1}^{n} x_i)^2 - \sum_{i=1}^{n}(x_i)^2} \\ b = \dfrac{\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n}(x_i \times y_i) - \sum_{i=1}^{n}(x_i)^2 \times \sum_{i=1}^{n} y_i}{(\sum_{i=1}^{n} x_i)^2 - \sum_{i=1}^{n}(x_i)^2} \end{cases}$$

Практическое значение функции линейной корреляции состоит в том, что она позволяет на основе положения лексической единицы в тексте на одном языке прогнозировать положение соответствующей ей лексической единицы в тексте на другом языке. А это, в свою очередь, позволяет отсеять те «плохие» пары-кандидаты (x_i, y_i), для которых реальное (т. е. не вычисленное) значение не входит в доверительный интервал для значения y, вычисленного в соответствии с функцией линейной регрессии при $x=x_i$.

Мера доверия к прогнозируемым функцией линейной регрессии значениям y тем выше, чем больше в исходном материале билингвистических пар, элементы которых на самом деле являются соответствующими друг другу лексическими единицами. Даже в том случае,

Если в более ранних работах эвристики использовались напрямую, без какой-либо поддержки статистическими методами, то в настоящее время использование статистических методов играет ключевую роль в содержании большинства публикаций по проблеме автоматического выравнивания параллельных текстов. Наиболее общий метод статистической поддержки, инвариантный относительно того, какие эвристики были использованы при создании исходного списка пар-кандидатов, основан на использовании корреляционной зависимости между позициями точек (лексических единиц), составляющих пару.

Исходным материалом для выполнения алгоритма является предварительный список таких пар-кандидатов (w^1, w^2), что число вхождений первого компонента пары в текст на языке 1 равно числу вхождений второго компонента в текст на языке 2. Слова текста на языке нумеруются числами от 1 до длины текста в словах. Аналогичным образом слова второго текста нумеруются числами от 1 до длины текста на языке 2. Это позволяет каждому i-му вхождению лексической единицы w^1 (как первого элемента некоторой билингвистической пары) и i-му вхождению лексической единицы w^2 (как второго элемента билингвистической пары) сопоставить пару чисел (x^1_i, x^2_i), задающих номера их позиций в текстах на языке 1 и языке 2 соответственно. В некоторых работах [например, 62] расстояние от начала текста до текущего слова измеряется не в словах, а в литерах.

Во избежание недоразумений следует заметить, что кроме пар лексических единиц в состав исходного списка могут включаться перечисленные выше пары чисел, аббревиатур, дат, имен собственных и т. п. На первой стадии выполнения алгоритма сформированные пары следует рассматривать в качестве кандидатов на роль разделителей параллельных текстов на соответствующие друг другу сегменты.

Множество полученных пар $\{(x_i, y_i)\}_{i=1}^{n}$ служит основой для построения простейшей корреляционной связи между координатами вхождения соответствующих друг другу точек в каждый из двух текстов — функции линейной регрессии (точнее, функции линейной регрессии второго рода [112]) переменной y относительно переменной x, представленной в виде $y=a+bx$, где a и b подобраны так, чтобы выполнялось условие

$$f(a,b) = \sum_{i=1}^{n} (y_i - (a + bx_i)) = min$$

Другими словами, функция $f(a, b)$ определяется так, чтобы она давала в среднем по возможности лучшее приближение к наблюдаемым значениям случайной переменной y, принятой за зависимую.

Коэффициенты a и b находятся стандартным путем. Приравнивание нулю частных производных функции $f(a, b)$ по a и b приводит к системе линейных уравнений

$$\begin{cases} (\sum_{i=1}^{n} x_i)^2 \times a + n \times b = \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} (x)^2 \times a + (\sum_{i=1}^{n} x_i) \times b = \sum_{i=1}^{n} (x_i y_i) \end{cases},$$

решение которой дает искомые значения коэффициентов

$$\begin{cases} a = \dfrac{n \times \sum_{i=1}^{n} (x_i y_i) - \sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} y_i}{(\sum_{i=1}^{n} x_i)^2 - \sum_{i=1}^{n} (x_i)^2} \\ b = \dfrac{\sum_{i=1}^{n} x_i \times \sum_{i=1}^{n} (x_i \times y_i) - \sum_{i=1}^{n} (x_i)^2 \times \sum_{i=1}^{n} y_i}{(\sum_{i=1}^{n} x_i)^2 - \sum_{i=1}^{n} (x_i)^2} \end{cases}$$

Практическое значение функции линейной корреляции состоит в том, что она позволяет на основе положения лексической единицы в тексте на одном языке прогнозировать положение соответствующей ей лексической единицы в тексте на другом языке. А это, в свою очередь, позволяет отсеять те «плохие» пары-кандидаты (x_i, y_i), для которых реальное (т. е. не вычисленное) значение не входит в доверительный интервал для значения y, вычисленного в соответствии с функцией линейной регрессии при $x=x_i$.

Мера доверия к прогнозируемым функцией линейной регрессии значениям y тем выше, чем больше в исходном материале билингвистических пар, элементы которых на самом деле являются соответствующими друг другу лексическими единицами. Даже в том случае,

когда число вхождений первого компонента пары-кандидата в текст на языке 1 равно числу вхождений второго компонента в текст на языке 2 (что в реальности встречается довольно редко), гарантия того, что i-е вхождение первого компонента пары соответствует i-му вхождению второго элемента пары, отсутствует. Для устранения отмеченных недостатков, как правило, реализуются две процедуры, предшествующие построению и использованию функции линейной регрессии:

- предварительная обработка параллельных текстов на предмет получения информации, необходимой для выбора эвристик с целью составления предварительного списка билингвистических пар;
- предварительная фильтрация полученного списка на предмет удаления из него компонентов пар-кандидатов, входящих в текст изолированно, и выравнивания тем самым чисел вхождения каждого из компонентов пары в параллельные тексты.

Операции предварительной фильтрации, уточнения коэффициентов функции линейной регрессии и отбраковывания «плохих» пар-кандидатов могут выполняться итеративно, применительно к текстам в целом, и к сегментам, выделенным в ходе предыдущих итераций. При этом на разных итерациях могут использоваться разные наборы пар-разделителей.

После того, как параллельные тексты выровнены, т. е. разделены на достаточно мелкие соответствующие друг другу текстовые сегменты, делается попытка извлечь информацию о переводных эквивалентах. Один из простых статистических алгоритмов построения конкордансов, предложенный Меламедом [117], включает следующие действия:

- выбирается мера для оценки степени сходства между словами языков 1 и 2;
- для каждой пары соответствующих друг другу сегментов L_1 и L_2 вычисляются степени связи всех пар слов $(V, W) \in L_1 \times L_2 (V, W)$;
- выполняется сортировка пар по степени убывания значений меры связи;
- определяется пороговое значение степени связи слов;
- пары, для которых вычисленные значения ниже порогового, исключаются из списка, оставшиеся пары включаются в конкорданс.

Классическим вариантом статистической меры сходства между словами V и W параллельных текстов является MI-мера (Mutual Information)

$$MI(V,W) = log_2 \frac{P(V,W)}{P(V) \times P(W)}$$

В качестве вероятностей $P(v)$, $P(w)$, $P(v,w)$ принимаются:

$$P(V) = \frac{a+b}{a+b+c+d},$$

$$P(W) = \frac{a+c}{a+b+c+d},$$

$$P(V,W) = \frac{a}{a+b+c+d},$$

где a — число сегментов, в которых встречаются как первое, так и второе слово; b — число сегментов, в которых встречаются первое слово, но не встречается второе; c — число сегментов, в которых встречается второе слово, но не встречается первое; d — число сегментов, в которых не встречается ни первое слово, ни второе. Подстановка выражений для $P(v)$ и $P(v,w)$ в исходное определение MI-меры преобразует последнее к виду

$$MI(V,W) = log_2 \frac{a \times (a+b+c+d)}{(a+b) \times (a+c)}.$$

Другой популярной мерой сходства между словами является мера *t-score*, определяемой в [122] формулой

$$t\text{-}score(V,W) = \frac{P(V,W) - P(V) \times P(W)}{\sqrt{\frac{P(V,W)}{N}}},$$

где $P(v)$, $P(w)$ и $P(v,w)$ имеют тот же смысл, что и в выражении для MI-меры, а N — общее число словоформ в тексте. Несколько иная форма меры t-score приводится в работе [81]:

$$t\text{-}score(V,W) = \frac{P(V,W) - \frac{P(V) \times P(W)}{N}}{\sqrt{P(V,W)}}.$$

Критическим местом при разработке алгоритмов извлечения информации о переводных соответствиях двух языков на базе параллельных текстов является создание процедуры выравнивания параллельных текстов. И дело здесь не в трудностях алгоритмизации, а в изначальной «непараллельности» параллельных текстов. Смысл этой непараллельности Паскаль Фанг [100] формулирует в виде следующих пяти положений:

1. Слова имеют разные смыслы как в рамках пары параллельных текстов, так и в рамках корпуса в целом.
2. Слова имеют разные переводы как в рамках пары параллельных текстов, так и в рамках корпуса в целом.
3. Некоторые переводные единицы одного текста могут отсутствовать в другом.
4. Частоты вхождения лексических единиц в один и другой текст являются несопоставимыми (т. е. могут существенно различаться).
5. Позиции вхождения лексических единиц, составляющих билингвистическую пару и действительно являющихся соответствующими друг другу, часто являются несопоставимыми (т. е. сильно различающимися).

На основании сформулированных положений она делает вывод о том, что задача извлечения билингвистической терминологии из непараллельных текстов представляет собой менее трудную задачу, чем ее извлечение из параллельных текстов. Возможность эффективного использования непараллельных текстов для извлечения билингвистической терминологии Фанг обосновывает следующими двумя положениями.

1. Для текстов одной и той же предметной области и назначения лексические единицы имеют вполне сопоставимые контексты.
2. Лексические единицы разноязыких текстов одной и той же предметной области, написанных примерно в один и тот же период времени, имеют вполне сопоставимые используемые шаблоны.

Идея представленного в работе алгоритма DKvec является развитием идеи [101], состоящей в использовании контекста слов для измерения с помощью MI- и t-score мер степени их соответствия друг другу.

Цель предлагаемого алгоритма состоит в нахождении переводов (или по крайней мере кандидатов на эту роль) тех слов, которые отсутствуют в online словаре. Кроме словаря, в качестве исходного материала рассматриваются два корпуса текстов по соответствующей тематике, написанных в определенный (по отношению к настоящему) период времени. В рамках корпуса рассматриваются текстовые сегменты (последовательности слов), удовлетворяющие определенным условиям.

Предлагаемый автором алгоритм, по сути дела, представляющий собой фильтрацию списка изначально выбранных кандидатов, предполагает выполнение для каждого слова V, перевод которого в словаре отсутствует, следующих пяти шагов:

1. Составить список слов, входящих во все контексты слова V в рамках корпуса. На основании полученного списка составить вектор параметров контекста слова V.
2. Составить векторы параметров контекста для всех слов W на языке 2, которые являются кандидатами на перевод слова V.
3. Для всех пар (V, W) вычислить значение меры их сходства S.
4. Упорядочить пары по убыванию в соответствии с полученным значением меры сходства.
5. Выбрать первые M пар, имеющих наибольшие значения, в качестве более узкого списка кандидатов.

Термином «контекст слова» здесь обозначается сегмент текста, содержащий определенное число слов, включая данное. Под термином «вектор параметров контекста некоторого слова U» понимается N-мерный вектор (N — число слов словаря), каждая i-я координата которого содержит значение меры $TF\text{-}IDF$ i-го слова словаря, если это слово входит хотя бы в один контекст слова U, и значение 0 — в противном случае. Во избежание недоразумений следует отметить, что слова, присутствующие в контексте или контекстах слова U, но отсутствующие в словаре, в векторе параметров контекста слова U никак не отображаются.

$$t\text{-}score(V,W) = \frac{P(V,W) - \frac{P(V) \times P(W)}{N}}{\sqrt{P(V,W)}}.$$

Критическим местом при разработке алгоритмов извлечения информации о переводных соответствиях двух языков на базе параллельных текстов является создание процедуры выравнивания параллельных текстов. И дело здесь не в трудностях алгоритмизации, а в изначальной «непараллельности» параллельных текстов. Смысл этой непараллельности Паскаль Фанг [100] формулирует в виде следующих пяти положений:

1. Слова имеют разные смыслы как в рамках пары параллельных текстов, так и в рамках корпуса в целом.
2. Слова имеют разные переводы как в рамках пары параллельных текстов, так и в рамках корпуса в целом.
3. Некоторые переводные единицы одного текста могут отсутствовать в другом.
4. Частоты вхождения лексических единиц в один и другой текст являются несопоставимыми (т. е. могут существенно различаться).
5. Позиции вхождения лексических единиц, составляющих билингвистическую пару и действительно являющихся соответствующими друг другу, часто являются несопоставимыми (т. е. сильно различающимися).

На основании сформулированных положений она делает вывод о том, что задача извлечения билингвистической терминологии из непараллельных текстов представляет собой менее трудную задачу, чем ее извлечение из параллельных текстов. Возможность эффективного использования непараллельных текстов для извлечения билингвистической терминологии Фанг обосновывает следующими двумя положениями.

1. Для текстов одной и той же предметной области и назначения лексические единицы имеют вполне сопоставимые контексты.
2. Лексические единицы разноязыких текстов одной и той же предметной области, написанных примерно в один и тот же период времени, имеют вполне сопоставимые используемые шаблоны.

Идея представленного в работе алгоритма DKvec является развитием идеи [101], состоящей в использовании контекста слов для измерения с помощью MI- и t-score мер степени их соответствия друг другу.

Цель предлагаемого алгоритма состоит в нахождении переводов (или по крайней мере кандидатов на эту роль) тех слов, которые отсутствуют в online словаре. Кроме словаря, в качестве исходного материала рассматриваются два корпуса текстов по соответствующей тематике, написанных в определенный (по отношению к настоящему) период времени. В рамках корпуса рассматриваются текстовые сегменты (последовательности слов), удовлетворяющие определенным условиям.

Предлагаемый автором алгоритм, по сути дела, представляющий собой фильтрацию списка изначально выбранных кандидатов, предполагает выполнение для каждого слова V, перевод которого в словаре отсутствует, следующих пяти шагов:

1. Составить список слов, входящих во все контексты слова V в рамках корпуса. На основании полученного списка составить вектор параметров контекста слова V.
2. Составить векторы параметров контекста для всех слов W на языке 2, которые являются кандидатами на перевод слова V.
3. Для всех пар (V, W) вычислить значение меры их сходства S.
4. Упорядочить пары по убыванию в соответствии с полученным значением меры сходства.
5. Выбрать первые M пар, имеющих наибольшие значения, в качестве более узкого списка кандидатов.

Термином «контекст слова» здесь обозначается сегмент текста, содержащий определенное число слов, включая данное. Под термином «вектор параметров контекста некоторого слова U» понимается N-мерный вектор (N — число слов словаря), каждая i-я координата которого содержит значение меры $TF\text{-}IDF$ i-го слова словаря, если это слово входит хотя бы в один контекст слова U, и значение 0 — в противном случае. Во избежание недоразумений следует отметить, что слова, присутствующие в контексте или контекстах слова U, но отсутствующие в словаре, в векторе параметров контекста слова U никак не отображаются.

Мера *TF-IDF* является произведением частоты слова (*TF* — term frequency) и обратной частоты документа (*IDF* — inverse document frequency). Она используется для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. В рамках описываемого алгоритма в качестве *TF* рассматривается число контекстов слова *U*, в которых присутствует *i*-е слово словаря, а значение *IDF* определяется по формуле

$$IDF = log\frac{n_{max}}{n_i},$$

где n_i — число контекстов слова *U*, в которых присутствует *i*-е слово словаря, а $n_{max} = max\{n_i\}_{i=1}^{N}$.

Для вычисления степени сходства слова *V* и слов-кандидатов его перевода рассматривается наиболее общая из разновидностей косинусной меры (Cosine Measure)

$$S(V,W) = \frac{|V \times W|}{|V||W|} = \frac{\sum_{i=1}^{N}(V_i \times W_i)}{\sqrt{\sum_{i=1}^{N} V_i^2} \times \sqrt{\sum_{i=1}^{N} W_i^2}},$$

где $V_i = TF_i \times IDF_i$, а $W_i = TF_i \times IDF_i$

Другие разновидности косинусной меры представлены в [7].

Все известные из публикаций алгоритмы, направленные на извлечение из корпусов текстов информации о переводных соответствиях двух языков и создание лексических конкордансов, на настоящий момент решают задачи выбора и максимального сужения билингвистических пар-кандидатов на эту роль.

Развитие теоретической базы направления осуществляется в рамках работ по созданию программных средств конкретного функционального назначения. Краткий обзор основных методов, разработанных в ходе их выполнения в период с 1995 по 2012 г., приведен в [94].

Среди многочисленных мер сходства (или взаимосвязи) лексических единиц выделяются два принципиально различных: скалярный и векторный. Первый основан на значениях частотных характеристиках самих слов (в рамках корпуса, отдельного текста или документа, входящих в его состав сегментов), второй — на значениях характеристик слов, входящих в его контексты.

Методы извлечения терминологии используются для автоматического построения иерархии терминов, которая является промежуточным звеном между текстом и его моделью.

На этом пути кроме сложностей, связанных с отсутствием симметричности терминологических систем разных языков, особую проблему представляет отбор переводов для корпуса параллельных текстов, поскольку их качество часто является сомнительным. Поэтому обращение к псевдопараллельным (сопоставимым) корпусам текстов, при организации которых возможна экспертная оценка текстов на сопоставляемых языках, вполне естественно.

Кроме того, при использовании корпусов сопоставимых текстов вопрос о выравнивании переходит в особую плоскость. В случае параллельных корпусов текстов основным является выравнивание по предложениям, которое опирается на формальные показатели границ и частей предложений, соответствие объемно-прагматических структур текстов. При всех возникающих технических и лингвистических сложностях этот процесс вполне реализуем. В случае текстов сопоставимых возможно только терминологическое выравнивание, опирающееся на выявление характерных для обоих массивов корпуса однословных терминологических единиц и их сопоставление в качестве кандидатов в переводные эквиваленты, а также поиск устойчивых словосочетаний с этими однословными терминами в качестве ядер [14]. Дальнейший сопоставительный анализ требует привлечения знаний из переводных автоматизированных словарей, позволяющих верифицировать выбранные пары терминов. Извлечение многокомпонентных терминов может производиться на основе автоматического синтаксического анализа на уровне функциональных сегментов — именных групп.

Задачей правил синтаксического разбора (независимо от формы их реализации) на разных уровнях анализа предложения является свертка фрагмента (именной или глагольной группы, функционального сегмента, простого предложения и т. п.) до его главного слова [37]. В случае корректного установления границ простых именных групп таким главным (ядерным) словом является крайний правый элемент. В рамках конкретного текста структура именных групп может расширяться как за счет уточнения характеристик ядерного элемента, так и элементов, его определяющих, которые в свою оче-

редь тоже могут последовательно уточняться в рамках именной группы.

В то же время расширение границ именной группы не может быть бесконечным, поскольку приводит к нарушению ее единства, затрудняет понимание и восприятие и специалистом, и переводчиком.

Так, например, выявление кандидатов из корпусов однословных терминов может опираться на семантические характеристики этих слов, извлекаемые из различных автоматизированных баз данных. При использовании в качестве справочного массива словарей предметно-ориентированных систем машинного перевода, словарные статьи которых содержат базовые синтаксические и семантические характеристики выявленных слов, такое соотнесение можно автоматизировать.

3. АСПЕКТЫ СОПОСТАВИТЕЛЬНОГО АНАЛИЗА ТЕРМИНОЛОГИЧЕСКИХ СИСТЕМ

3.1. ИЗВЛЕЧЕНИЕ КАНДИДАТОВ В ТЕРМИНЫ В ЗАДАЧАХ ПЕРЕВОДНОЙ ЛЕКСИКОГРАФИИ

Работы в области переводной лексикографии и в области лингвистического обеспечения систем машинного перевода оказываются в целом нескоррелированными. Создание терминологических баз данных, как правило, опирается на анализ и оцифровывание уже опубликованных словарных источников и (менее) на результаты извлечения терминологии из параллельных и сопоставимых корпусов текстов, но не на те огромные словарные ресурсы, которые накоплены в различных системах МП.

Автоматический словарь системы МП является ее ядерной частью, так как именно на основе заключенной в нем информации реализуется все программное обеспечение лингвистических алгоритмов. При создании практической системы МП отбор лексики в словарь производится на основе опубликованных словарей по соответствующей предметной области, но и прежде всего распределений лексических единиц в представительной выборочной совокупности текстов. При этом, в отличие от словарей «бумажных», словари систем МП постоянно пополняются и модернизируются по результатам работы системы и редактирования ее результатов.

При отборе лексики в словари систем машинного перевода учитывается не только (и не столько) терминологический статус соответствующей лексической единицы — слова или словосочетания (машинного оборота), но ее распространенность в конкретном языке для специальных целей (ЯСЦ). Автоматические словари систем переработки информации, и словари систем МП в том числе, в своей исходной части не являются словарями нормативными, поскольку

в качестве заглавия словарной статьи используются все зафиксированные в обучающем корпусе текстов номинации объектов, процессов и явлений. Перевод при этом соответствует тому, который рекомендован в соответствующем ЯСЦ.

