

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ
Минский государственный лингвистический университет

АКТУАЛЬНЫЕ ПРОБЛЕМЫ СОВРЕМЕННОЙ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

*Сборник научных статей,
посвященный 80-летию доктора филологических наук,
профессора академика Международной
академии информатизации А. В. Зубова*

Минск МГЛУ
2017

УДК 800:37
ББК 81.1
А437

Рекомендован Редакционным советом Минского государственного лингвистического университета. Протокол от 03.11.2016 № 4(45)

Редакционная коллегия: Л. Н. Беляева, Т. В. Бусел, Р. В. Детскина, А. В. Зубов, Т. П. Карпилович, С. Е. Кунцевич, Д. В. Степанова, И. В. Совпель (*ответственный редактор*)

Актуальные проблемы современной прикладной лингвистики:
А437 сб. науч. статей / редкол.: / Л. Н. Беляева, Т. В. Бусел, Р. В. Детскина [и др.]. – Минск : МГЛУ, 2017. – 328 с.
ISBN-978-985-460-776-4.

Сборник содержит статьи по проблемам современной прикладной лингвистики: теоретическим аспектам, автоматическому выявлению отзывов и мнений в сети Интернет, автоматическому анализу и синтезу текста и речи, проблемам перевода текстов, компьютерной лингвистике, созданию словарей различных типов. В него включены статьи ученых, аспирантов и магистрантов из 8 стран, в том числе 56 статей из 7 учреждений высшего образования и Академии наук Беларуси.

Для научных работников, аспирантов и магистрантов.

УДК 800:37
ББК

ISBN-978-985-460-776-4

© Минский государственный лингвистический университет, 2017

СОДЕРЖАНИЕ

Теоретические аспекты прикладной лингвистики

<i>Богданова Н. А.</i> Формализация логико-семантической структуры художественного текста: проблемные вопросы создания метаязыка	6
<i>Богушевич Д. Г., Иванова-Мицкевич И.В., Москалёва М.И.</i> Формализация процесса построения семантической конфигурации предложения	11
<i>Вербиненко Ю. И.</i> Discourse particles: problems of linguistic description	17
<i>Головня А. И.</i> Системность омонимии в словосложениях	21
<i>Гурбик Е. А.</i> Английский язык в средствах массовой информации Пакистана	27
<i>Дарчук Н. П.</i> Компьютерная грамматика украинского языка АГАТ и перспективы ее использования	32
<i>Детскина Р. В.</i> Структурные модели англоязычных именных словосочетаний	38
<i>Копачева Н. А.</i> Прагмастилистические функции словопорядка в английских сложных предложениях	43
<i>Кунцевич С. Е.</i> Основные направления исследования проблемы жанра в лингвистике	46
<i>Лосик Г. В.</i> Антропологическая информация о вариативности сообщения	53
<i>Маракулина П. А.</i> Понятие речевого конфликта	56
<i>Степанова Д. В.</i> Анализ методов автоматического выделения терминов из научно-технических текстов.....	62
<i>Третьяк А. О.</i> Основное содержание текста и его автоматическое определение	68
<i>Шимуля Роберт.</i> Метафорические названия жанров театрализованных представлений в современном русском политическом дискурсе	73
<i>Юдкин-Рипун И.Н.</i> Реферирование и комментирование в сценаристике как текстовые трансформации	79
<i>Boldak-Brial A. A.</i> L'aspect dialogique des processus cognitifs dans l'écrit de vulgarisation scientifique	84

Автоматическое выявление отзывов и мнений в сети Интернет

<i>Белюга М. А.</i> Многоуровневый подход к измерению степени смыслового подобия текстов	88
<i>Березюк Д. В.</i> Средства формирования тональности письменного текста	93
<i>Дзибук Т. В.</i> Автоматическое извлечение фактов из текстов досье	97
<i>Семак А. С.</i> Статический аспект содержания текстов оценочных отзывов на предприятия гостиничного типа	103
<i>Шумская М. А.</i> Автоматическое формирование ответа на вопрос пользователя как базовая функция вопросно-ответной поисковой системы	108
<i>Bobkov A., Gafurov S., Krasnoproshin V., Vissia H.</i> Semantic-based information extraction in BuzzTalk	113