Рассмотрим возможность сопоставления именных машинных оборотов с ядерными словами *method* и *метод* для англо-русского и русско-английского автоматических словарей, ориентированных на подъязык ЯСЦ «Филология» лингводидактика.

Обе эти ЛЕ относятся к общенаучной лексике, могут быть выявлены в любом научном тексте и соответственно в ядерной позиции могут служить опорой для установления кандидатов в термины. В анализируемых АС зафиксировано 34 словосочетания — машинных оборотов с ядерным словом *method*, большинство из которых (20) состоят из 2-х компонентов и создано по модели A+N (см. табл. 5). Если рассматривать варианты перевода самого ядра, то из 34 именных терминологических словосочетаний только в одном из них при переводе английского слова method использован переводной эквивалент *способ*.

При рассмотрении русско-английского варианта автоматического словаря оказывается, что при переводе слова *метод* на английский язык возникает более сложная ситуация. В автоматическом словаре зафиксировано 65 именных словосочетаний с ядерным словом *метод* (табл. 5), что почти вдвое больше количества словосочетаний со словом *method* в ядре.

Таблица 5

Словосочетания с ядерным словом *method*

Термин	Перевод
acceptable word method	метод оценки с учетом слов, соответствующих контексту
adult method	принятый у взрослых метод
alternative method	альтернативный метод
audiolingual method	аудиолингвальный метод
audiovisual method	аудиовизуальный метод
basic direct access method	базисный прямой метод доступа

Термин	Перевод
classroom method	метод организации образовательного процесса
communicative method	коммуникативный метод
comparative method	сравнительно-исторический метод
complementary method of measurement	метод дополнения
cross sectional method	метод поперечного среза
deductive method	дедуктивный метод
ELT method	метод преподавания английского языка
exact word method	метод оценки с учетом только точно восстановленных слов
example-based method	метод перевода по образцам
exhaustive method	метод полного перебора
grammar translation method	грамматико-переводной метод
inductive method	индуктивный метод
interview method	метод интервью
inverse transformation method	метод обратного преобразования
language teaching method	метод преподавания иностранных языков
learning method	метод обучения
linguistic method	лингвистический метод
mass comparison method	метод массового сравнения
natural method	натуральный метод
sampling method	выборочный метод
scoring method	способ подведения результатов
sending questionnaire method	метод рассылки анкет
teaching method	метод преподавания
test method	способ тестирования

Термин	Перевод
testing method	метод тестирования
trial and error method	метод проб и ошибок
trial-and-error method	метод проб и ошибок
variable ratio method	тест на понимание текста с нефиксированным пропуском слов

При этом ЛЕ *метод* соответствуют в переводных эквивалентах английские лексические единицы *method* (32), *approach* (2), *teaching* (5), *instruction* (1), *analysis* (1), *learning* (1), *model(ling)* (2), *fashion* (1), *technique* (3), *programme* (1). Для остальных 16 словосочетаний нет прямого соответствия в переводных эквивалентах, что свидетельствует в том числе и о более широком значении ЛЕ *метод* в русском языке (табл. 6).

Таблица 6

Словосочетания с ядерным словом *метод*

Термин	Перевод
альтернативный метод	alternative method
аудиовизуальный метод	audiovisual method
аудиолингвальный метод	ALM
аудиолингвальный метод	audiolingual method
аудиолингвальный метод	audiolingualism
базисный прямой метод доступа	basic direct access method
выборочный метод	sampling method
грамматико-переводной метод	grammar translation
грамматико-переводной метод	grammar translation method
дедуктивный метод	deductive method
индуктивный метод	inductive method
классический метод	classical approach
коммуникативный метод	communicative method
коммуникативный метод обучения	communicative instruction

Термин	Перевод
коммуникативный метод обучения	communicative teaching
коммуникативный метод обучения языкам	CLT
коммуникативный метод обучения языкам	communicative language teaching
комплексный метод	companion analysis
лингвистический метод	linguistic method
метод «общины»	community language learning
метод «тихого» обучения	silent way
метод «черного ящика»	black box model
метод «черного ящика»	black box modelling
метод анализа в глубину	depth first fashion
метод ведения родительского дневника	parental diary technique
метод дополнения	complementary measurement
метод дополнения	complementary method of measurement
метод закрытия предложения	sentence completion
метод измерений замещением	substitution measurement
метод интервью	interview method
метод Карена	community language learning
метод коммуникативных заданий	task-based learning
метод коммуникативных заданий	TBL
метод массового сравнения	mass comparison method
метод математического моделирования	mathematical model
метод обратного преобразования	inverse transformation method
метод обучения	learning method

Термин	Перевод
метод организации образовательного процесса	classroom method
метод осознания языковой формы	awareness raising technique
метод оценки с учетом слов, соответствующих контексту	acceptable word method
метод оценки с учетом только точно восстановленных слов	exact word method
метод перевода по образцам	example-based method
метод погружения	immersion programme
метод полного перебора	exhaustive method
метод поперечного среза	cross sectional method
метод преподавания	teaching method
метод преподавания английского языка	ELT method
метод преподавания иностранных языков	language teaching method
метод проб и ошибок	trial and error method
метод проб и ошибок	trial-and-error method
метод рассылки анкет	sending questionnaire method
метод с конечным числом состояний	finite state technique
метод с конечным числом состояний	finite state technology
метод синхронизации частоты основного тона	pitch synchronous approach
метод тестирования	testing method
метод усечения дерева	pruning technique
натуральный метод	natural method
педагогический метод	instructional practice
принятый метод работы	working practice
принятый у взрослых метод	adult method
ситуативный метод обучения	situational language teaching

Термин	Перевод
ситуативный метод обучения	SLT
слепой метод исследования	blind study
сравнительно-исторический метод	comparative method
устный метод	oral approach

Системы синтаксических и семантических признаков, используемые при описании лексических единиц (слов и словосочетаний) в различных системах переработки текстовой информации, отличаются друг от друга, поскольку ориентированы на различные алгоритмы парсинга и семантического анализа.

Так, например, набор семантических признаков существительного, разработанный для русского-английского варианта системы Word+, включает 37 базовых единиц (например, антропонимы, свойства и качества, аспекты отношений) и 11 комплексных единиц (например, омоним типа вещество/финансовые параметры). Набор семантических признаков, реализованный для англо-русского варианта системы Word+, совпадает с ним и является единым для обоих вариантов системы. Соответственно, эта информация может использоваться для сопоставления лексических единиц в одноязычных версиях сопоставимого корпуса текстов [ср. 55].

В основе большинства современных разработок лежат идеи извлечения знаний и автоматизации создания баз данных разного типа на основе применения средств извлечения терминов и поиска их переводных эквивалентов. Необходимость использовать весь потенциал информационных технологий на этапах создания и ведения современных лексикографических систем определяется сегодня потребностями научного и технического сообщества, задачами исследования терминологии, развитием множества (более 300) [44, с. 23] языков для специальных целей (ЯСЦ), а также потребностями современной лексикографической работы.

Автоматическое извлечение терминов (как универбов, так и многокомпонентных лексических единиц — коллокаций) основано на предварительном выравнивании текстов на разных языках, идентификации терминологических единиц в текстах на одном языке и дальнейшем установлении их переводных эквивалентов или скорее

кандидатов в возможные переводные эквиваленты. Хотя утверждается, что подобная задача хорошо решена для разных языковых пар в случае анализа параллельных текстов, ее решение еще требует исследований в случае сопоставления языков с различной графикой.

Большинство автоматизированных систем извлечения терминов используют либо статистический, либо лингвистический подход [109; 120]. При этом используется частота лексической единицы в тексте, отношение правдоподобия для двусловных терминов, мера, основанная на полном количестве информации. Для оценки коллокаций, состоящих более чем из двух слов, в качестве единственного статистического параметра используется частота кандидата в термины в корпусе текстов [99].

В последнее время появились гибридные подходы, использование которых представляет собой попытку преодоления ограничений односторонних подходов к решению задачи извлечения терминов на основе как лингвистических, так и статистических элементов [126].

Одним из методов оценки степени терминологичности является независящий от предметной области метод автоматического выявления многокомпонентных терминов в тексте. В качестве исходного материала для анализа при этом используется корпус текстов исходного языка, на основе которого формируется список кандидатов в многокомпонентные термины. Эти термины упорядочиваются по степени терминологичности, которую принято называть C-value. Получаемый в результате список оценивается экспертом в конкретной предметной области. Поскольку кандидаты в многокомпонентные термины оцениваются по степени их терминологичности, эксперт может просматривать списки, начиная от верхней части сверху вниз, работая с ним столько, сколько позволяет время и/или затраты [99].

Подход с использованием C-value основан на объединении лингвистической и статистической информации, причем особое значение имеет именно статистическая компонента. Лингвистическая информация состоит из грамматической разметки корпуса текстов по частям речи, лингвистический фильтр ограничивает тип извлекаемых терминов и включает список стоп-слов (в другой терминологии — антипризнаков).

Лингвистическая база, необходимая для реализации этого метода, включает следующие компоненты:

1. Информацию о части речи, извлекаемую из результатов парсинга и грамматической разметки корпуса текстов.
2. Собственно лингвистический фильтр, применяемый к размеченному корпусу текстов, чтобы исключить те цепочки, извлечение которых не требуется по формальным признакам. К моделям таких цепочек относится неразрешенная комбинаторика частей речи. При этом возможно применение как «закрытого» фильтра, разрешающего извлечение цепочек слов только конкретных типов, так и «открытого» фильтра, в котором перечисляются только неразрешенные типы цепочек.
3. Список слов-антипризнаков.

В рамках собственно статистического анализа исследуются статистические характеристики цепочки лексических единиц, являющихся кандидатами в термины. Необходимость гибридного подхода определяется тем, что доступная для анализа статистическая информация без специальной лингвистической фильтрации не является достаточной для получения достоверных и /или полезных результатов. Так, например, без учета лингвистической информации бессмысленные в терминологичском смысле цепочки слов типа *is a* также будут извлекаться.

Выбор конкретного лингвистического фильтра и его наполнение зависят от того, как терминолог предпочитает сбалансировать полноту и точность: предпочтение точности над полнотой, вероятно, потребует использовать закрытый фильтр, в то время как предпочтение полноты определяет использование открытого фильтра [99, с. 586–587].

Мера C-value строится достаточно прямолинейно на основе характеристик цепочек лексических единиц, являющихся кандидатами в термины. К таким характеристикам относятся:

1. Суммарная частота цепочки лексических единиц, являющихся кандидатами в термины, в корпусе текстов.
2. Частота цепочки лексических единиц, являющихся кандидатами в термины, как часть других, более длинных кандидатов в термины.
3. Количество таких более длинных кандидатов в термины.
4. Длина цепочки лексических единиц, являющихся кандидатами в термины (в количестве слов).

Значение C-value вычисляется в зависимости от длины коллокаций, начиная с самых длинных и заканчивая биграммами.

Большинство лингвистических подходов основано на использовании синтаксических моделей и систем фильтров. Как правило, термины описываются регулярным выражением из меток частей речи, извлекаемых на основе анализа последовательности слов текста. Примером такой системы является система TERMS [108], аналогично работает и система LEXTER. Несмотря на провозглашаемый лингвистический подход, обе системы используют и некоторую базовую статистическую информацию.

Методы, используемые в статистических системах, варьируются от простых подсчетов частот до вычисления сложных статистических индикаторов для измерения силы связи элементов коллокаций, встретившихся в структуре кандидата на роль термина. Основные проблемы, возникающие при применении этих подходов, состоят в том, что частые слова или сочетания слов с высоким индексом связи не обязательно являются терминами. В некоторых статистических подходах привлекаются лингвистические данные, которые должны позволить преодолеть эти ограничения.

Слова и словосочетания, извлеченные подобными системами из предметно-ориентированных корпусов текстов, не всегда релевантны с терминологической точки зрения, хотя могут быть лексическими единицами, зафиксированными в этих корпусах. Поскольку выделяемые лексические единицы не всегда достоверны как термины, их принято называть *term candidates* — кандидатами в термины (КТ).

Чтобы установить, какие кандидаты действительно представляют собой единицы перевода, т. е. для данного типа исследований термины, а какие должны быть отброшены, часто используется методика определения веса термина, позволяющая количественно определить потенциал КТ быть реальным термином. Чаще всего для определения веса ЛЕ и оценки степени ее терминологичности используется подход, основанный на сравнении корпусов. Этот подход, который принято называть Contrastive Automatic Term Extraction [97; 102], представляет собой процедуру автоматического извлечения терминов из сопоставляемых корпусов текстов (контрастивное автоматическое извлечение терминов — КАИТ), подход основан на сравнении

частот кандидатов в термины в двух различных корпусах текстов (общий корпус текстов и специализированный корпус текстов), позволяющем оценить то, что рассматривается как «норма» и фиксируется в национальных корпусах текстов. Количественный анализ этого отклонения и используется как показатель степени терминологичности.

Главными недостатками большинства методов извлечения терминов из текста являются следующие:

1) Неточность результата. Многие из рассматриваемых кандидатов не являются реальными терминами и должны отбраковываться в результате ручного постредактирования получаемого списка. Эта проблема в основном относится к лингвистическим системам и связана с недостаточным и неполным описанием и/или исчислениием синтаксических моделей для выявления и обнаружения терминов.
2) Неполнота результата. Некоторые реальные термины не обнаруживаются системой. Чаще всего эта проблема возникает при применении только статистической информации.

Методы, используемые для извлечения терминов, дают возможность реализовать соответствующую структуру для объединения ряда приемов, обычно используемых для решения этой задачи. Фактически, комбинация множества классификаторов является именно той методикой, которая успешно применяется в ряде задач обработки текстов на естественном языке, например, в задачах грамматической разметки, синтаксического анализа, или классификации и фильтрации текстов, приводя к существенному улучшению результатов, получаемых на основе отдельно применяемых методов [126, с. 516].

Работа терминолога и лексикографа при создании переводных словарей непосредственно связана с осуществлением перевода. Постредактирование результатов МП и получение окончательного варианта перевода текста требует обращения к словарным и энциклопедическим базам данных, выбранным переводчиком, а также к заранее выбранным корпусам текстов. При решении вопроса о выборе перевода конкретной терминологической единицы необходимо привлечение миниконкорданса. В результате работы на этапе собственно перевода должен формироваться пользовательский словарь, характеризующий терминологические особенности конкретного

текста. Этот словарь на этапе поддержки АРМ добавляется в его лингвистические ресурсы.

Рассмотрим возможность реализации гибридного подхода на примере системы, разрабатываемой в Институте прикладной лингвистики университета Pompeu Fabra в Барселоне [126]. Цель создаваемой там системы извлечения терминов состоит в анализе набора текстовых документов в области медицины и создании списка кандидатов в термины (КТ), упорядоченного по степени их терминологичности. Для решения этой задачи предлагается следующая последовательность действий:

1) Построение набора отдельных систем извлечения терминов, работающих на основе различного вида информации;
2) Объединение результатов (т. е. создание набора кандидатов и индекса достоверности для каждого кандидата) применения каждого из методов.

После того как список КТ получен, его можно обработать различными способами в зависимости от предполагаемого использования извлеченных терминов, диапазон следующей обработки может варьироваться от автоматического принятия тех кандидатов, уровень терминологичности которых превышает некоторый порог, до ручного анализа терминологом ТК с лучшими показателями.

В гибридной системе предполагается выполнение работ на 3 основных этапах:

1. Выбор кандидатов в термины. Этот этап состоит в выборе последовательностей единиц, которые могут быть потенциально терминологическими.
2. Анализ выбранных кандидатов в термины. Этот этап состоит из применения ряда процедур выбора терминов, которые должны оценить меру терминологичности кандидатов в термины.
3. Объединение результатов различных систем. На этом этапе объединяются результаты работы различных систем извлечения терминов для того, чтобы произвести окончательный выбор кандидатов в термины.

Сервисы выделения терминов опираются на различные лексические характеристики предметных областей (языков для специальных целей), установленные и зафиксированные в словарях и глоссариях,

частотные характеристики лексических единиц в больших универсальных корпусах текстов и т. п. Рассмотрим некоторые из них [126, с. 518–519].

Принцип анализа семантики контента основан на идее о том, что чем больше вероятность того, что компоненты КТ принадлежат предметной области и входят в лексику языка для специальных целей, тем больше вероятность того, что и сам кандидат в термины является реальным термином. Принадлежность конкретного слова заранее определенной предметной области может опираться на различные глоссарии и словари. Для английского языка и ряда европейских языков наиболее развитым источником информации является EuroWordNet (EWN) — многоязычная лексическая база данных общего назначения, разработанная на основе системы WordNet Принстонского университета и охватывающая испанский, французский и другие европейские языки. Словари WordNet структурированы как наборы лексико-семантических единиц, связанные основными семантическими отношениями. Разработка системы RusNet позволяет предположить возможность ее использования для решения тех же задач для терминологического анализа текстов на русском языке.

Подобные системы используются для того, чтобы определить, принадлежит ли слово конкретной предметной области. В системе EWN есть несколько важных ограничений, связанных с тем, что эта система является универсальной онтологией и в ней недостаточно информации, связанной с принадлежностью слова конкретным предметным областям.

Однако в случае конкретной предметной области этот недостаток можно преодолеть. Например, в условиях работы с предметной областью «медицина» недостаток информации о предметной области слова компенсировался тем, что изначально было установлено и маркировано около 30 медицинских подобластей. Затем было принято предположение о том, что все гипонимы, которые попадают в эти подобласти, принадлежат области медицины. Так, можно сказать, что *disease* (болезнь) составляет подобласть медицины, поскольку все болезни, зарегистрированные в EWN, являются гипонимами этого синсета.

Для анализа многозначных лексических единиц вычисляется специальный коэффициент принадлежности к предметной области «ме-

дицина» (КПМ), предназначенный для измерения степени терминологичности каждого кандидата в термины. Этот коэффициент может вычисляться достаточно прямолинейно как отношение между количеством «медицинских» значений и общего числа значений слова, зарегистрированных в EWN. Несмотря на простоту и элементарность, этот способ вычисления коэффициента работает достаточно хорошо. Этот коэффициент используется для определения порогового значения принадлежности слова к области медицины. Предполагается, что любое существительное с коэффициентом КПМ выше порогового имеет значение для предметной области «медицина». Этот коэффициент может рассматриваться как мера специализации слова.

Кроме этого в системе реализован принцип анализа греческих и латинских форм. Отраслевые словари и глоссарии включают слова, которые могут быть разложены на греческие и латинские компоненты. Эта характеристика слова важна для извлечения терминологии, поскольку подобные сложные слова, как правило, специализированы и не встречаются в общей лексике. Такие слова достаточно легко можно разложить на части и вычислить значение выделенных частей. У этого метода анализа точность достаточно высокая, хотя и очень ограниченный охват. Следовательно, его необходимо объединять с другими методами извлечения терминологии.

Принцип анализа контекста также представляет важную составляющую описываемого метода, его основой являются следующие положения:

a) слова, окружающие «основной» КТ (т. е. уже известные термины или КТ с высоким баллом), могут быть полезными ключами для других терминов, и

b) те контекстно-связанные слова, которые являются «основными» КТ и семантика которых подобна семантике рассматриваемых кандидатов в термины, дают дополнительную информацию, подтверждающую терминологичность КТ.

Таким образом, вычисляемый в процессе анализа коэффициент контекста (КК) состоит из двух компонентов: лексического коэффициента контекста и семантического коэффициента контекста. Лексический КК определяется словами, которые окружают «основных» кандидатов, а семантический КК зависит от концептуального расстояния между КТ и «основными» терминами, появляющимися в их контексте. На этом этапе анализа необходимо выбрать:

1) «основного» кандидата в термины;
2) окно контекста;
3) соответствующую меру подобия.

«Основными» кандидатами в термины считаются кандидаты, имеющие максимальное значение коэффициента контекста (вместо этого можно использовать исходный список «истинных» КТ). При выборе конкретных значений параметров и коэффициентов терминологу или лексикографу приходится сначала экспериментировать с разными размерами окон контекста и различными относительными весами лексических и семантических коэффициентов.

Для оценки кандидатов можно использовать традиционные статистические методы, использующие различные меры ассоциации между словами, входящими в многословные термины. Интуитивно понятно, что два компонента, между которыми существует тесная связь в предметной области, формируют термин. Как правило, в описываемой системе используются три критерия связи (логарифмическая функция правдоподобия, взаимное количество информации и взаимное количество информации в кубе).

Ни один из этих методов не накладывает ограничения на длину и состав КТ, следовательно, все они могут применяться в любой модели анализа терминологии.

Различные методы извлечения терминов из текстов в принципе доказали свою работоспособность, но результаты их использования в значительной степени зависят от корректности выбранных текстов. При создании переводного словаря на основе корпуса параллельных текстов необходимо:

- определить принципы формирования выборочной совокупности для создания исследовательского параллельного корпуса текстов и ее необходимый и достаточный объем;
- установить требования к разметке текстов в корпусе и получить базовую лингвистическую информацию;
- определить необходимые для работы средства информационных технологий, включая системы машинного перевода, переводческой памяти, средства выравнивания параллельных текстов, редакторы разметки (тэггирования), средства формального извлечения терминов из текста и т. д.

Средства информационных технологий могут быть ориентированы на выбор и обработку терминов и/или понятий, на работу с конкретной языковой парой, многоязычными или одноязычными ресурсами. Подобные средства целесообразно включать в автоматизированное рабочее место лексикографа, на основе которого можно осуществлять запись, обработку, сохранение и использование различных лингвистических и лексикографических данных.

В современной корпусной лингвистике принято различать корпуса параллельных текстов (parallel corpora) и корпуса псевдопараллельных текстов (comparable corpora). Их различие, кроме всего, связано с принципами отбора текстов. В случае псевдопараллельных текстов их отбор может осуществляться на основе достаточно ясных критериев. При создании корпуса параллельных текстов для последующего терминологического анализа и извлечения пар типа термин–перевод важным условием «успешности», адекватности созданного ресурса является качество выбранных переводов. При создании подобного корпуса для других целей качество перевода не столь важно.

Соответственно, проблема оценки качества перевода оказывается кардинальной для формирования корпуса параллельных текстов, критериями такой оценки могут быть:

- последовательность использования номинаций в тексте перевода;
- соблюдение переводчиком норм языка перевода;
- сохранение логической структуры исходного текста и, в целом;
- экспертная оценка.

Рассмотрение корпуса параллельных и/или псевдопараллельных текстов в качестве лексикографической базы предполагает необходимость его дополнения корпусом машинных переводов текстов, что позволяет явным образом выделить те лексические единицы, которые должны быть введены в словарь или перевод которых требует модификации.

Приемы извлечения терминологии из параллельных корпусов текстов основаны на исследовании дистрибутивной семантики, т. е. на гипотезе о том, что у семантически близких слов есть тенденция появляться в одних и тех же контекстах [100]. В общем случае идея идентификации переводов терминов в сопоставимых корпусах текстов осуществляется в три этапа. Для реализации идей дистрибутивной семантики, кроме корпуса сопоставимых текстов, формируется база обучающей информации, включающая базовый переводной словарь общенаучных терминов и частотных терминов конкретной предметной области.

На первом этапе производится автоматическое вычисление контекста каждого кандидата в термины в исходном корпусе текстов и в корпусе текстов на языке перевода. Контекст термина Ò описывается вектором, указывающим число раз, когда Ò появляется совместно с каждым конкретном словом W_c в данном контекстном окне. Величина контекстного окна зависит от типа языка, но чаще всего составляет контекст длиной 7 слов: три слова слева и три слова справа.

На втором этапе слова в векторах исходного контекста переводятся на язык перевода на основе исходного двуязычного словаря. На третьем этапе сравниваются векторы исходного языка и языка перевода: чем более подобны друг другу векторы, тем вероятнее, что фрагмент текста на переводном языке и термины являются переводами друг друга. Вопрос о степени подобия в разных системах определяется эвристически. В конечном счете результатом работы алгоритма выравнивания является список кандидатов в переводные эквиваленты типа «один ко многим»: каждое входное слово связано с упорядоченным списком вариантов перевода, которые упорядочиваются от самого вероятного до наименее вероятного. Результаты оцениваются путем исследования лучшего варианта перевода (Top 1), десяти лучших вариантов перевода (Top 10) или двадцати лучших вариантов перевода (Top 20) [96].

Недостаток извлечения терминологии из сопоставимых корпусов текстов состоит в том, что полученные лексиконы не так надежны, как те, что извлечены из параллельных текстов. Извлечение лексикона из параллельных текстов в идеале дает одно-однозначное выравнивание терминов с высокими показателями точности. Например, в исследовании [113] приведены оценки с точностью от 85% до 90%. Напротив, системы, которые извлекают лексиконы из сопоставимых корпусов текстов, дают в результате выравнивание типа «один ко многим»: входная ЛЕ, связанная с набором наиболее вероятных переводных эквивалентов. Как следствие, полученные списки

кандидатов в термины и их переводы должны подвергаться постредактированию до введения в базу терминов или в модуль обработки любого другого языка. Например, при выравнивании по однословным терминам на двадцати самых вероятных кандидатах, вычисленных из больших генеральных корпусов текстов объемом в 100 млн слов или более [100], достигается точность до 80%.

Существующие инструментальные средства валидизации выравнивания терминов не оценивают подобные результаты. Например, применение программного продукта iView из комплекта iTools [118] предлагает пользователю список одно-однозначных соответствий терминов, по отношению к которым ему следует оценить, настолько эти пары правильны и насколько они характерны для предметной области. Дополнительная информация о парах терминов состоит из предложений, выбранных из текстов, и некоторых статистических данных.

Подобным же образом работает программный продукт фирмы Heartsome Europe GmbH Araya Bilingual Term Extractor [121], являющийся частью переводческого инструментария Araya Server Translation Tool. Программа создает список одно-однозначных соответствий, сопровождаемых оценками качества: парам приписываются вероятности от 1 (максимальное правдоподобие, что пара связана отношением переводческой эквивалентности) до 0,5 (минимальное правдоподобие, что пара связана отношением переводческой эквивалентности). Кроме того, в результате анализа двух текстов, заранее представленных в формате UN8, можно получить информацию о числе сегментов в исходном и переводном тексте, в которых появляются и исходный и переводной компонент пары. В качестве сегмента может использоваться как отдельное предложение, так и фрагмент текста. Кроме того, предоставляется информация о том, в скольких сегментах исходного и переводного текста встретились исходный термин и его перевод [121].

При работе системы учитываются следующие параметры, задаваемые пользователем. К ним относятся:

- Минимальное/максимальное количество слов в составе термина.
- Минимальная/максимальная частота термина в массиве.
- Максимально возможное количество фиксируемых переводов.

- Возможность представления терминов в нижнем регистре.
- Игнорирование уже установленных терминов.

В процессе работы программы формируется специальный файл для извлеченных и проверенных пар терминов, что дает возможность не учитывать их при анализе новых текстов. Однако остается не ясным, что происходит, когда уже установленный термин входит в новое словосочетание как его часть.

Программа XTerm, также предназначенная для извлечения терминов и их переводов, предлагает пользователю 2 сервиса [93, с. 118]:

- извлечение терминологии из текстов документов. Эта процедура реализована для извлечения терминов из технической документации на французском, русском и английском языках. Пользователь имеет возможность отбора предложенных ему системой терминов, используя графический интерфейс;
- навигация по тексту на основе выявленных терминов. Эта процедура дает возможность пользователю увидеть термин или, скорее, кандидата в термины, в контексте предложения или в более широком контексте.

В задачи, решаемые системой Xterm, входят следующие:

- анализ корпусов текстов и формирование терминологических ресурсов, имеющих иерархическую структуру;
- обеспечение эксперта (переводчика, лексикографа, терминолога) возможностью оценивать конкретные терминологические единицы;
- порождение иерархии классов в формальной модели и передача их в сервер знаний Troeps на основе потока XML;
- обеспечение обратного перехода от модели к документам [93].

Извлечение терминологии из сопоставимых корпусов текстов определило потребность в новом виде инструмента для валидизации терминологии. Этот новый лингвистический инструмент должен анализировать лексические соответствия типа «один ко многим» или даже «многие со многими», если термин исходного языка представляет собой набор словосочетаний, отражающих несколько вариантов одного и того же термина. Кроме того, предполагается [96], что характеризующие контекст предложения и статистические данные являются важной, но недостаточной информацией для того, чтобы помочь терминологу или профессиональному переводчику оценить

кандидатов в термины и их переводы должны подвергаться постредактированию до введения в базу терминов или в модуль обработки любого другого языка. Например, при выравнивании по однословным терминам на двадцати самых вероятных кандидатах, вычисленных из больших генеральных корпусов текстов объемом в 100 млн слов или более [100], достигается точность до 80 %.

Существующие инструментальные средства валидизации выравнивания терминов не оценивают подобные результаты. Например, применение программного продукта iView из комплекта iTools [118] предлагает пользователю список одно-однозначных соответствий терминов, по отношению к которым ему следует оценить, настолько эти пары правильны и насколько они характерны для предметной области. Дополнительная информация о парах терминов состоит из предложений, выбранных из текстов, и некоторых статистических данных.

Подобным же образом работает программный продукт фирмы Heartsome Europe GmbH Araya Bilingual Term Extractor [121], являющийся частью переводческого инструментария Araya Server Translation Tool. Программа создает список одно-однозначных соответствий, сопровождаемых оценками качества: парам приписываются вероятности от 1 (максимальное правдоподобие, что пара связана отношением переводческой эквивалентности) до 0,5 (минимальное правдоподобие, что пара связана отношением переводческой эквивалентности). Кроме того, в результате анализа двух текстов, заранее представленных в формате UN8, можно получить информацию о числе сегментов в исходном и переводном тексте, в которых появляются и исходный и переводной компонент пары. В качестве сегмента может использоваться как отдельное предложение, так и фрагмент текста. Кроме того, предоставляется информация о том, в скольких сегментах исходного и переводного текста встретились исходный термин и его перевод [121].

При работе системы учитываются следующие параметры, задаваемые пользователем. К ним относятся:

- Минимальное/максимальное количество слов в составе термина.
- Минимальная/максимальная частота термина в массиве.
- Максимально возможное количество фиксируемых переводов.

- Возможность представления терминов в нижнем регистре.
- Игнорирование уже установленных терминов.

В процессе работы программы формируется специальный файл для извлеченных и проверенных пар терминов, что дает возможность не учитывать их при анализе новых текстов. Однако остается не ясным, что происходит, когда уже установленный термин входит в новое словосочетание как его часть.

Программа XTerm, также предназначенная для извлечения терминов и их переводов, предлагает пользователю 2 сервиса [93, с. 118]:

- извлечение терминологии из текстов документов. Эта процедура реализована для извлечения терминов из технической документации на французском, русском и английском языках. Пользователь имеет возможность отбора предложенных ему системой терминов, используя графический интерфейс;
- навигация по тексту на основе выявленных терминов. Эта процедура дает возможность пользователю увидеть термин или, скорее, кандидата в термины, в контексте предложения или в более широком контексте.

В задачи, решаемые системой Xterm, входят следующие:

- анализ корпусов текстов и формирование терминологических ресурсов, имеющих иерархическую структуру;
- обеспечение эксперта (переводчика, лексикографа, терминолога) возможностью оценивать конкретные терминологические единицы;
- порождение иерархии классов в формальной модели и передача их в сервер знаний Troeps на основе потока XML;
- обеспечение обратного перехода от модели к документам [93].

Извлечение терминологии из сопоставимых корпусов текстов определило потребность в новом виде инструмента для валидизации терминологии. Этот новый лингвистический инструмент должен анализировать лексические соответствия типа «один ко многим» или даже «многие со многими», если термин исходного языка представляет собой набор словосочетаний, отражающих несколько вариантов одного и того же термина. Кроме того, предполагается [96], что характеризующие контекст предложения и статистические данные являются важной, но недостаточной информацией для того, чтобы помочь терминологу или профессиональному переводчику оценить

соответствие терминов. Не менее важно условие, что необходим формат обмена данными, приспособленный для обмена автоматически созданными лексиконами.

Использование корпусов текстов для лексикографических исследований и работ по автоматизации извлечения терминов привели к тому, что в переводной лексикографии возобладало представление о том, что задача дефинирования термина не является первостепенной, поскольку отношение термин/понятие не всегда определено, и терминам конкретной предметной области свойственно меняться и по структуре, и по составу. Соответственно, при установлении пар терминов в системах извлечения информации характеристика типа термин/терминоид/номен не определяется, задачей является описание употребления и значения ЛЕ, а не формирование его дефиниции.

Таким образом, можно утверждать, что результатом установления переводных терминов [96] является список пар лексических единиц, извлеченных из корпуса текстов. По отношению к этим парам предполагается, что они являются правильными переводами друг друга и тем самым в переводческом аспекте рассматриваются как переводческие эквиваленты. Соответственно, в результате работы подобных систем пользователь получает «сырую» информацию, извлеченную из сопоставимого или параллельного корпуса текстов. Эта информация может быть извлечена в результате выравнивания терминов и тогда является дополнительным продуктом этого процесса, или она может быть восстановлена из специальных предметно-ориентированных ресурсов или сетевых ресурсов общего пользователя типа Wikipedia или Wiktionary. Тогда подобная информация предназначена для помощи комментатору при анализе поведения анализируемых терминов.

В предоставляемую пользователю (лексикографу, терминологу, переводчику) информацию включается:

- нормализованная форма термина (обычно результат лемматизации, что при извлечении многосоставных терминов из флективных языков может приводить к неясности синтаксической структуры и некорректному установлению значения термина и его возможного перевода);
- информация о части речи с точностью до снятой конверсионной омонимии;
- частота или количество вхождений в корпус текстов;
- дефиниция, извлекаемая из словаря системы или сетевых ресурсов;
- многословные сочетания с указанным термином в ядре;
- «мягкие» варианты типа акронимов и орфографических вариантов;
- более сильные варианты типа синтаксических или морфосинтаксических вариантов;
- термины, у которых одинаковая основа с описываемым;
- термины, которые появляются в сходных контекстах;
- контексты, в которых появляется термин: раздел текста с указанием связи, ведущей к исходному документу [96].

Термины, извлекаемые различными автоматизированными системами, естественно представляют собой некоторый «первопродукт», требующий дальнейшего анализа.

3.2. СОПОСТАВИТЕЛЬНЫЙ АНАЛИЗ АНГЛО-РУССКОЙ И РУССКО-АНГЛИЙСКОЙ ТЕРМИНОЛОГИЧЕСКИХ СИСТЕМ

Несмотря на то что термин является центральным элементом, интересующим терминолога и переводчика, работающего в каждой конкретной предметной области, – термин является единицей словаря, объектом перевода и извлекаемым из корпуса элементом текста, при создании терминологической базы данных необходимо соотносить получаемые термины с концептуальным описанием данной предметной области, как это реализуется, например, в рамках многоязычного проекта EuroTermBank [124]. В качестве универсальных описаний в больших проектах используются, например, тезаурус Eurovoc или онтологии, включающие уже несколько тысяч понятий (ср. OMEGA, SUMO, DOLCE и др.). Подобное описание, созданное для каждого из рабочих языков, позволяет выделить ядерные понятия исследуемой области, а также оценить сбалансированность корпуса текстов, выделить зоны, требующие особого внимания в связи с появлением новых понятий. Последующая гармонизация терминологий может осуществляться по схеме, реализуемой в рамках проекта Штутгартского университета TTC (рис. 2) [125].

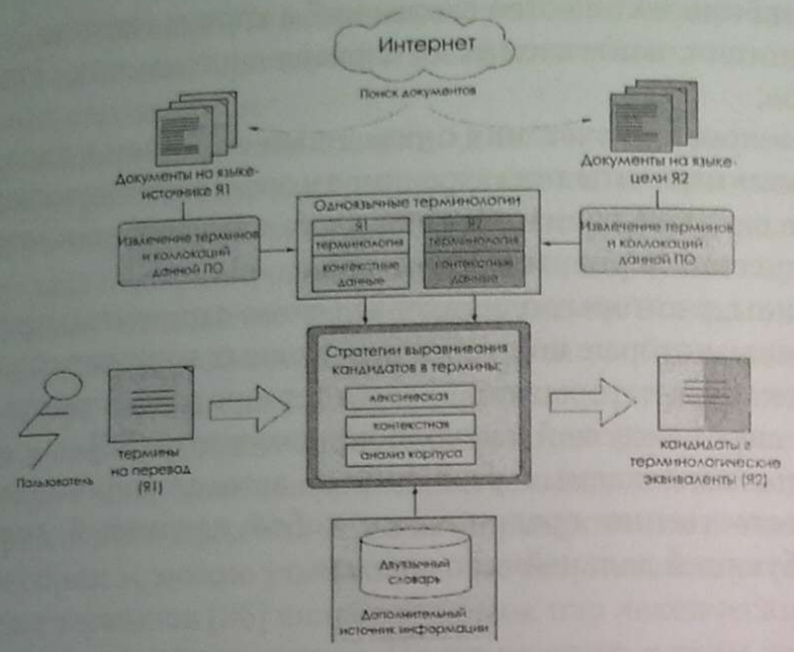


Рис. 2. Структура процесса гармонизации терминологии

Из текстов на исходном языке, которыми располагает переводчик или исследователь, можно извлечь начальный список слов для дальнейшего целенаправленного поиска документов, которые составят основу корпуса (или нескольких корпусов) текстов. Такая же процедура проводится и для текстов на языке перевода. Данный начальный список слов может также послужить основой терминологической базы и онтологии для конкретных предметных областей. При автоматическом выделении опорных терминоэлементов следует с осторожностью относиться к критерию частотности, так как в верхней части списка могут оказаться не только строевые элементы текста, но и общенаучные или общетехнические лексические единицы, «ненужность» которых довольно сложно заранее вычислить и указать в стоп-листе.

Если набор исходных текстов небольшой, то многие термины, подходящие на роль опорных для данной предметной области, могут оказаться низкочастотными. Эта проблема может быть решена путем разбиения частотного списка на зоны и путем их сравнения на предмет выявления релевантных терминоэлементов. Например, при построении онтологии по предметной области «лингводидактика» в результате наблюдений над частотностью слов в трех подкорпусах, в качестве наиболее перспективной была выделена среднечастотная зона, а наиболее и наименее частотные зоны списка отсеяны [1, с. 61]. Для выравнивания терминологий в проекте TTC используется комбинация статистических и лингвистических методов, с помощью которых лексические единицы, извлеченные из текстов на обоих языках, группируются в эквивалентные пары. При этом этапы интернет-поиска и извлечения терминов полностью автоматизированы, а последующие этапы автоматизированы частично [125].

Традиционными пользователями англо-русских и русско-английских терминологических словарей в предметной области «филология» являются специалисты в гуманитарных областях знаний и студенты соответствующих направлений подготовки. Некоторые особенности устройства этих словарей отражают стремление преодолеть неудобство существующего бумажного формата. Как правило, применяется алфавитно-гнездовая система, ориентирующая пользователей словарей в системных связях между терминами. Для удобства поиска специализированные словари снабжаются указателем русских терминов, связанных цифровыми обозначениями с основной частью словаря.

Электронный формат представления словарной информации позволяет существенно изменить микроструктуру словаря – словарную статью. Целесообразно структурировать словарную статью «концентрически», с возможностью ступенчатого извлечения информации. Минимальный блок информации должен содержать информацию о произношении, вариантах написания данного термина в американском и британском вариантах английского языка, семантизацию заголовочной единицы, краткую информацию о системных отношениях и сочетаемостную характеристику. При помощи гиперссылок отдельные элементы минимального блока должны обеспечивать доступ пользователя к другим словарным статьям, связанным с данной статьей, а также к более детальной информации о системных отношениях термина, к классификациям номенклатурных единиц, графическому иллюстративному материалу. В случае использования близких по значению терминов в отношении одних и тех же номенклатурных единиц в словарной ста-

тье к подобным терминам необходима «экспертная справка» об использовании вариантов термина.

При сопоставительном анализе корпусов сопоставимых текстов выявляется несколько «подводных» камней. Во-первых, «английские» тексты в своем большинстве написаны на глобальном английском языке, что в реальности означает нарушение синтаксической структуры предложения, вызванное влиянием родных языков, и отсутствие гармонизации терминологии, в результате чего термины часто представляют собой переводы соответствующих лексических единиц (ЛЕ) родного языка автора, а не стандартизированные номинации. «Русские» тексты, в свою очередь, «отягощены» безумным научным канцеляритом, частотным использованием синтаксических структур с объектом в первой позиции предложения и отсутствием явных границ между именными группами, номинирующими термины и выполняющими разные роли в предложении. Рассмотрим это на примере предложения из статьи на русском языке (табл. 7).

Таблица 7

Пример работы со сложным предложением из научной статьи на русском языке

Исходное предложение	**Машинный перевод системой Word+**	**Отредактированный вручную перевод**
Методика состоит из трех этапов:	Technique consists of three phases:	The technique consists of three phases:
1) поиск оптимальных параметров вязкого демпфирования для линейной динамической модели с одной степенью свободы (ЛДМ);	1) search of optimal parameters of viscous damping for linear dynamic model with single degree of freedom (ЛДМ);	1) searching the optimal parameters of viscous damping for linear dynamic model (LDM) with single degree of freedom;

Исходное предложение	**Машинный перевод системой Word+**	**Отредактированный вручную перевод**
2) **построение соответствующих различным конструктивным параметрам семейства силовых характеристик упругопластических демпферов;**	2) construction of force characteristic family corresponding various design data of elastic plastic dampers;	2) modelling the elastic plastic dampers force characteristic family corresponding to various design values;
3) выбор силовой характеристики демпфера, обеспечивающей демпфирование, эквивалентное оптимальному вязкому демпфированию.	3) selection of force characteristic of damper, ensuring damping, to equivalent to optimal to viscous damping.	3) choosing the force characteristic of damper, ensuring the damping equivalent to optimal viscous damping.

Во фрагменте этого предложения, выделяемой системой машинного перевода (МП) в качестве функционального сегмента (построение соответствующих различным конструктивным параметрам семейства силовых характеристик упругопластических демпферов), построение ЛЕ представляет собой общенаучную лексическую единицу. Весь выделенный фрагмент является синтаксической загадкой как для системы МП, так и для человека-переводчика, поскольку его можно трактовать либо как:

- построение семейства силовых характеристик, соответствующих различным конструктивным параметрам упругопластических демпферов,

либо как

- построение упругопластических демпферов, соответствующих различным конструктивным параметрам семейства силовых характеристик.

Правильное решение может принять только специалист в соответствующей предметной области на основе своего профессионального тезауруса.