Автоматический анализ и синтез текстов и речи

<i>Аникеева Т. В., Панова И. И.</i> Константные и вариативные признаки в просодии английской публичной речи представителей различных профессий	118
<i>Болдак И. А.</i> Социальная детерминированность научной биографии	123
<i>Бусел Т. В.</i> Современные подходы к порождению текстов на естественном языке	126
<i>Зубова И. И.</i> Теоретические аспекты понимания письменного текста компьютером	134
<i>Карпилович Т. П.</i> Распознавание смысла аналитического медиатекста формальными методами	139
<i>Корневская Е. Б., Долматова Е. Д.</i> Сочетания идентичных согласных как потенциальная трудность распознавания слова в английской спонтанной речи	145
<i>Крывальцэвіч А. В., Качан Я. С., Гецэвіч Ю. С.</i> Апрацоўка колькасных выказаў у мабільным прыкладанні турыстычнай тэматыкі як сродак вивучэння культурна-маўленчых асаблівасцяў краіны	150
<i>Молоткова А. В.</i> Средства выражения именованных сущностей и фактов в письменном тексте	154
<i>Нагарнович Э. В.</i> Автоматическая коррекция однобуквенных опечаток во франкоязычных запросах пользователей Интернет	159
<i>Панова И. И., Коваленко А. Д.</i> Параметрические значения просодической дистантности орфоэпического стандарта английского языка в Шотландии	165
<i>Сиразитдинов З. А.</i> О моделировании словоизменительной системы именных частей речи башкирского языка парными сочетаниями	170
<i>Скрябина А. В.</i> Определение основного содержания англоязычных публицистических текстов	175
<i>Чапля А. И.</i> Использование статистических методов в стилистических исследованиях	178
<i>Ющенко С. С.</i> Численная оценка коллизийности текстов	183
<i>Яковичин В. С.</i> Formal proto-language in a text processing system	189
<i>Fiodorov A. M.</i> 'SCORAPICE': adaptation of russian shorthand for the english language	195

Создание словарей различных типов

<i>Беляева Л. Н., Новикова К. В.</i> Переводной словарь в синергетическом аспекте: потенциал и применение	201
<i>Василевская В. А., Жданович А. Е., Ильина Е. И., Марковская Е. В.</i> Специфика организации информации в англо-русских терминологических словарях по информационным технологиям	210
<i>Василенко Е. С., Зыгмантович Н. В., Филимонова Т. А., Шалимо Н. В.</i> Специфика организации информации в русско-белорусских терминологических словарях по информационным технологиям	214
<i>Грибач Т. А.</i> Построение лексических словарей для системы автоматического создания транспортных документов	218
<i>Грязнухина Т. А., Заика Н. М., Любченко Т. П., Широков В. А.</i> Интегрированная лексикографическая система как инструмент автоматизированного формирования новых электронных словарей и как база лингвистических исследований в области лексической семантики	221

<i>Зубов А. В.</i> Общие принципы создания терминологических словарей для автоматического перевода специализированных текстов	226
<i>Карпова О. М.</i> Терминологические словари Шекспира нового поколения	228
<i>Меньшенина И. А.</i> О подходах к отбору лексического минимума для обучения специальности	234
<i>Остапова И. В.</i> Этимологический словарь в цифровой среде	237
<i>Рычкова Л. В.</i> Отражение семантической эквивалентности в словаре социально значимой специальной лексики белорусского языка	242
<i>Сянкевіч Н. М.</i> Канкарданс як сродак выяўлення асаблівасцей мовы мастацкай літаратуры	248
<i>Щербин В. К.</i> Сравнительная лексикография и параллельная текстография: анализ взаимосвязей	254

Корпусные исследования

<i>Жубанов А. К.</i> Теоретические основы семантической разметки знаменательных классов слов в Национальном корпусе казахского языка	260
<i>Зубань О. Н.</i> Задачи и методы автоматического морфемного анализа в корпусе украинского языка	265
<i>Камшилова О. Н., Сосунова Т. А.</i> Корпусный анализ грамматического состава учебных диалогов	272
<i>Кошчанка Уладзімір, Смалюк Антон.</i> Праграмае забеспячэнне і працэс падрыхтоўкі тэкстаў для Нацыянальнага корпусу беларускай мовы	277

Проблемы перевода текстов

<i>Горбунова О. Ю.</i> Обучающий французско-русский автомобильный тезаурус в аспекте дидактики технического перевода.....	282
<i>Козеренко Е. Б.</i> Лингвистические представления в интеллектуальных многоязычных системах	286
<i>Макарич М. В.</i> Пути достижения адекватности перевода технических терминов.....	292

Компьютерная лингводидактика

<i>Горбунов Ю. И.</i> Тезаурусная лексикография и компьютерная лингводидактика	297
<i>Ефремова А. А.</i> Дистанционное обучение иностранному языку юристов-заочников	302
<i>Нымм В. Р., Пиотровская К. Р., Макогон И. С., Еремеева Ю. П.</i> Создание и развитие средств электронной поддержки обучения иностранному языку на базе вычислительного эксперимента	308
<i>Пиотровская К. Р.</i> Технологии текст-майнинга как элемент информационной и математической культуры студентов-филологов	312
<i>Потапова Р. К.</i> Перспективы использования гибридных систем распознавания речи (применительно к обучающим системам)	315
<i>Хайчук Роман.</i> Компьютерная лингводидактика (несколько указаний по вопросу о глоттодидактическом приоритете)	319
Авторы статей сборника	324

ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ ПРИКЛАДНОЙ ЛИНГВИСТИКИ

Н. А. Богданова

ФОРМАЛИЗАЦИЯ ЛОГИКО-СЕМАНТИЧЕСКОЙ СТРУКТУРЫ ХУДОЖЕСТВЕННОГО ТЕКСТА: ПРОБЛЕМНЫЕ ВОПРОСЫ СОЗДАНИЯ МЕТАЯЗЫКА

Формализованное представление объектов естественного языка помогает выявлять закономерности лингвистической организации произведений устной и письменной речи. Формализация структурно-содержательных аспектов художественных текстов – одна из самых сложных и спорных задач прикладных исследований. С одной стороны, не всегда очевидна необходимость самой формализации, с другой стороны, художественные тексты демонстрируют самый высокий барьер отторжения к точности метаязыков формализации. Тем не менее, применение в исследованиях художественных текстов методов точных наук помогло по-новому взглянуть на текст как на диалектическое единство общего и индивидуального. Логико-математические методы, обладающие большим эвристическим потенциалом, успешно используют для изучения индивидуального стиля писателей, исследования семантического пространства и композиции сюжета нарративов, построения моделей активного типа, для порождения ритмических структур различных стихотворных форм.

Один из путей исследования композиционных принципов построения художественных текстов, принадлежащих одному жанру, писателю, направлению базируется на построении и последующем анализе логико-семантической формулы текста (ЛСФТ) [1]. Для построения ЛСФТ строятся таблицы основного статического содержания текста (ТОСС) с использованием алфавитно-частотных словарей. ТОСС состоят из наименований главных, второстепенных или прочих объектов, субъектов, действий, признаков, времени и места действия. С помощью ТОСС можно получить наиболее общее представление о содержании текста по опорным словам. Для построения ЛСФТ необходим метаязык логико-семантического описания содержания текстов, который учитывал бы типизированные формы передачи информации с минимальной долей субъективных оценок исследователя.

Разработанный в рамках коммуникативной грамматики русского языка подход Г. А. Золотовой интегрирует в себе весь накопленный опыт изучения текста как речевого произведения с точки зрения его композиционной структуры. Автор предлагает в качестве типизированных форм передачи информации «речевой регистр» (РР), под которым понимается модель речевой действительности, обусловленная точкой зрения говорящего и его коммуникативными намерениями, располагающая определенными языковыми средствами и реализованная в конкретном фрагменте текста [2, с. 314]. Автором выделены следующие РР:

1. Изобразительный (репродуктивный) РР. В данном РР говорящий/пишущий воспроизводит непосредственно сенсорно наблюдаемое, в конкретной длительности или сменяемости действий, состояний, находясь – в реальности или воображении – в хронотопе происходящего.

2. Информативный РР предлагает сообщения о фактах, событиях, свойствах, поднимающихся над наблюдаемым в данный момент, отвлеченные от конкретной длительности единичного процесса. Это сфера не прямого наблюдения, а знания, полученного либо в результате неоднократного наблюдения, опыта, узуса, либо в результате логических мыслительных операций.

3. Генеритивный РР оформляет композиционные отрезки текста, в которых говорящий/пишущий обобщает информацию, соотнося ее с универсальным опытом, облекая высказывания в форму умозаключений, афоризмов, сентенций, пословиц.