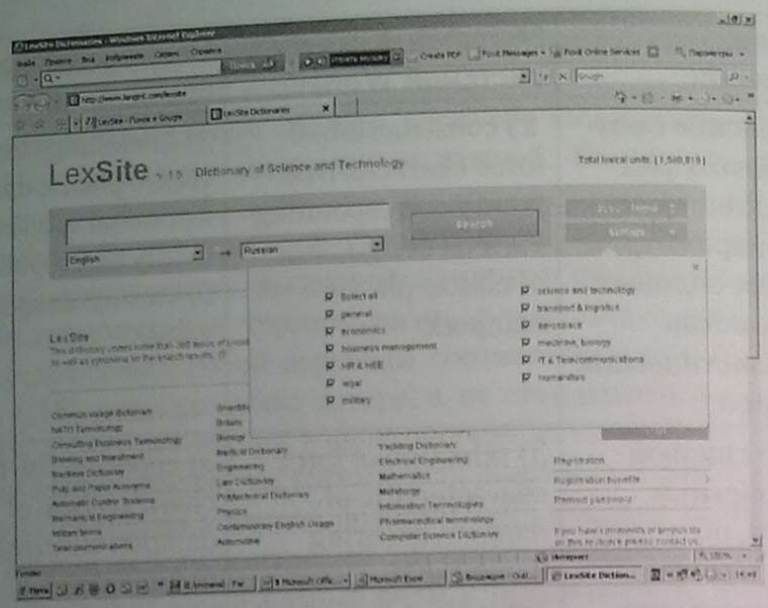


Рис. 3. Общая структура интерфейса системы LexSite

Наиболее популярными и наиболее мощными англо-русскими сетевыми словарями являются Мультитран, ABBYLingvo и Мультилекс [43, с. 151–153], причем Мультитран превосходит остальные ресурсы по наполнению, но уступает им в быстродействии [43, с. 154].

Сетевой словарь LexSite [75] представляет собой открытый лексический ресурс, основанный на базе данных компании Language Interface. Эта база содержит общую и специальную лексику, а также уникальные термины, накопленные в результате работы над проектами в различных областях знаний. Корпус параллельных текстов, составляющий второй компонент базы данных, включает пары типа оригинал-перевод для российских нормативных и юридических документов [43, с. 156]. Результат поиска ранжирован по 300 предметным областям, из которых 13 являются базовыми (рис. 3, 4). Словарные статьи снабжены синонимической информацией, которая выводится на экран по запросу пользователя.

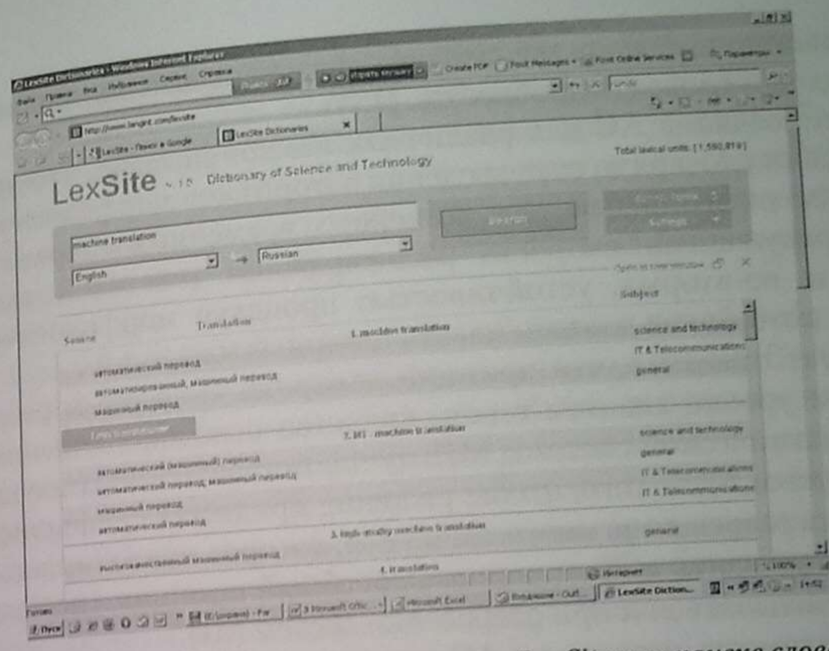


Рис. 4. Пример интерфейса системы LexSite при поиске слова

3.3. АВТОМАТИЧЕСКИЙ СЛОВАРЬ СИСТЕМЫ МАШИННОГО ПЕРЕВОДА КАК БАЗА ПЕРЕВОДНОЙ ЛЕКСИКОГРАФИИ ДЛЯ ПОДЪЯЗЫКА ФИЛОЛОГИИ

Автоматические словари (АС) систем машинного перевода представляют собой базу данных, используемую при реализации алгоритмов перевода на всех его уровнях. Поэтому правильный выбор состава и структуры АС во многом определяет результат работы системы в целом.

Необходимость настройки АС системы МП на конкретную предметную область признается практически всеми разработчиками систем МП. Столь же несомненным является требование к объему АС, обеспечивающему достаточно полное распознавание элементов текста (как минимум 90 % лексических единиц текста должны быть правильно идентифицированы). Естественно, объем АС зависит от типа (уровня аналитизма) языка и реализованных в системе алгоритмов морфологического анализа.

Эти алгоритмы определяют и способы хранения информации в АС, и структуру самого автоматического словаря — включение

в нее машинных основ, словоформ, фрагментов слов или единиц, больших чем слово (подробнее см. [13]).

Опыт создания АС для различных языков показал, что уровень компрессии словаря за счет введения алгоритмов морфологического анализа должен определяться, во-первых, временными параметрами, характеризующими скорость распознавания текстовой единицы по АС, и, во-вторых, устойчивостью процедур морфологического анализа, которые не должны давать неверных членений.

Следует отметить, что принцип контроля «неверных решений» на любом уровне анализа текста является основополагающим для любой практической системы МП: лучше оставить многозначный вариант перевода для того, чтобы решение принимал специалист, чем предлагать разрешение многозначности, которое может оказаться неверным и исказить информацию. Врачебный принцип «не навреди» должен учитываться и при разработке систем МП.

Необходимость введения в АС как отдельных лексических единиц, так и их сочетаний, также признается всеми как данность. Проблема создания и использования системой словарей словосочетаний (автоматических словарей оборотов — АСО) в другом: какова типология таких сочетаний, как они должны храниться в АС, как и когда отдельные обороты (или типы оборотов) должны распознаваться при анализе текста.

С типологической точки зрения словосочетания (машинные обороты) можно описать как:

- собственно иконические обороты, т. е. неизменяемые конструкции, перевод и функциональные характеристики которых не зависят от контекста (сюда относятся во всех языках сложные предлоги, союзы, наречные конструкции и т. п.);
- иконические изменяемые обороты, представляющие собой линейно неразрывные последовательности, перевод и функциональные характеристики которых не зависят от контекста, а определяются синтаксической функцией в предложении. Этими оборотами чаще всего являются термины;
- условно иконические обороты, представляющие собой неизменяемые конструкции, перевод и функциональные характеристики которых зависят от контекста, в частности наличия знака препинания, выделяющего вводную конструкцию (например, «в целом» (рус.) — «as a whole» (англ.) и т. д.);

- разрывные обороты, представляющие собой последовательность лексических единиц, между которыми могут «вклиниваться» другие элементы и даже конструкции. В языках со свободным порядком слов элементы таких оборотов могут занимать произвольную позицию в тексте, по своему морфологическому типу эти обороты могут быть изменяемыми и неизменяемыми;
- обороты с эмфазой, когда часть оборота в какой-либо определенной форме физически не присутствует в тексте. В частности, это касается оборотов с глаголом «быть» в русском языке, где сам глагол-связка в настоящем времени в тексте отсутствует, например, оборот «быть неисправным» должен переводиться «be in default» (англ.), соответственно предложение «Прибор неисправен» должно переводиться как «The device is in default». Эти обороты могут также быть как разрывными, так и инверсионными.

Естественно, что на уровне первичного анализа текста могут выделяться только собственно иконические и иконические изменяемые обороты, три оставшиеся типа должны специально анализироваться и распознаваться на различных уровнях анализа.

Автоматический словарь системы МП является ее ядерной частью, так как именно на основе заключенной в нем информации реализуется все программное обеспечение лингвистических алгоритмов. Поэтому особое внимание уделяется структуре информации, приписываемой каждому элементу системы, и способу ее хранения в словарной статье (СлСт). Опыт разработки практических систем МП показал нецелесообразность разработки СлСт *ad hoc*, сразу и для всех возможных ситуаций. При всей привлекательности для лингвиста выполнения процедуры «портретирования» слова в реальности в СлСт должна включаться та и только та информация, использование которой связано с реализацией конкретных алгоритмов.

Такой подход, конечно, должен сочетаться с созданием условий для свободного пополнения и модификации любой СлСт и любого элемента в ней, но это уже вопрос сервиса.

При создании практической системы МП целесообразно различать собственно лингвистическую информационную базу (ЛИБ), включающую удобно организованный словарь и средства его орга-

низации и развития, ориентированный на задачи, решаемые лингвистами-разработчиками системы, и словарные файлы, создаваемые как результат программно выполняемой конверсии ЛИБ в формат, предназначенный для использования программами системы МП. При таком подходе формат словарных файлов может меняться по мере развития системы, а ЛИБ (при достаточном сервисе) может развиваться автономно, без обязательной перестройки каждый раз, когда меняется структура файлов для системы перевода.

Машинный перевод представляет собой ту область, в которой квантитативные методы могут использоваться как на этапе проектирования практической системы МП, так и на этапе исследования процесса перевода. При создании практической системы МП квантитативный анализ должен использоваться для решения задач создания словаря:

- определения структуры и состава самих АС и словарных статей на основе статистического исследования распределений слов в текстах конкретной предметной области;
- выбора основной терминологии, которая должна включаться в АС на основании изучения распределений в представительной выборочной совокупности текстов.

При отборе лексики в словари систем МП учитывается не только терминологический статус [85] лексической единицы (единицы перевода) — слова или словосочетания (машинного оборота), но и ее распространенность в конкретном языке для специальных целей. АС в своей исходной части не является словарем нормативным, поскольку в качестве заглавия словарной статьи используются все встречающиеся варианты номинации объектов, а перевод соответствует рекомендуемому для языка перевода.

Следовательно, специализированные словари систем машинного перевода отражают состояние терминологии более корректно, чем словари бумажные. На основе этого предположения была получена лексикографическая версия автоматического словаря системы машинного перевода WORD+, предназначенного для перевода текстов предметной области «филология». После исключения из словаря словарных статей для служебной лексики и общенаучной лексики был получен словарь объемом 15 680 словарных статей.

Все единицы словаря имеют семантические коды, соответствующие принятой в системе параметризации.

Выделенный словарь был исследован с точки зрения распределения типов лексических единиц в сопоставлении структур английских и русских терминов. Полученный материал подтверждает несоответствие размерности не только синонимических рядов [24, с. 171] терминов и терминологических словосочетаний английского и русского языка филологии, но и несоответствие структурное, связанное с более расширенным переводом ЛЕ на русский язык. Общее распределение структур показано в табл. 8, частотные характеристики представлены в табл. 9.

Лексикологическое несоответствие объемов понятий (ср., например, термины *навык* и *умение* в русскоязычной лингводидактике и обобщающий термин *skills* в английской), а также системно-структурные расхождения, связанные с компрессией номинации в английском языке vs. явной выраженности всех сем в русской терминологии, можно рассматривать как системный диссонанс, свойственный терминологии филологии (ср. [24, с. 172]).

Таблица 8

Сводная информация о соотношении структур английских и русских терминов (фрагмент)

Часть речи английского термина	Длина в словоформах	Часть речи русского соответствия	Длина в словоформах	Количество терминов такого типа	Пример английского термина	Пример русского термина
A	1	A	1	1668	achronistic	ахронический
A	1	A	2	22	comprehensible	доступный усвоению
A	1	A	3	8	antepenultimate	третий от конца
A	1	G	4	1	postconsonantal	следующий за согласным звуком

Часть речи английского термина	Длина в словоформах	Часть речи русского соответствия	Длина в словоформах	Количество терминов такого типа	Пример английского термина	Пример русского термина
A	1	P	5	2	adultocentric	ориентированный на изучение языка взрослых
A	1	Ap	1	1	developmental	развития
A	1	Ap	2	14	preverbal	перед глаголом
A	1	Ap	3	7	observational	на основе наблюдений
A	1	Ap	4	1	postsecondary	после окончания средней школы
A	1	Ap	5	1	machinable	в пригодной для ЭВМ форме
A	1	G	1	22	classificatory	классифицирующий
A	1	G	2	14	forgettable	легко забывающийся
A	1	G	3	13	implicational	имеющий скрытый смысл
A	1	G	4	5	audile	обладающий хорошей слуховой памятью
A	1	P	1	45	perceptible	воспринимаемый
A	1	P	2	6	italic	набранный курсивом
A	1	P	3	3	accentual	связанный с ударением
A	1	P	4	1	glottal	образованный в голосовой щели
A	2	A	1	57	East Slavic	восточнославянский
A	2	A	2	69	context free	контекстно-свободный

Часть речи английского термина	Длина в словоформах	Часть речи русского соответствия	Длина в словоформах	Количество терминов такого типа	Пример английского термина	Пример русского термина
A	2	A	3	13	consonant like	подобный согласному звуку
A	2	A	4	6	domain specific	характерный для предметной области
A	2	A	5	2	perceptually silent	немой с точки зрения восприятия
A	2	Ap	2	7	loop-free	без циклов

Таблица 9

Частотные соотношения структур английских и русских терминов (фрагмент)

Часть речи английского термина	Длина в словоформах	Часть речи русского соответствия	Длина в словоформах	Количество терминов такого типа	Пример английского термина	Пример русского термина
N	2	N	2	5268	complex preposition	сложный предлог
N	1	N	1	3012	accentology	акцентология
A	1	A	1	1668	achronistic	ахронический
N	2	N	3	1056	substitution drill	отработка по шаблону

Часть речи английского термина	Длина в словоформах	Часть речи русского соответствия	Длина в словоформах	Количество терминов такого типа	Пример английского термина	Пример русского термина
N	3	N	3	578	theme rheme relation	тема-рематическое отношение
N	1	N	2	535	abessive	абессивный падеж
V	1	V	1	499	adjectivize	адъективировать
N	3	N	2	359	suppositional subjunctive mood	гипотетическое наклонение
N	2	N	4	347	ritual language	язык ритуального речевого поведения
N	2	N	1	318	vowel mutation	переогласовка
N	3	N	4	277	subject relative clause	субъектное относительное придаточное предложение
N	3	N	5	119	phrase structure rule	правило развертывания по непосредственно составляющим
N	2	N	5	117	native signer	пользующийся языком жестов с рождения
N	4	N	4	100	standard average European Language	усредненный европейский литературный язык

Часть речи английского термина	Длина в словоформах	Часть речи русского соответствия	Длина в словоформах	Количество терминов такого типа	Пример английского термина	Пример русского термина
N	1	N	3	98	complementixier	морфема комплетивного отношения
A	2	A	2	69	context free	контекстно-свободный
A	2	A	1	57	East Slavic	восточнославянский
N	4	N	3	57	adjective-based adverbial phrase	отадъективная наречная составляющая

3.4. ПРОБЛЕМЫ СООТНЕСЕНИЯ ТЕРМИНОВ В МИКРООБЛАСТИ ТЕСТОЛОГИИ

Тестирование языковой компетенции представляет собой относительно новую подобласть прикладной лингвистики и лингводидактики с относительно автономно развивающейся терминологией. Особое значение корректность перевода и/или передачи терминов этого микрополя получила с внедрением в практику оценивания языковой компетенции и уровня достижений единого государственного экзамена. Некорректность некоторых широко используемых переводов, отсутствие специализированных словарей и справочников приводят к рассогласованию в понимании конкретных методических и лингвистических критериев, аспектов и принципов тестологии.

Сегодня в нашей стране принципам и практикам тестирования в целом начинает уделяться серьезное внимание. Необходимость

разработки подходов к способам контроля и оценивания была осознана с внедрением единого государственного экзамена (ЕГЭ). Результаты ЕГЭ являются де-факто пропуском для учащихся в высшие учебные заведения. Эта значимость ЕГЭ предъявляет его разработчикам повышенные требования с точки зрения основных характеристик теста. Особое внимание к ЕГЭ вполне оправдано и тем фактом, что ЕГЭ стал частью общественной жизни, а следовательно, население должно испытывать доверие к процедуре и результатам, которые представляет этот тест.

В зарубежной традиции единый государственный экзамен рассматривался бы как *high stake test*, т. е. тест, результаты которого имеют высокую степень важности для испытуемых. К созданию апробации, администрированию и безопасности таких тестов предъявляются особые требования. Тест, результаты которого имеют низкую степень значимости для испытуемых, в зарубежной традиции называется термином *low stake test*. Возможно, что учет степени значимости в описании единого государственного экзамена привел бы к тому, что уже на начальном этапе подход к ЕГЭ был бы иным. Рассмотрим (табл. 10) варианты перевода терминов, определяющих степень значимости теста.

Таблица 10

Варианты перевода терминов *high stake test* и *low stake test*

Русский термин	**Термин на английском языке**
high stake test	Особо значимый тест
low stake test	Незначимый тест

Внедрение ЕГЭ не было успешным, вызвав критику со стороны родителей, испытуемых, учителей, преподавателей вузов. Тем не менее ЕГЭ все же выполняет роль единого стандарта образования и является реализацией общего права на получение высшего образования, поэтому его форма проведения в виде теста является на данный момент единственно возможной. Основная проблема состоит не столько в формате ЕГЭ, сколько в том, какие данные этот экзамен предоставил, насколько качественным является собственно наполнение теста. Не будем забывать, что ЕГЭ создавался в условиях отсутствия традиций разработки больших тестовых комплексов, поэтому неудивительно, что опыт ЕГЭ оказался не вполне удачным.

Рассмотрим далее проблемные области, которые были выделены по результатам ЕГЭ:

- результаты тестов не являются надежными;
- процедура тестирования не вызывает доверия у населения и экспертного сообщества;
- низкий уровень развития тестологической науки;
- недостаточное использование зарубежного опыта при разработке тестов.

Кроме этих проблемных областей отдельно стоит вопрос о терминологическом соответствии систем русской и зарубежной традиции. Рассмотрим некоторые аспекты сложности этих соотношений.

Прежде всего, попытаемся проследить соответствие между такими терминами, как *testing*, *assessment* и *evaluation* и русским термином *тестирование*. Отметим некоторые основные положения тестирования.

Тестирование должно предоставить информацию о следующем:

- об уровне владения иностранным языком (сведения о степени сформированности навыков и умений, в соответствии с выбранным критерием адекватности);
- о прогрессе, который сделал тестируемый для достижения поставленной цели;
- о знании языка (насколько оправдался прогноз испытуемых о собственном знании иностранного языка) [95, с. 10].

Исходя из типов информации, которые предоставляют процедуры тестирования, можно выделить функции, которые выполняет лингвистическое тестирование в обучении и жизни — контроль и сертификация уровня владения языком. Под сертификацией понимается процесс прохождения тестирования, построенного на авторитетных языковых стандартах и документации его результатов в случае преодоления необходимого минимума [116, с. 13–15]. Сегодня именно ЕГЭ является примером сертификации, которая оказывает серьезное влияние на будущее тестирование. Каждый этап теста, используемого для тестирования, требует точности при разработке теста, соответствия целям теста, процесса выверки и стремления к сбалансированности самого теста. Таким

образом, к понятию «тестирование» еще относится и процесс создания тестов.

Помимо этого тестирование также является методом контроля, который определяет самую суть тестирования. Это заставляет нас внимательно проанализировать понятие «контроль». Прежде всего, контроль рассматривается как обязательный компонент учебной деятельности. Благодаря контролю совершенствуются приемы управления процессом овладения знаниями, навыками и умениями на иностранном языке. Контроль позволяет получить обеим сторонам процесса обучения информацию о том, какова степень сформированности изучаемых компонентов, а также выявить основные трудности освоения языка [63, с. 81].

Контроль может также рассматриваться и как процесс определения уровня знаний, навыков, умений обучаемого, устанавливаемого в результате выполнения им устных и письменных заданий, тестов и формулирование на этой основе определенной оценки за определенный раздел программы, курса, периода обучения [2, с. 123].

При более общем подходе контроль понимается как процесс, в ходе которого должны быть получены результаты, позволяющие выявить на определенном этапе обучения уровень сформированности тех или иных навыков и осуществить коррекцию имеющихся ошибок у обучающихся. В первом определении принципиальным является включение тестовой практики в понятие контроля, во втором случае важным является упоминание о возможности коррекции ошибок обучаемых, что подразумевает временную протяженность осуществления контроля и его включение в систему обучения.

Временная протяженность контроля позволяет ему реализовывать следующие функции:

- функция обратной связи;
- функция управления;
- функция оценки;
- мотивационная функция.

Выделение функций контроля является очень важным моментом. Любая из функций контроля реализуется в различных формах контроля. И именно разное соотношение этих функций в процессе контроля и определяет отличия форм контроля друг от друга. Соотношение функций контроля может служить базой для определения

различных терминов, связанных с процессом тестирования (например, тест от контрольной работы при их формальном совпадении).

Для определения термина *тестирование* под контролем мы будем понимать вид учебной деятельности, направленный на установление индивидуального качественного уровня владения иностранным языком, на основе количественной оценки речевых действий обучаемого, выявления объектов для коррекционных мероприятий и разработки дальнейшей стратегии иноязычного образования [88, с. 23]. Выбор данной трактовки связан с тем, что в этом определении присутствуют все ключевые компоненты, которые составляют суть процесса контроля:

- включение в процессе учебной деятельности;
- оценка речевых действий обучаемых;
- возможность коррекции траектории обучения на основе полученной количественной и качественной информации;
- указание на все основные функции контроля, мотивации, оценки, управления.

Это определение поможет выявить и отрицательные формы контроля. Например, для объективности контроля важной является проблема субъективности оценки, что, в основном, зависит от выбора объекта контроля. Если объектом контроля является коммуникативная компетенция, то определенная степень субъективности будет заложена в контроль изначально [127, с. 17].

Определив термин *контроль*, мы можем перейти к анализу термина *тест*. Различные определения этого термина, как правило, имеют существенные недостатки. Так, например, в английском языке *test* может обозначать как тестовый комплекс, так и любую другую форму контроля, просто контролирующее мероприятие. По аналогии с английским языком термин *тест* используют для обозначения различных заданий, выполняемых в классах, включая перевод с русского на английский язык, ответы на вопросы и т. д. [89, с. 36]. Включение таких заданий в понятие *тест* приводит к его перегруженности, требующей постоянного уточнения в контексте.

Нельзя забывать о том, что тест, как и любая другая форма контроля, характеризуется целью, структурой, способом анализа результатов. При выборе одной из этих характеристик в качестве доминаты под тестом понимается:

- одна из наиболее доступных форм использования математических методов для оценки процесса усвоения знаний и формирование умений [40, с. 39];
- объективный контроль, охватывающий самые различные языковые аспекты, диагностирующий количественные и качественные отклонения от ожидаемого на конкретном этапе уровня владения языковыми шкалами [87, с. 39];
- метод контроля знания, позволяющий одновременно проверить всех испытуемых, которые выполняют данный тест в течение одинакового временного отрезка с одинаковым по объему и сложности материалом.

Данные определения представляются нам очень узкими, подчеркивающими только какой-то определенный компонент понятия «тест». Подобное сужение понятия выпускает из рассмотрения измерительную специфику теста и его структурные особенности. Данные трактовки представляют разрозненные реализации функций контроля. Для определения термина *тест* необходимо также рассмотреть его отличия от таких терминов, как *тестовое задание* и *задание в тестовой форме, контролирующее задание*. Итак, под тестом мы будем понимать форму контроля, использующуюся для выявления у испытуемых степени лингвистической и/или коммуникативной компетенции и представляющую комплекс заданий, который состоит из частей для тестирования отдельных видов речевой деятельности, прошел апробацию и предлагает получение результатов, которые могут быть оценены по установленным критериям [46, с. 40]. Тест может реализовывать все функции контроля. Это определение предпочтительно тем, что в нем задана цель тестирования, его структура, отмечена измерительная специфика, показана обязательность наличия четких критериев оценивания. Выделение этих необходимых критериев оценивания приводит нас к необходимости вычисления валидности и надежности теста, а также апробации тестовых комплексов. Это определение успешно вписывается в понятие контроля, которое мы рассмотрели ранее. Оно помогает разграничить понятия *тест* и *контрольная работа*, *тестовое задание* и *задание в тестовой форме* или просто *задание*. Контрольная работа будет отличаться от теста тем, что, во-первых, она контролирует сформированность коммуникативной компетенции на гораздо более узком участке, чем

тест, контрольная работа обычно включает в себя меньше материала. В отличие от контрольной работы тест является не просто набором заданий, но использует статистически просчитанное предположение о достоверной информации. Наличие апробации также говорит о том, что тесты имеют тенденцию проверять более широкий спектр компетенций, в то время как контрольная работа направлена на проверку специфической компетенции. Именно поэтому в зарубежной традиции термину *test* противопоставляется термин *quiz* для обозначения «контрольной работы».

На основе данного рабочего определения можно также развести понятия *тестового задания* и *задания в тестовой форме*. Тестовые задания будут являться частью измерения, у них будет свой вычисленный уровень сложности, они будут попадать в тест с серьезной доказательной базой в виде статистических данных. Задания в тестовой форме и тестовые задания также различаются с точки зрения цели. Тестовые задания, являясь частью теста, не только контролируют, но и мотивируют обучаемых. Использование же заданий в тестовой форме необходимо для реализации управляющей формы контроля и функции обратной связи. Для этого удобна тестовая форма. Но использование тестовой формы не сделает из контрольной работы надежный измерительный инструмент. Это важное различие позволяет нам разобраться с употреблением таких терминов, как *test task*, *item* (тестовое задание), *task* (задание или задание в тестовой форме).

Еще одну группу терминов, которая требует серьезных комментариев, составляют варианты и соотношения терминов *assessment*, *evalution* и *testing*. В широком смысле эти слова являются синонимичными, тем не менее следует отметить наличие важных различий. Так, наиболее общим из этих трех терминов является термин *evaluation*, который обозначает целые системы контроля для курсов, учебных программ. *Assessment* и *testing* используются более узко, для именования использования конкретных средств в процессе контроля.

При этом понятию *assessment* более соответствует термин оценивание, обозначначающий варианты контроля, которые не входят в структуру тестирования. Термин *testing* следует переводить как тестирование, в том значении, которое было рассмотрено выше.

Результаты сопоставления русских и английских терминов представлены в табл. 11.

Таблица 11

Сопоставление основных терминов, связанных с понятием тестирования

Русский термин	Термин на английском языке
test	тест
evaluation	контроль
assessment	оценивание
testing	тестирование
quiz	контрольная работа
test task, item	тестовое задание
task	задание или задание в тестовой форме

От четкости определения основных характеристик теста и связанных с ними понятий зависит установление соотношений между другой группой терминов: *validity*, *валидность*, *надежность* и их производные. Рассмотрим эти основные характеристики. Традиционно применительно к тестам используют две основные характеристики — валидность и надежность. Первая из них представляет характеристику теста, которая исследует тест на его эффективность в различных его аспектах. В самом общем виде, валидность определяют исходя из соответствия результата теста с одной или несколькими заданными целями. Например, тест достижений должен представлять все данные для того, чтобы сделать вывод о достижениях испытуемых [91].

Процедурам валидизации должно подвергаться содержательное наполнение теста. В ходе этой процедуры мы должны понимать, насколько полно тест охватывает материал, для проверки которого он был создан. Как правило, в лингвистических тестах содержательная валидность определяется использованием соответствующего языкового материала.

Кроме того, валидность исследуется в области конструкта (*test construct*). Под конструктом понимается способность или признак, который измеряется в ходе теста. Тестовые задания должны стремиться раскрыть требуемые свойства конструкта и следовать его основным психофизическим характеристикам.

При разработке теста важно понимать, что тест может оказывать на испытуемого стрессовое воздействие. В тесте не должно быть ничего такого, чтобы вызывало у испытуемого стресс. Задания должны быть правильно расположены и оформлены, инструкции должны быть четкими и ясными, сам формат тестирования должен быть нейтральным. Здесь оценка валидизации связана с внешними факторами теста.

Таким образом, следует различать разные типы валидности: в области содержания тест должен соответствовать содержательной валидности, в области конструкта — конструктной валидности, в области соответствия результата и цели — сопоставительной или прогностичной валидности, в области оформления — внешней валидности [92]. Описанным видам валидности соответствуют следующие англоязычные термины и их русские эквиваленты (табл. 12).

Таблица 12

Соответствие английских и русских терминов, связанных с валидностью

Термин на английском языке	Русский термин
criterion-related validity	валидность внешнего критерия
discriminant validity	дискриминантная валидность
divergent validity	дивергентная валидность
convergent validity	конвергентная валидность
content validity	содержательная валидность
construct validity	конструктная валидность
prognostic/ predictive validity	прогностическая валидность
concurrent validity	сопоставительная валидность
face validity	внешняя валидность

В табл. 12 представлены те английские термины, которые в нашей традиции употребляются редко. Прежде всего, это термин *criterion-related validity*, которому соответствует перевод *валидность, валидность внешнего критерия*. Фактически, этот термин может быть использован как синоним термину *сопоставительная валидность*,

которая может обозначать сравнение с определенным внешним критерием.

Другие три термина (*divergent validity, discriminant validity, convergent validity*) связаны с понятиями дивергентной и конвергентной валидности, они называют аспекты конструктной валидности, показывая, что данный тест находится в близкой связи с результатами теста, измеряющего схожий конструкт (конвергентная валидность) и насколько далеко он находится от результатов теста, который измеряет несходный конструкт (дивергентная валидность, дискриминантная валидность).

Каждый из рассмотренных терминов представляет валидность как сумму различных характеристик, как составные части общей валидности теста. Такой подход требует определения валидности как единой тестовой характеристики, что позволяет обеспечивать социальную значимость тестирования.

Обобщенный подход к валидности, состоящий в том, что валидность представляет собой единство содержательного, процессуального структурного обобщенного внешнего и последовательного аспектов, предлагает, в частности, С. Мессик [119, с. 75–83].

Еще одной важной характеристикой оценки тестов является *надежность*. Для начала определим то, что мы понимаем под надежностью. Надежность — это характеристика, которая позволяет определить точность и устойчивость измерений теста. Надежность характеризует тест, с одной стороны, с позиции его баланса, т. е. тест должен равномерно распределять задания, чтобы не было очевидных провалов с точки зрения правильности/неправильности. С другой стороны, надежность показывает, насколько тест зависит от условий его проведения и зависит ли. Надежность для теста необходима, но надежность не считается достаточным условием для того, чтобы считать тест валидным. Более того, вполне возможна ситуация, когда валидный тест не обладает высокой степенью надежности.

Обыкновенно степень надежности выражается в коэффициенте надежности [106, с. 370]. Чем ближе этот показатель к единице, тем более надежным считается тест. Однако следует отметить, что коэффициент надежности для тестов, измеряющих разные навыки и умения, будет существенно отличаться. В частности, тест, измеряющий грамматические навыки, будет считаться достаточно надежным, если

он будет иметь показатель коэффициента 0.95. А для теста, проверяющего умения говорения, высокой степенью надежности будет считаться результат 0.85.

В зарубежной терминологии термину *надежность* соответствует термин *reliability*. С термином *reliability* связан ряд терминов, которые требуют уточнения: *split half reliability* и *inter-rater reliability*. Термин *inter-rater reliability* соответствует термину *надежность рейтеров*. Под надежностью рейтеров понимается степень доверия к тому, как эксперты оценивают поведение тестируемых. Особенно актуальна разработка этого термина в рамках поиска критериев и процедур коммуникативного оценивания.

Термин *split half reliability* соответствует методике расщепления теста. Методика предполагает, что тест должен быть разделен на две половины. Если коэффициент корреляции между двумя этими половинами близок к единице, то делается вывод о том, что тест является надежным.

Интересным видится ситуация с вариантами перевода термина Item Response Theory (IRT). Следование концепции этой теории сделало возможным разработку тестовых комплексов, предоставляющих надежную информацию о конструктах теста. Оценивание, осуществляющееся по законам IRT, требует от разработчиков измерительных мероприятий доказательств эффективности применяемых средств и их связи с целями, одним из которых является надежность.

Перевод терминологической аббревиатуры IRT предлагает много разных вариантов. Эти варианты связаны с тем, что популярность и распространение теории сделали его скорее брендом, номеном, чем терминологическим сочетанием. Варианты перевода на русский язык представлены в табл. 13.

Таблица 13

Варианты перевода аббревиатуры IRT на русский язык

Термин	Перевод
Item Response Theory	теория латентных черт
	теория характеристических кривых заданий
	теория моделирования и параметризации педагогических тестов

Термин	Перевод
Item Response Theory	современная теория тестов
	теория ответа на задание
	математическая теория педагогических измерений

Рассмотрим эти эквиваленты. Перевод этого термина как *теория ответа на задание* выглядит как просто пословный перевод тех слов, которые входят в название. Полученный русский вариант не раскрывает специфику этой теории. Именно поэтому эквивалент использовать не рекомендуется.

Термин *современная теория тестов* также считается не очень удачным, поскольку эта теория является не единственной, она имеет свои достоинства и недостатки. Кроме того, другие теории создания тестов пока еще не потеряли возможности развития. Да и само понятие «современности» является достаточно условным.

Использование в качестве варианта перевода *теории латентных черт* имеет серьезные недостатки. Дело в том, что этот вариант уводит нас от основной концепции теории, концентрируясь на том, что в широком смысле теория IRT — это частная реализация методов латентно-структурного анализа. Но, как мы определили выше, концепция теории IRT включает не только математико-статистический аппарат. *Теория моделирования и параметризации тестов* — также неудачный вариант для перевода названия теории и фиксации его терминологического статуса. Так как следует учитывать, что теория IRT рассматривается гораздо шире, чем просто теория создания тестов, сегодня она включает в себя и оценивание. Самым лучшим представляется вариант *математическая теория педагогических измерений*. Этот вариант предполагает оценивание как измерение, в нем отмечена специфика метода измерения, выделена необходимость педагогической составляющей.

Мы рассмотрели лишь некоторые примеры сложностей, возникающих при сопоставлении терминов в области лингвистического тестирования. Без сомнения, требуется дальнейшая работа по уточнению наполнения терминов в сфере тестирования. Гармонизация терминологии в этой области приведет к более простой адаптации и использованию зарубежного опыта тестирования в российских условиях. Важным следствием такой гармонизации станет пересмотр принципов, которые сегодня используются отечественными специалистами для составления тестов.

3.5. ХАРАКТЕРИСТИКА СЛОВАРЯ ПОДЪЯЗЫКА «ПРИКЛАДНАЯ ЛИНГВИСТИКА», ИЗВЛЕЧЕННОГО ИЗ МАССИВА СПЕЦИАЛЬНЫХ ТЕКСТОВ

Язык филологии, как все специальные языки, включает как общенаучную (*проблема, система, анализ, синтез*), общефилологическую (*язык, текст, парадигма, синтагма, синтаксис*), так и отраслевую (*фонема, аллофон, граммема, синтаксема, жанр, стиль, поэтика, разметка, корпус, лингвистический автомат*) терминологическую лексику. То же верно и для подъязыков филологии, при этом очевидно, что специальные узкоотраслевые тексты, тематика которых к тому же ограничена условиями дискурса (материалы конференций, тематические сборники или предметно-ориентированные журналы), так называемые псевдопараллельные тексты, представляют собой релевантную эмпирическую базу для извлечения специализированной терминологии и создания отраслевого словаря. В качестве такой базы был создан массив (корпус) текстов научных докладов VI Международной конференции «Прикладная лингвистика в науке и образовании», проходившей 5–7 апреля 2012 г. в Санкт-Петербурге.

На основе массива, объем которого около 60 000 с/у, был определен состав микрополя «прикладная лингвистика», включающий 8 вершинных узлов, соответствующих 8 направлениям прикладной лингвистики, проблемы которых составили тематику конференции (рис. 5).

Структура микрополя открывает возможность описания частных подъязыков, например подъязыка лингводидактики или подъязыка лингвостатистики. Однако на данном этапе поставим задачу более широко: определим некоторые общие характеристики извлеченного списка терминов и другой специальной лексики (в частности номенов) с точки зрения задач наполнения отраслевого словаря, а именно:

- каноничности (общепринятости) термина;
- представленности в других лингвистических словарях;

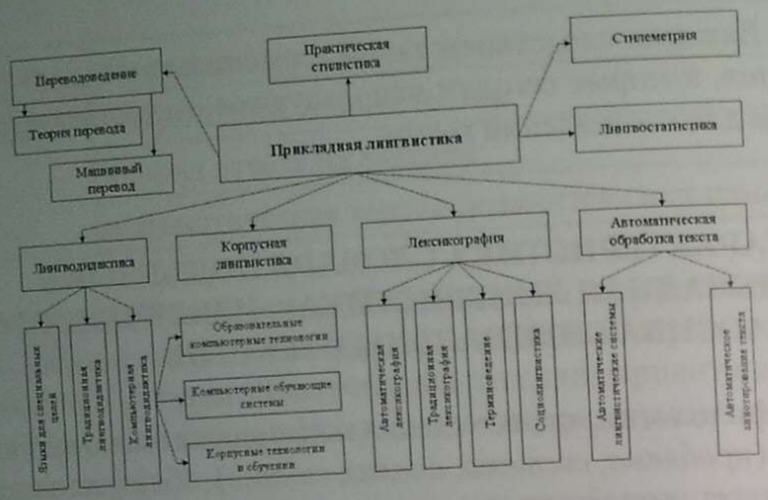


Рис. 5. Карта микрополя «Прикладная лингвистика»

- характера словарной единицы, места и способа ее описания в отраслевом словаре.

Извлечение терминов, кандидатов в термины и другой специализированной лексики было осуществлено вручную на основе полученных с помощью инструмента AntConc 3.2.1.w частотного списка и конкорданса. В отредактированный алфавитный список попали 1736 единиц (универбов — существительных, прилагательных, причастий и, реже, глаголов; словосочетаний и аббревиатур). Для того чтобы определить, насколько полно этот список отражен в русскоязычных и переводных специализированных словарях, была составлена сводная таблица данных извлеченного из массива словаря и данных англо-русского словаря по лингвистике и семиотике [3]. Поскольку сводный словарь получился достаточно объемный, для получения предварительных результатов анализу был подвергнут фрагмент на букву «А». В него вошли 190 терминологических слов и словосочетаний и кандидатов в термины.

Как и ожидалось, в извлеченном словаре оказались термины общенаучные (*автор, адекватный, активный, алгоритм, альтернатива, анализ* и т. п.), общефилологические (*абзац, автосемантический, агглютинативный, адресат, адресант, акроним, акронимия, акцент, аллофон, аллюзия, алфавит, анаграмма* и т. п.) и отраслевые (*лингвистический автомат, автоматический словарь, автоматическое реферирование, автоматизм восприятия, анализатор, аудирование, антипризнак* и др.).

Еще одна группа извлеченных словосочетаний представляет собой сочетание двух общенаучных терминов: *алгоритмическая модель, алгоритмическая процедура, альтернативная реальность, альтернативное значение.*

Именно в сфере общенаучной и общефилологической терминологической лексики совпадений в сопоставляемых словарях, как и ожидалось, больше. Это говорит также и в пользу каноничности (общепринятости) совпавших терминов (см. табл. 14).

Таблица 14

Совпадения в словаре русских терминов из массива текстов «Прикладная лингвистика в науке и образовании» (2012) и Англо-русском словаре по лингвистике и семиотике (фрагмент)

Слова и словосочетания в массиве	Слова и словосочетания в Англо-русском словаре по лингвистике и семиотике	Перевод в Англо-русском словаре по лингвистике и семиотике
абзац	абзац	paragraph
автор	автор	author
автосемантический	автосемантический	autosemantic
агглютинативный	агглютинативный	agglutinative
агент	агент	agent II
адаптация	адаптация	adaptation I
адаптивный	адаптивный	adaptive
адекватность	адекватность	adequacy
адекватный	адекватный	adequate
адресант	адресант	addressant
		locutionary source
		producent
		speaker

Слова и словосочетания в массиве	Слова и словосочетания в Англо-русском словаре по лингвистике и семиотике	Перевод в Англо-русском словаре по лингвистике и семиотике
адресат	адресат	addressee
		audience
		hearer
		narratee
анализ текста	анализ текста	text analysis

Совпадения встречались и в области отраслевой лексики, например:

Таблица 14 (продолжение)

Слова и словосочетания в массиве	Слова и словосочетания в Англо-русском словаре по лингвистике и семиотике	Перевод в Англо-русском словаре по лингвистике и семиотике
автоматизированный перевод	автоматизированная система перевода	machine translation system
автоматический	автоматический	computational
автоматическое аннотирование	автоматическое аннотирование	automated abstracting
автоматическое реферирование	автоматическое реферирование	automated abstracting
анализатор	анализатор	analyzer

Как видно из продолжения табл. 14, отраслевые термины и терминологические словосочетания, совпавшие в обоих словарях, также принадлежат к разряду общепринятых. Однако, как показал анализ фрагмента сводного словаря на «А», большинство лексем и словосочетаний, определенные как термины или кандидаты в термины

в исследуемом массиве специальных текстов, не находят совпадений в Англо-русском словаре по лингвистике и семиотике. Следовательно, необходимость в новом отраслевом словаре очевидна.

Ненашедшие соответствий словосочетания отражают современное положение дел в прикладной лингвистике, свидетельствуют о новых направлениях, методологиях, понятиях, процедурах и т. п. и в рамках специальных текстов приобретают терминологический характер (устойчивость сочетания, воспроизводимость в других текстах («история и география» использования) и его частотные характеристики): *автоматизированный информационный поиск, автоматизированные библиотечно-информационные системы, автоматический морфологический анализ, автоматическое извлечение/вычленение, автоматное программирование, аннотированный текст, аудиторная работа, аудиторное общение, аутентичность иностранной речи, асессор* и т. п.

Для утверждения статуса терминологичности словосочетания-кандидаты в термины должны пройти метрическую оценку (см. раздел 2.5.). Только после таких процедур можно сказать, насколько высока степень «терминоподобия» у таких выделенных из массива сочетаний, как, например, *аргументная лексика, аналитическая деятельность* (как вид обучающей деятельности), *асинхронная многозначность термина, аспектная атрибуция термина, аттестационная справка, аудиторный постер* и др.