4. Волюнтативный РР оформляет высказывания, содержащие побуждение к речевому действию (вопросы), а также к другому действию, которые могут иметь вид побудительных предложений, советов, рекомендаций, наставлений.

5. Реактивный РР характеризуется выражением реакции говорящего на коммуникативную ситуацию (прежде всего в диалоге). В данный РР входят также контактно-этикетные устойчивые фразы, предписываемые общественным узусом.

Понятие речевого регистра связано с понятием «рематическая доминанта» (РД) текстового фрагмента. Категории актуального членения – тема и рема – одни из главных организаторов связности текста, движения мысли от предложения к предложению. За рамками предложения рема вступает в контакт с ремами соседних предложений, создавая при этом РД текстового фрагмента, сигнализирующую о его семантической общности и способствующую членению текста. Выделяют следующие РД [3, с. 315]: *предметная* (описание портрета, пейзажа), *качественная* (характеристика персонажа, предмета), *акциональная* (динамика действия), *статуальная* (описание состояния природы, среды, лица), *статуально-динамическая* (констатация изменения состояния природы, среды, лица), *импрессивная* (субъективно-оценочное восприятие действительности).

Таким образом, при помощи таблиц основного стаического содержания текстов, набора речевых регистров и рематических доминант можно построить схему предельно обобщенного описания содержания художественного текста (табл. 1, 2).

В данных таблицах представлены в самом общем виде возможные информационные блоки текста с учетом речевых регистров, аспекта описания, принадлежности фрагмента текста к диалогической, монологической речи. Как видно из таблиц, состав опорных слов в каждом случае остается неизменным: главные, второстепенные и прочие субъект, объект, признак, время, место. В зависимости от того, в каком логико-семантическом аспекте употребляется опорное слово во фрагменте текста и каков фокус внимания авторов этом фрагменте, информация разделяется по речевым регистрам и рематическим доминантам.

Таблица 1

Схема обобщенного описания содержания художественного текста (РР 1-3)

Рематическая доминанта	Речевой регистр		
	изобразительный (описательный и повествовательный)	информативный (описательный и повествовательный)	генеритивный
Предметная	Описание ГС, ВС, ПС, ГО, ВО, ПО, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Сообщение о ГС, ВС, ПС, ГО, ВО, ПО, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Рассуждение о ГС, ВС, ПС, ГО, ВО, ПО, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ
Качественная	Описание качеств, свойств ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Сообщение о качествах, свойствах ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Рассуждение о качествах, свойствах ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ
Статуальная	Описание состояния ГС, ВС, ПС, ГО, ВО, ПО, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Сообщение о состоянии ГС, ВС, ПС, ГО, ВО, ПО, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Рассуждение о состоянии ГС, ВС, ПС, ГО, ВО, ПО, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ
Импрессивная	Описание ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ, отражающее эмоционально-оценочное восприятие действительности	Сообщение о ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ, отражающее эмоционально-оценочное восприятие действительности	Рассуждение о ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ, отражающее эмоционально-оценочное восприятие действительности
Акциональная	Описание ГД, ВД, ПД ГС, ВС, ПС	Сообщение о ГД, ВД, ПД ГС, ВС, ПС	Рассуждение о ГД, ВД, ПД ГС, ВС, ПС
Статуально-динамическая	Описание изменения состояния ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Сообщение об изменении состояния ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Рассуждение об изменении состояния ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ

Таблица 2

Схема обобщенного описания содержания художественного текста (РР 4-5)

Рематическая доминанта	Речевой регистр	
	волюнтаривный регистр	реактивный регистр
Предметная	Вопрос о ГС, ВС, ПС, ГО, ВО, ПО, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Высказывание о ГС, ВС, ПС, ГО, ВО, ПО, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ
Качественная	Вопрос о качествах, свойствах ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Высказывание о качествах, свойствах ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ
Статуальная	Вопрос о состоянии ГС, ВС, ПС, ГО, ВО, ПО, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ	Высказывание о состоянии ГС, ВС, ПС, ГО, ВО, ПО, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ
Импрессивная	Вопрос ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ, отражающий эмоционально-оценочное восприятие действительности	Высказывание о ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ отражающее эмоционально-оценочное восприятие действительности

Акциональная	Вопрос о ГД, ВД, ПД ГС, ВС, ПС Побуждение ГС, ВС, ПС к ГД, ВД, ПД	Высказывание о ГД, ВД, ПД ГС, ВС, ПС
Статуально-динамическая	Вопрос об изменении состояния ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ Побуждение ГС, ВС, ПС к изменению состояния	Высказывание об изменении состояния ГС, ВС, ПС, ГО, ВО, ПО, ГП, ВП, ПП, ГВ, ВВ, ПВ, ГМ, ВМ, ПМ

При помощи данного метаязыка текстовые фрагменты кодируются как микротемы, из которых и строится ЛСФТ. Микротема – это фрагмент текста, который характеризуется принадлежностью к одному РР, единой РД и единым составом действующих лиц.