Для представления терминов в специализированном отраслевом словаре важно определить, что кроме терминов (универбов и многокомпонентных лексических единиц — коллокаций) в нем могут и должны быть представлены определенные конкретной предметной областью номены и номенклатурные списки (стандарты описания, например). Это предположение возникло после анализа особой части извлеченного из массива словаря, а именно списка иноязычной лексики, вошедшей в русский текст.

В конечный список попали слова, аббревиатуры и акронимы на латинице, которые остались в тексте после удаления формул (буквенные символы), таблиц и списков использованной литературы. Эта часть списка составила словник иноязычной лексики из 111 единиц. В составе словника — английские термины, кандидаты в термины и номены, представленные отдельными лексемами, словосочетани-

ями и сокращенными формами — аббревиатурами и акронимами. Попадание иноязычной лексики в русский текст вызывает особенный интерес с точки зрения ее функции в специальном тексте.

Как показал контекстный анализ лексики в массиве, появление английских лексем в русском тексте не случайно и часто обусловлено отсутствием русского эквивалента. Обращение авторов текста с подобными единицами можно свести к следующим способам:

1) автор использует параллельные термины: русский термин дублируется английским (часто в скобках) и наоборот:

«Возможно использование трех мер сходства: косинусного коэффициента ***(cosine similarity)****, коэффициента Дайса* ***(Dice similarity)****, коэффициента Жаккарда* ***(Jaccard similarity)****»;*

«... имеются специальные инструменты, позволяющие измерять силу не только синтагматических, но и парадигматических связей на основе дистрибуции лексем в корпусе: тезаурус ***(thesaurus)****, кластеризация* ***(clustering)*** *и дифференциация* ***(differences)****...»;*

«... в научном стиле наиболее употребительна конструкция ***short passive*** *(без упоминания/выражения субъекта)...».*

2) в английской лексике в тексте автором дается толкование или собственный перевод:

*«****Knowledge Grid*** *— это интеллектуальная человеко-машинная среда, которая позволяет людям или виртуальным агентам эффективно собирать, координировать, публиковать, совместно использовать и управлять ресурсами знаний»;*

«Тег ***<settingDesc>*** *используется для того, чтобы указать, в какой окружающей обстановке происходит речевой акт»;*

«Оценка значимости термина является важным требованием к термину как входной единице словаря и подчеркивает необходимость эмпирических исследований прагматики словарного пользования и результатов процесса словарного обсуждения ***(actual dictionary consultation)****»;*

3) английская лексема (как правило, номен) не переводится и не транслитерируется, но предваряется именем класса:

«Методика была проверена путем ее имплементации в виде системы ***«RAlign»*** *на языке программирования* ***Delphi****»;*

«Система ***Sketch Engine*** *широко используется при составлении словарей»;*

«В настоящее время на основе международного опыта выработаны де-факто стандарты представления метаданных, как лингвистических, так и экстралингвистических, базирующиеся на описаниях текстов и корпусов в рамках проектов ***Text Encoding Initiative (TEI)****,* ***ISLE Project (International Standards for Language Engineering)*** *и на рекомендациях* ***EAGLES (Expert Advisory Group on Language Engineering Standards)****»;*

4) перевод или толкование на русском языке отсутствуют:

«Идея риторической структуры также нашла отражение в ***CST (cross-document structure theory)****»;*

«Студенту-филологу необходимо уметь работать с данными программами, так как все курсовые и выпускные квалификационные работы пишутся в ***Microsoft Office Word****, а на защите этих работ обязательным требованием является наличие презентации, созданной в* ***Microsoft Office Power Point****»;*

«Представленная в статье методика разработана в рамках гибридного подхода к АОТ и сочетает такие статистические методы, как вычисление ***n-gram*** *и частотности».*

Использование в русских текстах англоязычной лексики достаточно показательно для сегодняшнего состояния терминосистемы «прикладная лингвистика»: преимущество зарубежных (в том числе международных) разработок, инструментов, метрик, систем и стандартов и их популярность в профессиональном сообществе делают их имена узнаваемыми ярлыками, «логотипами», которые иногда получают расшифровку: *CES (Corpus Encoding Standard), CST (cross-document structure theory), TEI, DIS (ISO/DIS)* и т. п. В русскоязычной среде они часто адаптируются фонетически к «латинизированному» прочтению: /теи/, /исо/. Востребованность этих продуктов все же не исключает представленности их названий в специализированном словаре по прикладной лингвистике, где в соответствующей словарной статье могли бы найти место их существенные характеристики и области применения. Основанием для такой информации могут стать комментарии в профессиональных текстах (см. табл. 15).

Количество номенов в полученном иноязычном словнике составило чуть больше половины — 59 единиц (53, 15%), 9 из них используются без перевода, 2 сопровождаются переводом: ISO/TC 37 — Комитет Международной организации по стандартизации, RuTLC

(Russian Translation Learner Corpus) — Русский учебный корпус переводов, остальные (48 единиц) содержат толкование в тексте, как правило, в виде обобщающего слова — имени класса. В эту часть словника вошли:

1) имена продуктов интеллектуальной деятельности:
   a. имена разного рода программных продуктов (инструменты, системы, сервисы, приложения, платформы, автоматические словари);
   b. названия научных и научно-технических разработок (теории, стандарты и форматы представления данных, языки программирования, кодирования и разметки, корпусы текстов, технологии, меры);
   c. названия печатных изданий (учебники и сборники).
2) названия организаций, проектных групп и издательств.

Таблица 15

Англоязычные номены в текстах сборника «Прикладная лингвистика в науке и образовании» (2012). Фрагмент словника иноязычной лексики

Англоязычные номены	Элементы толкования в тексте
American National Corpus	корпус
AntConc 3.2.1.w	конкордансер
BNC	корпус
Booster	учебник
CDIF (Corpus Document Interchange Format)	стандарт представления метаданных
CES (Corpus Encoding Standard)	стандарт представления метаданных
China Knowledge Grid Research Group	[проектная группа] Институт компьютерных технологий академии наук Китая
Corpus Linguistics	сборник

Англоязычные номены	Элементы толкования в тексте
CST (cross-document structure theory)	—
Delphi	язык программирования
DIS (ISO/DIS)	стандарт описания системы элементов данных
DOS	формат
EAGLES (Expert Advisory Group on language engineering Standards)	—
e-mail	—
Exel	программа
Google Docs	сервис
Google.Translater	он-лайн переводчик
Hot Potatoes	инструментальная программа-оболочка
HTML5	технология
ICE (International Corpus of English)	корпус
IELTS	квалификационный тест
Interaction (English for Social and Cultural interaction)	учебник
ISLE Project (International Standards foe Language Engineering)	стандарт представления метаданных
ISO/TC 37	Комитет Международной организации по стандартизации
Knowledge Grid	Среда для поддержки географически распределенной системы параллельных и распределенных платформ для добывания знаний

Англоязычные номены	Элементы толкования в тексте
logDice	статистическая мера
Macmillan	—
Oxford University Press	—
Cambridge University press	—
Collins	—
Cornelsen Verlag	—
Manatee/Bonito	система
Manifold Ranking	алгоритм для ранжирования документов
Microsoft WCF RIA Services	—
MONDILEX	славистический проект
MULTEXT – East Version 4 (multilingual morphosyntactic specifications)	Многоязыковые морфосинтаксические спецификации
Power Point	приложение
PROMT	переводчик
RAlign	система
RuTLC (Russian Translation Learner Corpus)	Русский учебный корпус переводов
Semantic Web	естественноязыковой инструмент Интернета
SGML	язык кодирования
Silverlight	платформа
Sketch Engine	система
Skype	—
StanfordCoreNLP	—
SUMMARIST	система автоматического реферирования

Англоязычные номены	Элементы толкования в тексте
SUMMONS	система автоматического реферирования
SynAF	система синтаксического аннотирования
TEI (Next Encoding Initiative)	стандарт представления метаданных
TMX	формат для обмена файлами памяти переводов
Translate.Ru	он-лайн переводчик
txt	формат
UNICOD	формат
WordNet	тезаурус
XCES (Corpus Encoding Standard for XML)	стандарт представления метаданных
XML	язык разметки
Yandex.Translate	он-лайн переводчик
YouTube	видеохостинг

Термины и кандидаты в термины насчитывают 31 (27,92%) единицу словника иноязычной лексики (табл. 16). Большинство из них сопровождается параллельным переводом, из них 24 представляют фиксированный перевод[1].

В трех случаях можно наблюдать авторский перевод:

«Оценка значимости термина является важным требованием к термину как входной единице словаря и подчеркивает необходимость эмпирических исследований прагматики словарного пользования и результатов процесса словарного обсуждения ***(actual dictionary consultation)»***

[1] Под фиксированным переводом будем понимать такой, который встречается в других текстах данной или смежных областей, в том числе в научных публикациях, словарях, поисковых запросах и т. п.

или авторскую интерпретацию англоязычного термина при наличии общепринятой:

*«Прогресс каждой дисциплины невозможен без уточнения ее концептуально-логического аппарата и совершенствования собственного метаязыка **(LSP)**»*[2]

*«...e.g. **source text** – исходный текст в подлиннике: метафора движения на пути к цели перевода...».*

Имя класса (толкование) употребляется для обозначения типов электронных носителей (CD, DVD и т. п.), без перевода употреблен 1 термин (n-gram), который означает группу из n последовательных символов (слов) и в других текстах массива встретился в иной форме с разными написаниями: *n-грамма, пграма.*

Таблица 16

Англоязычные термины в текстах сборника «Прикладная лингвистика в науке и образовании» (2012). Фрагмент словника иноязычной лексики

Англоязычные термины	Перевод
Фиксированный перевод	
CALL	обучение иностранному языку с помощью компьютера
case-study	кейс-метод
clustering	кластеризация
Conclusion	Заключение
coursebook	учебник
Dice similarity	коэффициент Дайса
differences	дифференциация (тип парадигматической связи на основе дистрибуции лексем в корпусе, ср. кластеризация, тезаурус)
Discussion	Обсуждение
hedges	лексические загородки, хеджи
Inf	инфинитив
Jaccard similarity	коэффициент Жаккарда
instantiation	объективация
key words	ключевые слова
LGP	общелитературный язык
learner corpus	учебный корпус текстов
long passive	форма с упоминанием субъекта действия
Methods and Results	Методы и Результаты
NP	именная группа
References	Список цитируемых источников
rule-based systems	системы МП, основанные на лингвистических правилах; МП, основанный на лингвистических правилах
short passive	форма без упоминания субъекта
teacher's book, teacher's quide	книга для учителя
thesaurus	тезаурус
workbook	рабочая тетрадь
Авторский перевод /интерпретация / толкование (имя класса)	
actual dictionary consultation	словарное обсуждение, процесс словарного обсуждения
AudioCD	формат компакт-диска
CD –ROM	компакт-диск
DVD-ROM	компакт-диск
LSP	метаязык (язык для специальных целей)
source text	исходный текст в подлиннике

[2] ср. общепринятый перевод – «язык для специальных целей», «профессиональный язык»: *«Переносясь в сферу профессиональной коммуникации и пытаясь адаптировать высказанные выше требования к текстам **LSP**, можно констатировать следующее...»*

Англоязычные термины	Перевод
	Без перевода
n-gram	–

Особую группу иноязычной лексики составляют обозначения тегов текстовой разметки для разных стандартов. Номенклатура тегов как правило описывается в объяснительных записках к аннотированным текстовым массивам, например корпусам текстов. Современная тенденция в представлении текстовой информации определяет стандарты методов и языка разметки, поэтому резонно полагать, что номенклатура стандартизованных тегов должна быть представлена в справочных изданиях, возможно, какая-то часть (универсальные теги) должна быть включена в отраслевые словари.

Таблица 17

Список тегов с описаниями.
Фрагмент словника иноязычной лексики

Обозначение тега	Описание тега
<feat>	один из тегов стандарта разметки XCES, признак узла
<m-level>	тег для морфем
<NN>	тег, атрибут существительного
<particDesc>	тег [исп. в корпусах устной речи], который обслуживает дополнительную информацию о говорящих или, если это нужно, о лицах, упомянутых или обсуждаемых в письменном тексте
<p-level>	тег для абзаца
POS	–
<setting>	тег, исп. в корпусах устной речи
<settingDesc>	тег [исп. в корпусах устной речи] для того, чтобы указать, в какой окружающей обстановке происходит речевой акт
<sing>	тег, атрибут существительного в ед. ч.

Обозначение тега	Описание тега
<s-level>	тег для предложения
<teiHeader>	тег «заголовок» в системе TEI
<text>	тег «текст» в системе TEI
<w-level>	тег для слов

Употребление параллельных пар терминов и непереводных терминов в русских специальных текстах свидетельствует о современном состоянии русской терминосистемы подъязыка «прикладная лингвистика», а именно об ее становлении в активно развивающихся областях (компьютерная и корпусная лингвистика, машинный перевод, компьютерная лингводидактика). Это говорит в пользу включения извлеченных из профессиональных текстов параллельных терминов в отраслевой словарь.

Дублирование термина может выполнять информативную (ознакомительную) функцию в том случае, если соответствующий русский термин еще не имеет широкого распространения. С другой стороны, такой прием можно рассматривать как одну из стратегий хеджирования — стратегию солидаризации, установления контакта со «своей» аудиторией, указания на направление, методику, которой автор текста следует в своей работе.

3.6. ПРОГРАММА ГРАФЕМАТИЧЕСКОГО АНАЛИЗА ТЕКСТОВ, ИЗВЛЕКАЕМЫХ ИЗ СИСТЕМЫ ИНТЕРНЕТ

Компьютерно-опосредованная коммуникация как новое поле для лингвистических исследований. На сегодняшний день задача «отслеживания» новых терминов, необходимого для поддержания актуальности лексикографической системы, решается как на материале традиционных текстов различных предметных областей, существующих вне глобальной сети, так и на основе текстов, порожденных в результате коммуникации в сети Интернет.

Быстрый рост популярности социальных сетевых сервисов (иначе называемых социальные медиа, сервисы Web 2.0, социальные

сервисы сети Интернет и т. д. К социальным сервисам сети Интернет относятся, в частности, блоги, в том числе микроблоги, сетевые сообщества, медиахостинги, сервисы социальных закладок и другие площадки общения) — интернет-технологий, которые позволяют пользователям общаться между собой, привел к тому, что интенсивность коммуникации, опосредованной компьютером, стала сравнима с интенсивностью коммуникации вне сети, если не превзошла ее. Сегодня социальные сетевые сервисы играют все более важную роль, являясь безграничным источником информации и универсальной площадкой для обмена мнениями. В Интернете представлены не только тексты, порожденные в процессе бытового общения, но и тексты медицинского, делового, рекламного, массово-информационного, научного и всех прочих существующих разновидностей дискурса.

Эти тексты представляют собой уникальный материал для составления корпусов, которые являются базой данных для работы над созданием современных словарей различных предметных областей и используются для решения исследовательских и практических лексикографических задач. Снабженные соответствующей разметкой, такие корпусы текстов компьютерно-опосредованной коммуникации позволяют получить синхронный срез лексики, дают представление о современном узусе, о действительных особенностях употребления тех или иных лексем, являются уникальным хранилищем иллюстративного материала в помощь лексикографам и исследователям-языковедам.

Однако нужно помнить, что язык общения в социальных сервисах в любой отдельно взятой стране не равен языку общения вне сети, поскольку условия его существования существенно отличаются от обычных. Нельзя не учитывать эти отличия при обращении к социальным сетевым сервисам, какой бы ни была цель — общение, поиск информации или исследования любого рода.

Особенности обработки материала при исследовании компьютерно-опосредованной коммуникации. Изучение языка межличностной компьютерно-опосредованной коммуникации, в частности коммуникации в социальных сетевых сервисах, предполагает работу с большим количеством текстов на естественном языке, порожденных пользователями сервиса сетевого общения и представляющих собой опосредованные диалоги и полилоги. Объем исследуемого

материала вынуждает исследователя искать решения для автоматизации процедуры его обработки и использовать специальные программные средства, позволяющие выделить и систематизировать необходимую лингвистическую информацию. Так, например, автоматическое определение состава и распределения языковых единиц значительно облегчает задачу извлечения лексики из текстов, что является важным не только для исследований в области терминоведения и лексикографии, но и в областях, не связанных, на первый взгляд, с языкознанием (маркетинг, информационный поиск и пр.). Кроме того, программные средства позволяют подкрепить исследование объективными количественными данными.

Для исследования особенностей языка межличностного общения в русскоязычном сегменте сервисов сетевого общения важно анализировать тексты «как есть», т. е. со всеми ошибками, искажениями, уникальными элементами (такими как упоминания, ссылки, эмотиконы и пр.). Поэтому предварительная обработка исследуемых текстов осуществляется в основном на графематическом уровне, что позволяет распознавать элементы, присущие изучаемому языку, во всем их своеобразии. Исследование не предполагает нормализацию текстов, поэтому предварительная обработка не включает в себя лемматизацию и исправление орфографических ошибок.

На сегодняшний день графематические анализаторы являются неотъемлемой частью многих программных продуктов лингвистического анализа и обработки текста (Яндекс.Сервер, AskNet, Russian Context Optimizer). Описание некоторых находится в открытом доступе (например, компонент графематического анализа ГрафАн проекта АОТ). Однако для изучения языка коммуникации в сервисах сетевого общения требуется создать специальную программу, учитывающую цель исследования и особенности общения в сервисах сетевого общения (как пример такого сервиса используется система микроблогов Twitter).

Описание программы: состав и назначение. Разработанная программа представляет собой комплекс функциональных блоков, предназначенный для обработки и анализа текстов на естественном языке, а именно текстов сообщений системы микроблогов Твиттер.

Твиттер — система микроблогов, позволяющая пользователям размещать короткие публичные заметки объемом до 140 символов,

используя веб-интерфейс, средства мгновенного обмена сообщениями или сторонние программы-клиенты.

Программа выполняет первичную подготовку корпуса текстов для последующего исследования путем удаления нерелевантных сообщений (Step 1: Data filtering), проводит графематический анализ текстов (Step 2: Tokenizing) и генерирует отчеты, содержащие количественные характеристики анализируемого корпуса сообщений (Step 4: Report generating). Кроме этого, в программу на экспериментальной основе включен блок морфологического анализа, выполняемого программой mystem (Step 3: Mystem analysis). Полученные данные могут использоваться:

- для исследования особенностей функционирования естественного языка в условиях коммуникации в Твиттере;
- для изучения сочетаемостных предпочтений представленных в корпусе лексических единиц;
- для выявления кандидатов в термины (многокомнонетных в том числе) в предметно-ориентированном корпусе текстов на основе статистического анализа;
- для выявления лексических единиц, используемых для обозначения вновь появившихся понятий предметной области.

Программа подходит для применения в прикладных задачах изучения компьютерно-опосредованной коммуникации, в частности языка общения в Твиттере, позволяет осуществить начальный количественный анализ текстового материала и получить данные, которые послужат доказательной базой лингвистического исследования.

Описание программы: функциональные блоки

Блок 1: Чистка входных данных

Входные данные: текстовый файл в кодировке Windows-1251, содержащий только собственно тексты сообщений, в котором каждая строка является сообщением.

Выходные данные: текстовый файл, содержащий только отфильтрованные строки.

Задача на этом этапе — распознать и удалить все сообщения, не отвечающие заданным критериям. В файле twitkit.ini пользователь может самостоятельно настроить параметры фильтрации. Например,

для обеспечения достоверности результатов обработки текстов при исследовании языка русскоязычной межличностной коммуникации в Твиттере анализируемый материал не должен содержать сообщения рекламного характера, массовые рассылки, а также ретвиты (повторяющиеся сообщения) и сообщения, написанные кириллицей, но не являющиеся русскоязычными.

Программа чистки последовательно анализирует каждую строку. Сначала перебираются символы строки, каждый символ проверяется на принадлежность множеству символов русского и английского языков, а также множеству знаков препинания.

Если присутствуют символы других языков, например украинская і или є, диакритики, умлауты (кроме ё), и т. д., то строка отсеивается.

Далее в строке ищутся последовательности символов, указанные в файле twitkit.ini. Если найдено совпадение, то строка отсеивается. Для этого в файле twitkit.ini в строке кода «ExcludeTokens=» после знака равно указываются элементы, при обнаружении которых программа должна удалить все сообщение. Например, указав «ExcludeTokens=що http://t.co/», пользователь требует удаления всех сообщений, содержащих ùî (так как скорее всего такое сообщение будет на украинском языке) и ссылку http://t.co/ (так как скорее всего это сообщение будет рекламным).

Одновременно с поиском ExcludeTokens в строках ищутся вспомогательные символы машинной кодировки (переносы строк и др., указанные в CleanupTokens) и заменяются пробелами: CleanupTokens=\n /n > < & nbsp; quot.

В результате создается выходной текстовый файл с оставшимися после чистки строками. Алгоритм фильтрации прост, поэтому быстро работает на большом наборе входных строк.

Блок 2: Токенизация

Входные данные: текстовый файл в кодировке Windows-1251, содержащий только собственно тексты сообщений, в котором каждая строка является сообщением.

Выходные данные: бинарный файл с токенизированными строками и базой данных всех токенов.

На этом этапе программа выделяет минимальные лингвистически значимые элементы текста (токены) и приписывает каждому элементу условный тип (дескриптор) (табл. 18).

Таблица 18

Перечень графематических дескрипторов

Название дескриптора	Пояснение	Примеры
RW	Последовательность (словоупотребление) из кириллицы	Кошка
LW	Последовательность из латиницы	Love
RLW	Последовательность из кириллицы и латиницы одновременно	fashion-фотограф
P	Один или несколько подряд идущих знаков препинания	".", '[', ']', '(', ')', '-', ':', ';' и пр.
D	Цифровой комплекс, присваивается последовательностям, состоящим из цифр	1234
DRW	Цифро-буквенный комплекс, присваивается последовательностям, состоящим из цифр и кириллических букв	34й
DLW	Цифро-буквенный комплекс, присваивается последовательностям, состоящим из цифр и латинских букв	34th
DRLW	Цифро-буквенный комплекс, присваивается последовательностям, состоящим из цифр и кириллических и латинских букв	3D-телевизор
HT	Хэштеги (со знаком # в начале)	#СДнемРождения
R	Упоминания/ответы (со знаком @ в начале)	@masha
EA	Ссылки (с www, http и т. п.), электронные адреса	http://angrybirds.com
EM	Эмотиконы, смайлики	:)
OTHER	Последовательности, не обладающие вышеперечисленными признаками	!!!#

Анализ проводится в две стадии: вначале входные строки делятся на «черновые» токены — произвольные последовательности символов, разделенные одним или несколькими пробелами. Так, например, строка «Привет, как дела?)))))» разбивается на три «черновых» токена:

Привет,
как
дела?))))

Далее каждый из токенов анализируется посимвольно, начиная с первого символа. По мере анализа последовательно идущих символов токена выдвигается гипотеза о будущем типе токена, которая может меняться от более простых типов вроде RW, LW, D к более сложным RLW, DRW, DRLW и т. д.

В результате получаем «чистовые» токены всех типов, кроме P и EM, но с общим типом P_EM. Далее токены типа P_EM делятся на P и EM.

Для этого они анализируются отдельно с использованием эвристических методов, например:

- несколько закрывающихся скобок подряд, это эмотикон, гипотеза = EM;
- .!? в начале, это знак препинания, гипотеза = P, остаток токена анализируется отдельно;
- односимвольный токен, символ не принадлежит множеству символов, входящих в эмотиконы, гипотеза = P.

Для одиночных скобок отдельно проверяется их парность, т. е. непарность считается признаком эмотикона.

В результате строка «Привет, как дела?)))))» разбивается токены следующим образом:

Привет – RW
, – P
как – RW
дела – RW
? – P
)))) – EM

Как видно, один «черновой» токен может сразу оказаться чистовым, а может содержать несколько «чистовых».

Далее для токенов типов RW и LW вычисляется регистр написания (все заглавные – AA, все строчные – aa, первая прописная – Aa, смешанный – aA). Результаты записываются в файл *.tkn

На этапе токенизации программа создает список всех токенов из всех сообщений, каждый токен встречается столько раз, сколько он встречается в исходных строках. Этот список может быть упорядочен и отсортирован по любому критерию, например, по типу токена или по типу части речи (после выполнения морфологического анализа), по алфавиту и т. д.

Блок 3: Морфологический анализ с помощью программы mystem

Входные данные: бинарный файл, с токенизированными строками и базой данных всех токенов.

Выходные данные: текстовый файл, содержащий список слов текстов сообщений в алфавитном порядке с грамматическими пометами.

Большая часть данного этапа выполняется программой mystem,

наша программа осуществляет подготовку входного файла, чтение результатов и занесение их в базу данных.

Программа mystem производит морфологический анализ текста на русском языке. Для слов, отсутствующих в словаре, порождаются гипотезы [2].

Программа графематического анализа текстов сообщений системы микроблогов Твиттер подготавливает текстовый файл с упорядоченным по алфавиту списком токенов типа RW, который дается на вход программе mystem. Каждое слово анализируется отдельно, контекст употребления не учитывается. На выходе получаем список слов в алфавитном порядке с грамматическими пометами. Например:

- поэзии{поэзия=S,жен,неод=им,мн|=S,жен,неод=род,ед|=S,жен,неод=дат,ед|=S,жен,неод=вин,мн|=S,жен,неод=пр,ед}
- поэт{поэт=S,муж,од=им,ед}
- поэтическое{поэтический=A=им,ед,полн,сред|=A=вин,ед,полн,сред}
- поэтому{поэтому=ADVPRO=}
- появиоись{появиоись??}
- появится{появляться=V,нп=непрош,ед,изъяв,3- л,сов}
- появиться{появляться=V,нп=инф,сов}.

Тип части речи записывается в базу данных в *.tkn файле. При нескольких гипотезах для одной формы берется та часть речи, которая упоминается большее количество раз. Например, для формы БЛИЗКИЕ{близкие=S,мн,од=им |близкий=A=им,мн,полн|=A=вин,мн,полн,неод} часть речи будет определяться как прилагательное.

Данные, получаемые на этом и предыдущем этапе, используются при построении отчетов.

Блок 4: Построение отчетов

На этом этапе выполняется генерация текстовых файлов отчетов по собранной базе данных. База данных содержит результаты обработки текстов на этапах графематического и морфологического анализа. Если морфологический анализ не выполнялся, отчеты по частям речи не будут доступны.

Список отчетов:

- частотный список типов токенов;
- частотные списки элементов для каждого из типов токенов;

- отчет по языку сообщений (количество сообщений, написанных кириллицей, латиницей и смешанных);
- частотные списки n-грамм различной длины (последовательности из трех, четырех и пяти токенов);
- средняя длина сообщений в знаках, токенах и словах;
- средняя длина токенов каждого типа;
- частотный список частей речи;
- частотные списки слов каждой части речи.

Описание программы: интерфейс пользователя

Программа реализована на языке C++ и совместима с ОС Windows, интерфейс пользователя написан на языке C# (рис. 6).

Программа имеет простой интерфейс, состоящий из окна программы и четырех вкладок. Каждая вкладка соответствует отдельному функциональному блоку.

На вкладке Step 1: Data filtering в поле Input file пользователь указывает путь к текстовому файлу, содержащему сообщения, которые нужно подвергнуть предварительной обработке, т. е. удалить повторяющиеся, рекламные сообщения и пр. В поле Output file указывается папка, в которую следует поместить файл с результатами обработки. Запуск обработки осуществляется кнопкой Start. После

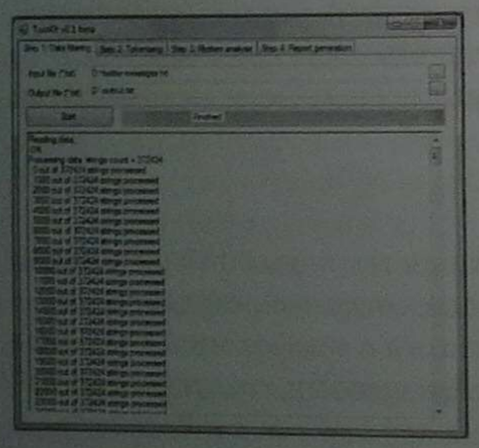


Рис. 6. Пользовательский интерфейс программы. Вкладка Step 1: Data filtering

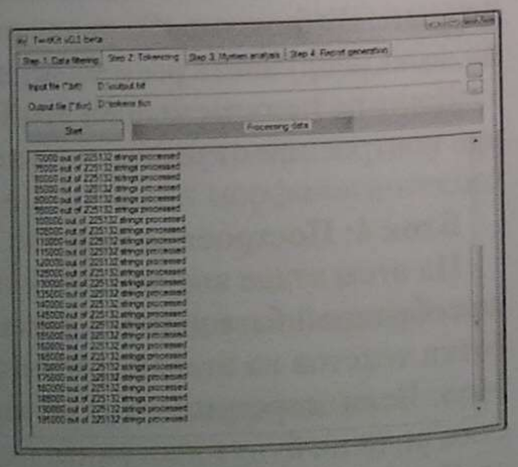


Рис. 7. Пользовательский интерфейс программы. Вкладка Step 2: Tokenizing

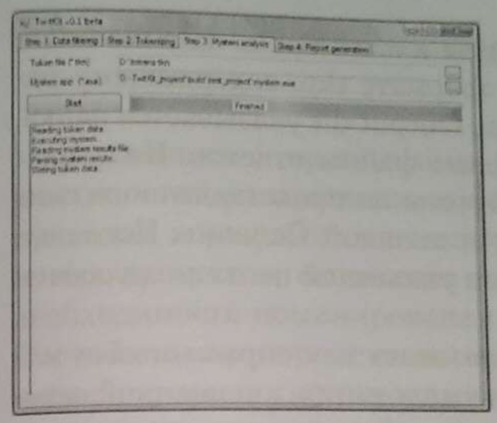


Рис. 8. Пользовательский интерфейс программы. Вкладка Step 3: Mystem analysis

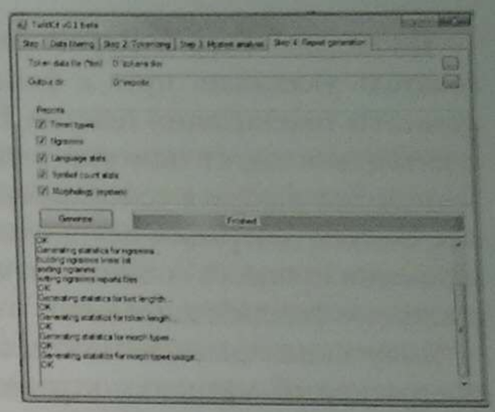


Рис. 9. Пользовательский интерфейс программы. Вкладка Step 4: Report generating

завершения обработки пользователь переходит ко второй вкладке (рис. 7).

На вкладке Step 2: Tokenizing в поле Input file пользователь указывает путь к текстовому файлу, содержащему отфильтрованные сообщения, который нужно подвергнуть токенизации, т. е. разделить тексты на минимальные лингвистически значимые элементы и определить тип таких элементов. В поле Output file указывается папка, в которую следует поместить файл с результатами обработки в формате tkn. Запуск обработки осуществляется кнопкой Start. После завершения обработки пользователь переходит к третьей или четвертой вкладке (рис. 8 и 9).

На вкладке Step 3: Mystem analysis в поле Token file пользователь указывает путь к файлу в формате tkn, содержащему результаты токенизации текстов. В поле Mystem app указывается исполняемый файл mystem, размещенный в папке программы. Запуск обработки осуществляется кнопкой Start. После завершения обработки в папке программы создается текстовый файл mystem_output, содержащий текст с морфологической разметкой.

Заключительным этапом обработки является создание отчетов, содержащих частотные списки всех типов токенов, слов и словосочетаний (последовательностей из 3–5 элементов), частей речи, средние длины сообщений в знаках, токенах и словах, средние длины токенов в знаках (рис. 9).

На вкладке Step 4: Report generating в поле Token data file пользователь указывает путь к файлу в формате tkn, содержащему результаты токенизации текстов. В поле Output dir указывается папка, в которую следует поместить текстовые файлы отчетов. Пользователь может выбрать генерируемые отчеты из представленного списка. Запуск генерации осуществляется кнопкой Generate. После завершения процесса создания отчетов в указанной папке появляются текстовые файлы с данными.

Полученные данные можно использовать как справочный и иллюстративный материал в рамках исследования письменной компьютерно-опосредованной коммуникации в русскоязычном сегменте сети Интернет, в частности, в системе микроблогов Твиттер, для выявления особенностей функционирования языка в условиях компьютерно-опосредованной коммуникации и оценки распространенности тех или иных языковых явлений.

Таким образом, программа графематического анализа текстов сообщений системы микроблогов Твиттер осуществляет начальный анализ текста на естественном языке и предоставляет данные, необходимые для дальнейшего лингвистического анализа.

ЗАКЛЮЧЕНИЕ

Постоянное и чрезвычайно быстрое изменение науки, техники и технологий, развитие новых направлений и новых отраслей знаний приводит к разработке новых устройств, технической реализации новых процессов, осознанию новых явлений. Это влечет за собой необходимость номинирования всего нового и ранее неизвестного. Соответственно, происходит процесс постоянного обновления терминологических систем, возникновения новых номенов и терминов, которые, несмотря на объективную глобализацию науки, получают различную номинацию в языках мира и языках для специальных целей. Эта ситуация приводит к постоянному устареванию специализированных переводных словарей, предназначенных для поддержки работы переводчиков и специалистов, а также для обучения языкам для специальных целей. Сетевые ресурсы, форумы переводчиков и т. п. не являются удобным и достоверным средством работы с новой терминологией.

Опыт работы с переводными словарями, как бумажными, так и автоматизированными, и их анализ подтверждают несоответствие отраслевых переводных словарей современному уровню науки и техники и основным направлениям развития знаний. Это определяется не столько естественным отставанием лексикографии, связанным с необходимостью оценивать изменения в лексической системе языка и описывать только устойчивые, сколько традиционным подходом к созданию словарей как к процессу объединения уже опубликованных материалов, а уже затем к результату фрагментарного анализа текстов, переведенных самими авторами словарей. Сегодня потребность в оперативном извлечении, анализе и переводе новой, постоянно возникающей терминологии и создании словарей хорошо осознается как лексикографами, так и специалистами в информационных технологиях, поскольку именно они должны помочь в разработке специальных методов, так как решение задачи оперативного извлечения терминов возможно только на основе использования корпусных технологий.

В современной лингвистике разработано более 80 различных метрик и методов извлечения терминологии из одноязычных и парал-

лельных или сопоставимых корпусов текстов. Эти методы, несмотря на принципиально эвристический характер, дают хорошие результаты на родственных языках с единой или сопоставимой графикой. В то же время само использование различных метрик и методов извлечения терминологии в случае сопоставительного анализа лексических систем русского и английского языков оказывается гораздо более сложным.

Дело в том, что большинство научных и технических «английских» текстов написано на глобальном английском языке, что в реальности означает нарушение синтаксической структуры предложения в целом и структуры составляющих его групп, вызванное влиянием родных языков. Кроме того, в таких текстах нет гармонизации терминологии, в результате чего термины часто представляют собой переводы соответствующих лексических единиц родного языка автора, а не стандартизированные номинации. «Русские» тексты, в свою очередь, «отягощены» научным канцеляритом, частотным использованием синтаксических структур с объектом в первой позиции предложения и отсутствием явных границ между группами, номинирующими термины и выполняющими разные роли в предложении. Сопоставительный анализ структуры именных групп в русском и английском языках позволяет сделать вывод о различиях в принципах номинирования сложных объектов и степени отражения особенностей этих внеязыковых объектов при расчлененной (многокомпонентной) номинации.

В случае использования параллельных корпусов текстов для выявления кандидатов в термины основным методом является выравнивание по предложениям, которое опирается на формальные показатели границ и частей предложений, соответствие объемно-прагматических структур текстов. При всех возникающих технических и лингвистических сложностях этот процесс вполне реализуем. В случае текстов сопоставимых возможно только терминологическое выравнивание, опирающееся на выявление характерных для обоих массивов корпуса однословных терминологических единиц и их сопоставление в качестве кандидатов в переводные эквиваленты, а также поиск устойчивых словосочетаний с этими однословными терминами в качестве ядер. Дальнейший сопоставительный анализ требует привлечения знаний из переводных автоматизированных словарей, позволяющих верифицировать выбранные пары терминов.

Большинство автоматизированных систем извлечения терминов используют либо статистический, либо лингвистический подход. При этом используется частота лексической единицы в тексте, отношение правдоподобия для двусловных терминов, мера, основанная на полном количестве информации. Для оценки коллокаций, состоящих более чем из двух слов, в качестве единственного статистического параметра используется частота кандидата в термины в корпусе текстов. Использование гибридных подходов представляет собой попытку преодоления ограничений односторонних подходов к решению задачи извлечения терминов на основе как лингвистических, так и статистических элементов.

В основе подобного формально-терминологического анализа лежит сопоставление частоты лексической единицы в проблемно-ориентированном корпусе и в корпусе текстов национального языка. Тогда критерием терминологичности оказывается степень различия этих частот, позволяющая установить особенности «поведения» отдельного слова или коллокации в языке для специальных целей или специально созданном корпусе текстов. Соответственно, различные метрики отличаются именно тем, как устанавливаются подобные критерии и каковы их численные значения.

Автоматизация терминологического анализа корпусов текстов является средством извлечения материала для дальнейшего лексикографического анализа, выполнить который может только специалист. Весь потенциал лингвистических технологий позволяет провести первичный анализ текстов и подготовить для специалиста — лексиколога, лексикографа, терминолога материал для оперативного анализа, тем самым кардинально сократив объем его работы и оставив человеку для решения творческие ментальные задачи. Никакая лингвистическая технология ни сейчас, ни в будущем не отменит необходимость работы специалиста.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. *Азарова И. В., Гордеев С. С.* Построение предметной онтологии на базе тематического корпуса текстов // Труды международной конференции «Корпусная лингвистика – 2011». 27–29 июня 2011 г. – СПб.: СПбГУ, Филологический факультет, 2011. – С. 59–62.

2. *Азимов Э. Г., Щукин А. Н.* Новый словарь методических терминов и понятий (теория и практика обучения языкам). – М.: ИКАР, 2009. – 448 с.

3. Англо-русский словарь по лингвистике и семиотике / А. Н. Баранов, Д. О. Добровольский, М. Н. Михайлов, П. Б. Паршин, О. И. Романова / под ред. А. Н. Баранова, Д. О. Добровольского. – 2-е изд. – М.: Институт русского языка им. В. В. Виноградова РАН, 2003. – 642 с.

4. *Андрющенко В. М.* Вычислительная лексикография и автоматические словари // Вопросы языкознания. – 1986. – № 3. – С. 42–53.

5. *Андрющенко В. М.* Концепция и архитектура машинного фонда русского языка. – М.: Наука, 1989.

6. *Анисимова Ю. А.* Системные характеристики английских юридических терминов и профессионализмов // Единицы языка и их функционирование: межвуз. сб. науч. тр. / отв. ред. С. П. Хижняк. – Саратов: Науч. кн., 2003. – Вып. 9. – С. 18–28.

7. *Антонова А. Ю., Клышинский Э. С.* Об использовании мер сходства при анализе документации // Труды 13-й Всероссийской научной конференции «Электронные библиотеки: методы и технологии, электронные коллекции». – Воронеж: RCDL, 2011. – С. 134–138.

8. *Апресян Ю. Д.* Дистрибутивный анализ значений и структурные семантические поля // Лексикографический сборник / гл. ред. С. Г. Бархударов. – М.: Гос. изд-во иностр. и нац. словарей, 1962. – Вып. 5. – С. 52–72.

9. *Бабушкина Н. В.* Исследование результатов машинного перевода герундия: особенности анализа и критерии редактирования: дис. на соиск. уч. степ. кан-та филолог. наук. – СПб., 2007.

10. *Баранов А. Н.* Введение в прикладную лингвистику. – 2-е изд., испр. – М.: УРСС, 2003. – 358 с.

11. *Беляева Л. Н.* Автоматизированная лексикография: гуманитарные технологии. – СПб.: Изд-во РГПУ им. А. И. Герцена, 2011. – 96 с.

12. *Беляева Л. Н.* Информационные технологии в прикладной лингвистике // Слово и словарь = Vocabulum et vocabularium: сб. науч. тр. по лексикографии. – Гродно: ГрГУ, 2009. – С. 33–37.

13. *Беляева Л. Н.* Лингвистические автоматы в современных информационных технологиях. – СПб.: Изд-во РГПУ им. А. И. Герцена, 2001. – 130 с.

14. *Беляева Л. Н.* Параллельный корпус текстов в задачах лексикографического анализа // Труды международной конференции «Корпусная лингвистика – 2013». – СПб.: СПбГУ, Филологический факультет, 2013. – С. 192–199.

15. *Беляева Л. Н.* Потенциал автоматизированной лексикографии и прикладная лингвистика // Известия РГПУ им. А. И. Герцена. – СПб., 2010. – № 134. – С. 70–79.

16. *Беляева Л. Н., Герд А. С., Убин И. И.* Автоматизация и лексикография // Прикладное языкознание: учебник / отв. ред. А. С. Герд. – СПб.: Изд-во СПбГУ, 1996. – С. 318–333.

17. *Беляева Л. Н., Джепа Т. Л., Зак Г. Н. и др.* Автоматизированное рабочее место филолога в структуре образовательного пространства современного вуза / Л. Н. Беляева, Т. Л. Джепа, Г. Н. Зак, О. Н. Камшилова, В. Р. Нымм, В. В. Разумова. – СПб.: ООО «Книжный Дом», 2013. – 127 с.

18. *Березин Ф. М., Головин Б. Н.* Общее языкознание. – М.: Просвещение, 1979. – 416 с.

19. *Большакова Е. И., Васильева Н. Э.* Формализация лексико-синтаксической информации для распознавания регулярных конструкций естественного языка // Международный журнал «Программные продукты и системы», 2008. – № 4. – С. 103–106.

20. *Бондарко Л. В., Алексеева Т. В., Асиновский А. С. и др.* Морфемный словарь как база автоматизации лингвистических исследований / / Л. В. Бондарко, Т. В. Алексеева, А. С. Асиновский, С. И. Богданов, Н. В. Богданова, О. Б. Ермолаева, Е. Б. Овчаренко, С. В. Степанова, Т. В. Шарыгина // Национальные лексико-фразеологические фонды: сб. ст. / отв. ред. Ф. П. Сороколетов. – СПб.: Наука, С.-Петерб. издат. фирма, 1995. – С. 183–188.

21. *Будагов Р. А.* Введение в науку о языке. – М.: Добросвет–2000, 2003. – 544 с.

22. *Вигурский К. В., Пильщиков И. А.* Филология и современные информационные технологии // Изв. АН. Сер. лит. и яз. – 2003. – Т. 62. – № 2. – С. 9–16.

23. *Виноградов В. С.* Введение в переводоведение (общие и лексические вопросы). – М.: Изд-во ин-та общего среднего образования РАО, 2001. – 224 с.

24. *Власенко С. В.* Отраслевой перевод: синонимизация терминологии как метод компенсации системного диссонанса англо-русских терминосистем // Вестник МГЛУ. — М.: Московский государственный лингвистический университет, 2007. — № 532. — С. 171–183.

25. *Власов Д. Ю., Пальчунов Д. Е., Степанов П. А.* Автоматизация отношений между понятиями из текстов естественного языка // Вестник НГУ. — Новосибирск: Новосибирский государственный университет, 2010. — Т. 8. — Вып. 3. — С. 23–33. (Информационные технологии).

26. *Гвишиани Н. Б.* К вопросу о метаязыке языкознания // Вопросы языкознания. — М., 1983. — № 2. — С. 64–72.

27. *Герд А. С.* Специальный текст как предмет прикладного языкознания // Прикладное языкознание: учебник / отв. ред. А. С. Герд. — СПб.: Изд-во СПбГУ, 1996. — С. 68–91.

28. *Герд А. С.* Типы знаков в языках для специальных целей // Прикладная лингвистика без границ: материалы международной конф., 25–26 марта 2004 г. / ред. Л. Н. Беляева. — СПб.: Инфо-да, 2004. — С. 21–25.

29. *Герд А. С., Богданов В. В., Азарова И. В., Аверина С. А., Зубова Л. В.* Автоматизация в лексикографии и словари-конкордансы // Филол. науки. — 1981. — № 1. — С. 72–78.

30. *Головин Б. Н.* Типы терминосистем и основания их различения // Термин и слово: межвуз. сб. / Горьк. гос. ун-т им. Н. И. Лобачевского; отв. ред. Б. Н. Головин. — Горький, 1981. — С. 3–10.

31. *Головин Б. Н., Кобрин Р. Ю.* Лингвистические основы учения о терминах. — М.: Высш. шк., 1987. — 104 с.

32. *Гринев С. В.* Введение в терминоведение. — М.: Московский Лицей, 1993. — 309 с.

33. *Гринев С. В.* Специальная лексика и терминологическая лексикография // Лексика и лексикография: сб. науч. тр. / отв. ред. Ю. Г. Коротких, А. М. Шахнарович. — М.: МИПС, 1991. — Вып. 1. — С. 16–24.

34. *Даниленко В. П.* Русская терминология: опыт лингвистического описания. — М.: Наука, 1977. — 246 с.

35. *Даниленко В. П., Скворцов Л. И.* Лингвистические проблемы упорядочения научно-технической терминологии // Вопросы языкознания. — М., 1981. — № 1. — С. 7–16.

36. *Дорот В. Л., Новиков Ф. А.* Толковый словарь современной компьютерной лексики. — 2-е изд., перераб. и доп. — Дюссельдорф; М.; Киев; СПб.: БХВ-Петербург, 2001. — 509 с.

37. *Ермаков А. Е.* Неполный синтаксический анализ текста в информационно-поисковых системах // Компьютерная лингвистика и интеллектуальные технологии: труды Международного семинара Диалог–2002: в 2 т. — М.: Наука, 2002. — Т. 2. Прикладные проблемы. — С. 180–185.

38. *Захаров В. П.* Корпусная лингвистика: учебно-методическое пособие. — СПб., 2005. — 48 с.

39. *Изергина И. А.* Синонимия в современной английской терминологии электроники: автореф. дис. ... канд. филол. наук. — Л.: Ленингр. гос. ун-т им А. А. Жданова, 1980. — 62 с.

40. *Ильин Е. П.* Умения и навыки: нерешенные вопросы // Вопр. психологии. — 1986. — № 2. — С. 138–148.

41. *Камшилова О. Н.* Международный язык современного информационного пространства (инновационные образовательные технологии в описании лингвистического объекта): научно-методические материалы. — СПб.: ООО «Книжный Дом», 2008. — 152 с.

42. *Карпинская Е. В.* Унификация, стандартизация, кодификация терминов. Понятие о гармонизации терминов и терминосистем // Культура русской речи: учебник для вузов / под ред. Л. К. Граудиной и Е. Н. Ширяева. — М., 2003. — С. 207–210.

43. *Кит М. С.* О стратегии построения высокоэффективных сетевых словарей (на базе разработки словаря LexSite) // Вестник РГГУ. — 2010. — № 9. — С. 149–160.

44. *Кияк Т. Р.* Лингвистика профессиональных языков и терминоведение // Терминология и знание. Матриалы I международного симпозиума. — М.: Институт русского языка им. В. В. Виноградова РАН, 2009. — С. 21–27.

45. *Козеренко Е. Б., Лунева Н. В., Морозова Ю. И., Ермаков П. В.* Проектирование многоязычного лингвистического ресурса для систем машинного перевода и обработки знаний // Системы и средства информатики. — М.: Наука, 2009. — Вып. 19. — С. 119–141.

46. *Коккота В. А.* Лингводидактическое тестирование. — М.: Высшая школа, 1989. — 127 с.

47. Корпус CELEX. — URL: http://www.ldc.upenn.edu/Catalog/

48. *Кудашев И. С.* Проектирование переводческих словарей специальной лексики. — Хельсинки, 2007. — 445 с.

49. *Кузубов В. Н.* Стандарты информационных технологий — основа интеграции автоматизированных учебных курсов дистанционнного обучения // Открытое образование. — 1997. — № 3. — URL: http://www.e-joe.ru/sod/97/3_97/st086.html(24.10.05)

50. *Кулагина О. С.* Исследования по машинному переводу. — М.: Наука, 1979. — 320 с.

51. *Лейчик В. М.* Предмет, методы и структура терминоведения. — М., 1989. — 89 с.

52. *Лейчик В. М.* Прикладное терминоведение и его направления // Прикладное языкознание: учеб. / под ред. А. С. Герда. — СПб.: Изд-во СПбГУ, 1996. — С. 276–286.

53. Лейчик В. М. Терминоведение: предмет, методы, структура. — 2-е изд. — М.: КомКнига, 2006. — 255 с.

54. *Лейчик В. М., Шеллов С. Д.* Лингвистические проблемы терминологии и научно-технический перевод — М.,1990. — 62 с.

55. *Лукашевич Н. В., Логачев Ю. М.* Комбинирование признаков для автоматического извлечения терминов // Вычислительные методы и программирование. — 2010. — Т. 11. — С. 108–115.

56. *Манерко Л. А.* Аспекты моделирования ментальных процессов при описании терминосистемы // Материалы I Международного симпозиума. — М.: Институт русского языка им. В. В. Виноградова РАН, 2009. — С. 65–77.

57. *Марчук Ю. Н.* Основы терминографии: метод. пособие. — М.: Изд-во Моск. ун-та, 1992. — 76 с.

58. *Марчук Ю. Н.* Проблемы машинного перевода. — М.: Наука, 1983. — 234 с.

59. *Митрофанова О. А., Захаров В. П.* Автоматизированный анализ терминологии в русскоязычном корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: труды международной конференции «Диалог–2009». — М., 2009. — С. 321–328.

60. *Никитина С. Е.* Семантический анализ языка науки: на материале лингвистики. — М.: ЛИБРОКОМ, 2011. — 146 с.

61. Описание системы RU.DICT: Версия 2. 2003. — URL: http://rudict.noolab.ru/about.phtml?2

62. *Павловский Э.* Введение в математическую статистику. — М.: Статистика, 1967. — 288 с.

63. *Пассов Е. И.* Основы коммуникативной методики обучения иноязычному общению. — М.: Русский язык, 1989. — 276 с.

64. *Петрова-Маслакова Т. Н.* Лексикографический анализ терминополя «спортивные танцы» (на материале русского и английского языков): дис. на соиск. уч. степ. канд. филол. наук. — СПб., 2005. — 326 с.

65. *Пиотровский Р. Г.* Лингвистический автомат (в исследовании и непрерывном обучении). — СПб.: Изд-во РГПУ им. А. И. Герцена, 1999.

66. *Пиотровский Р. Г.* Лингвистическая синэргетика: исходные положения, первые результаты, перспективы. — СПб.: Филологический факультет СПбГУ, 2006. — 158 с.

67. *Пиотровский Р. Г., Билан В. Н., Боркун М. Н., Бобков А. К.* Методы автоматического анализа и синтеза текста. — Минск: Вышэйш. шк., 1985. — 222 с.

68. *Пиотровский Р. Г., Рахубо Н. П., Хажинская М. С.* Системное исследование лексики научного текста. — Кишинев: Штиинца, 1981. — 159 с.

69. *Поминов А. В.* Некоторые вопросы построения многоязычных автоматических словарей // Труды Международной конф. «Диалог' 2003» / Ассоц. компьютер. лингвистики и интеллектуал. технологий. — М., 2003. — URL: http://www.dialog-21.ru/archive_article.asp?param=7023&y=2001&vol=6078>

70. *Послед Б. С.* Access 2002: Приложения баз данных: лекции и упражнения. — М. и др.: ДиаСофт, 2002.

71. *Рубашкин В. Ш.* Универсальный понятийный словарь: функциональность и средства ведения // КИИ—2002. Восьмая национальная конференция по искусственному интеллекту с международным участием: труды конференции. — М., 2002. — С. 231–237.

72. *Рябышкин В., Танков С., Киселев С., Ильин Н.* Технологии извлечения знаний из текста // Открытые системы. 31/08/2006. — № 6. — URL: http://www.osp.ru/os/2006/06/2700556/.

73. *Селегей В.* Электронные словари и компьютерная лексикография // Ассоциация переводчиков Lingvo. — URL: <http://www.lingvoda.ru/transforum/articles/pdf/selegey_a1.pdf> 2005.

74. *Семенов А. П.* Проблемы формирования и лексикографического описания терминологии новейших предметных областей: автореф. дис. ... канд. филол. наук / Воен. акад. экономики, финансов и права. — М., 1994. — 16 с.

75. Сетевой словарь LexSite. — URL: http://www.langint.com/lexsite

76. *Табанакова В. Д.* Идеографическое описание научной терминологии в специальных словарях: автореф. дис. ... д-ра филол. наук. — СПб.: СПбГУ, 2001. — 288 с.

77. *Татаринов В. А.* Общее терминоведение: энциклопедический словарь / Российское терминологическое общество РоссТерм. — М.: Московский Лицей, 2006. — 528 с.

78. *Татаринов В. А.* Теория терминоведения: в 3 т. — Т. 1. Теория термина: история и современное состояние. — М., 1996. — 311 с.

79. *Трембач В. М.* Компьютерные методы представления и формирования знаний для синтеза планов решений // Новости искусственного интеллекта. — 2005. — № 3. — С. 51–62.

80. *Хаютин А. Д.* Термин, терминология, номенклатура: учеб. пособие. — Самарканд, 1972. — 129 с.

81. *Хохлова М. В.* Исследование лексико-синтаксической сочетаемости в русском языке с помощью статистических методов (на базе корпусов текстов): автореф. на соиск. уч. ст. канд. филол. наук. — СПб., 2010. — 26 с.

82. *Циткина Ф. А.* Системный анализ в сопоставительном терминоведении // Изв. АН СССР. Сер. лит. и яз. — М., 1988. — Т. 46. — № 6. — С. 557–563.

83. *Шевчук В. Н.* Военно-терминологическая система в статике и динамике: автореф. дис. ... д-ра филол. наук / Воен. краснознам. ин-т. — М., 1985. — 43 с.

84. *Шелов С. Д.* О языковой природе термина // Науч.-техн. информ. — Сер. 2: Информ. процессы и системы. — 1982. — № 9. — С. 1–6.

85. *Шелов С. Д.* Термин. Терминологичность. Терминологические определения. — СПб.: Филологический фак-т СПбГУ, 2003. — 280 с.

86. *Шелов С. Д.* Терминологическая база знаний WinTerm: сводка лингвистических и компьютерных результатов // Терминология и знание: материалы международного симпозиума. — М.: Институт русского языка РАН, 2009. — С. 130–166.

87. *Шипицо Л. В., Сосенко Э. Ю.* Контроль устной речи: на начальном этапе обучения. — М.: Изд-во Московского университета, 1985. — 86 с.

88. *Федотова Л. А.* Рейтинговый контроль обучающихся на факультете довузовской подготовки // Перспективы науки = Science Prospects. — 2010. — № 11. — С. 23–26.

89. *Фоломкина С. К.* Тестирование в обучении иностранному языку // Иностранные языки в школе. — 1986. — № 2. — С. 16–20.

90. *Akhmanova O., Agapova G.* Terminology: Theory and Method / Moscow State Univ. — Moscow, 1974. — 207 p.

91. *Brown R. D.* Adding Linguistic Knowledge to a Lexical Example-Based Translation System // Proceedings of the Eighth International Conference on Theoretical and Methodological Issues in Machine Translation (TMI-99). — Chester, UK, August 1999. — P. 22–32.

92. *Camilli G. and Shepard L. A.* Methods for Identifying Biased Test Items. — Hollywood, CA: Sage Publications, 1994. — 181 p.

93. *Cerbach F., Euzenat J.* Using Terminology Extraction to Improve Tracebikity from Formal Models to Textual Representations // NLDB 2000, LNCS 1959. — Berlin Heidelberg, Springer Verlag 2001. — P. 115–126.

94. *Chacraborty V.* Survey on Comparable Corpora until June 2012 // Data Driven Machine Translation. — Somya Gupta, June 2012. — P. 1–12.

95. *Davies A.* Principles of Language Testing. — Oxford: Blackwell, 1990. — 147 p.

96. *Delpech E., Daille B.* Dealing with lexicon acquired from comparable corpora: validation and exchange // Proceedings, 9th Conference on Terminology and Knowledge Engineering (TKE). — Fiontar, Dublin City University, 2010. — P. 229–223.

97. *Drouin P. and Doll F.* Quantifying TH through Corpus Comparison // Madsen B. N. and Thomsen H. E. (eds). Managing Ontologies and Lexical Resources, TKE 2008. — Copenhagen, 2008. — P. 191–206.

98. *Evert Stefan.* Corpora and collocations. — A. Lüdeling and M. Kytö (eds). Corpus Linguistics. An International Handbook, article 58. Mouton de Gruyter, Berlin. — URL: http://purl.org/stefan.evert/PUB /Evert2007HSK_extended_manuscript.pdf

99. *Frantzi K. T., Ananiadou S., Tsujii J.* The C-value/NC-value Method of Automatic Recognition for Multi-word Terms // C. Nikolaou, C. Stephanidis (eds.): ECDL'98, LNCS 1513. — Berlin Heidelberg: Springer-Verlag, 1998. — P. 585–604.

100. *Fung P.* Finding Terminology Translations from Non parallel Corpora // Proceedings of the 5th Annual Workshop on Very Large Corpora (VLC'97). — Hong Kong, 1997. — P. 192–202.

101. *Gale W. and Church K.* Identifying word correspondences in parallel text // Proceedings of the Fourth Durpa Workshop of Speech and Natural Language. — Asilomar, 1991. — P. 118–132.

102. *Gillam L. and Ahmad K.* Pattern Mining Across Domain-Specific Text Collections // Lecture Notes in Computer Science. — Berlin/Heidelberg: Springer, 2005, 3587. — P. 570–579.

103. *Green D. R.* Consequential Aspects of the Validity of Achievement Tests: A Publisher's Point of View // Educational Measurement: Issues and Practice. Vol. 17, Issue 2. — June 1998. — P. 16–19.

104. *Gruber T. R.* A Translation Approach to Portable Ontologies // Journal on Knowledge Acquisition, 1993, 5 (2). — P. 199–220.

105. *Habash N., Dorr B.* Handling Translation Divergences: Combining Statistical and Symbolic Techniques in Generation-Heavy Macint Translation // D. Richardson (ed.). AMTA, 2002, LNAI 2499. — P. 84–93.

106. *Hagwood C., Kacker R., Yen J., Banks D., Rosenthal L., Gallagher L., Black P.* Reliability of Conformance Tests // COMPSAC98, The Twenty-Second Annual International Computer Software & Applications Conference. — Vienna Austria, August 19-21, 1998. — P. 368–372.

107. *Hutchins J.* The Origins of the Translator's Workstation // Machine Translation, 1998, 13. – P. 287–307.

108. *Justeson J. and Katz S.* Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text // Natural Language Engineering 1 (1), 1994. – P. 9–27.

109. *Kageura K. and Umino B.* Methods for Automatic Term Recognition: A Review // Terminology , 1996, 3 (2). – P. 259–289.

110. *Kelih E.* Preliminary analysis of a Slavic parallel corpus // Proceedings, SLOVKO 2009: NLP, Corpus Linguistics, Corpus Based Grammar Research. – Smolenice, Slovakia, 2009. – P. 175–183.

111. *Kudashev I., Carlson L. Kudasheva I.* TermFactory: Collaborative Editing of Term Ontologies // Proceedings, 9th Conference on Terminology and Knowledge Engineering (TKE). – Fiontar, Dublin City University, 2010. – P. 479–500.

112. *Langlais P., El-Beze M.* Alignement de Corpus Bilingues: Algorithmes et Evaluation. In: Ressources et Evaluations en Ingenierie de la Langue, Collection Actualite Scientifique. – Aupfel-Uref, Paris, France, 1999. – P. 18–43.

113. *Lefever E., Macken L. and Hoste V.* Language-independent bilingual terminology extraction from a multilingual parallel corpus // Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. – Athens, 2009. – P. 496–504.

114. *Lehrer A.* Semantic Fields and Lexical Structure. – Amsterdam: North-Holland; New York: American Elsevier, 1974. – 225 p.

115. *Madsen B. N., Erdman Thomsen H.* A tool for the Construction of Terminological Ontologies // Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA, 2009. – P. 279–282.

116. *McNamara T. F.* Measuring Second Language Performance. – London and New York: Addison Wesley Longman, 1996. – 323 p.

117. *Melamed I.* Bitext Maps and Alignment via Pattern Recognition // Computational Linguistics. – 1999. – Vol. 25. – № 1. – P. 107–130.

118. *Merkel M. and Foo J.* Terminology Extraction and Term Ranking for Standardizing Term Banks // 16th Nordic Conference of Computational Linguistics, Tartu, Estonai, 2007. – P. 349–354.

119. Messick S. Validity // Educational measurement (3rd ed.,). – New York: Macmillan, 1989. – P. 13–103.

120. *Morin E., Daille B., Takeuchi K. and Kageura K.* Bilingual Terminology Mining – Using Brain, not brawn comparable corpora // Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL'07) Prague, Czech Republic, 2007. – P. 664–671.

121. *Rapp R.* Identify Word Translations in Non-Parallel Texts // Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics ACL'95). – Boston, Massachussets, USA, 1995. – P. 320–322.

122. *Somers H.* Knowledge Extraction from Bilingual Corpora // Information Extraction. – Springer-Verlag, Berlin Heidelberg, 1999. – P. 120–133.

123. *Steinberger R.* Language Engineering Technologies and their use for TFUCLAF. A Report on JRC's Institutional Support Activities. – URL: http://langtech.jrc.it/Documents/Report-98_Steinberger_LangTech4OLAF.pdf]

124. Towards Consolidation of European Terminology Resources. Experience Recommendations from Euro TermBank Project. – Edited by: Signe Rirdance, Andrejs Vasiljevs. Riga: Tilde, 2006. – 123 p.

125. TTC Project: Terminology Extraction, Translation Tools and Comparable Corpora. – URL: http://www.ttc-project.eu/about-ttc/concept-and-objectives

126. *Vivaldi J., Marquez L., Rodriguez H.* Improving Term Extraction by System Combination Using Boosting // L. De Raedt and P. Flach (eds.): ECML 2001, LNAI 2167. – Berlin Heidelberg Springer-Verlag, 2001. – P. 515–526.

127. *Weir C. J.* Communicative Language Testing. – New York: Prentice Hall, 1990. – 218 p.

128. *Wong W., Liu W., Bennamoun M.* Determination of Unithood and Termhood for Term Recognition // M. Song and Wu, Y. (eds): Handbook of Research on Text and Web Mining Technologies, IGI Global, 2009. – P. 500–529.

Научное издание

Лариса Николаевна Беляева
Оксана Алексеевна Данилова
Татьяна Леонидовна Джепа
Ольга Николаевна Камшилова
Екатерина Владимировна Карнуп
Волдемар Рихардович Нымм
Сергей Владимирович Чумилкин

ЛЕКСИКОГРАФИЧЕСКИЙ ПОТЕНЦИАЛ СОВРЕМЕННЫХ ЛИНГВИСТИЧЕСКИХ ТЕХНОЛОГИЙ

Монография

Под редакцией Л. Н. Беляевой

Выпускающий редактор *А. С. Балуева*
Корректор *Н. И. Фалёва*
Дизайн, верстка *Т. В. Житкевич*

ООО «Книжный Дом», лицензия № 05377 от 16.07.2001
191186, Санкт-Петербург, М. Конюшенная ул., д. 5

Подписано в печать 18.03.2014.
Гарнитура Petersburg. Формат 60 x 84/16. Бумага офсетная.
Объем 11 печ. л. Тираж 300 экз. Заказ № 453

Отпечатано в типографии ООО «Инжиниринг Сервис»
190020, Санкт-Петербург, ул. Циолковского, д. 13.