Одной из самых сложных задач построения ЛСФТ является вопрос о критериях разграничения микротем. Для формализации логико-семантической организации художественного текста мы предлагаем следующее решение, позволяющее максимально формализовать логико-семантическую сегментацию текста. Микротема считается завершенной, если выполняется хотя бы одно из следующих условий:

- имеется знак препинания, свидетельствующий об окончании предложения;
- оканчивается абзац;
- меняется речевой регистр;
- меняется рематическая доминанта;
- меняется опорное слово.

Такое решение вопроса о разграничении микротем позволяет максимально объективно и тщательно проанализировать логико-семантическую структуру текста. К преимуществам данного подхода можно отнести следующие возможности кодирования текстовых фрагментов:

1. Возможность выделить парцеллированные фрагменты текста в отдельные микротемы. Если одной из авторских черт является подача информации «штрихами», дробными порциями, ЛСФТ отразит данную особенность повтором микротемы. В этом случае парцеллированные фрагменты текста – слова и словосочетания – будут выделены в отдельные микротемы. Если они содержат то же опорное слово, ту же РД и принадлежат тому же регистру, что и предыдущий текстовый фрагмент, микротема будет иметь тот же код, а формула отразит наличие симметричных точек в структуре текста.

2. Возможность выделить микротемы с разными рематическими доминантами. Зачастую при описании одного и того же объекта автором могут делаться разные акценты: на предметах, признаках, действиях. Например, автор в описании главного героя сначала сосредотачивается на предметах одежды, потом на действиях героя. Несмотря на то, что описание может относиться к одному опорному слову, микротемы будут разные, и ЛСФТ отразит смену фокуса внимания автора сменой микротемы.

3. Возможность разделить микротемы при смене регистра. Смена повествования рассуждением, описания – повествованием, рассуждения – описанием и тому подобные смены не настолько сложны в разграничении, как

смены схожих регистров. Нередко в нарративах встречаются фрагменты текста, в которых информация, продвигающая повествование, соседствует с дополнительной информацией, не участвующей в сюжетодвижении. Г. П. Эбот различает в этой связи в структуре нарратива два вида событий: конститутивные (constituent) и добавочные (supplementary), указывая на то, что события первого типа обязательны для нарратива, они развивают сюжет, продвигают его, а события второго типа необязательны, так как никуда не ведут [4, с. 20]. В терминах речевых регистров происходит смена изобразительного регистра на информационный. То есть, например, описание действий главного героя, направленное на развития сюжета, сменяется описанием его действий в прошлом или иных действий, отсылающих читателя за пределы событийной оси текста. При этом и опорное слово, и рематическая доминанта могут не изменяться. Тем не менее, это разные логико-семантические отрезки текста, поэтому ЛСФТ покажет границу таких фрагментов текста сменой микротемы.

4. Возможность разделить микротемы, состоящие из частей одного предложения, которые по авторской задумке разделены абзачными отступами. Иногда авторы прибегают к графическим выразительным средствам, выделяя части фразы или слова в отдельные абзацы. Этот способ схож с парцелляцией, так как отдельно выделенная информация воспринимается маленькими порциями, каждая из которых важна для читателя. ЛСФТ отразит эту особенность, отобразив каждую такую часть отдельной микротемой.

Тем не менее, наряду с достоинствами, данный подход при его практическом применении обнаружил и некоторые недостатки. Не все особенности информационного развёртывания текста удастся отобразить. Например, не учитываются некоторые стилистические фигуры: повторы, перечисления, вынесение за скобки. Если не происходит смены опорного слова, регистра или рематической доминанты, то в рамках одного предложения микротема не будет заканчиваться и дублироваться (как, например, в случае парцелляции).

Из вышеизложенного можно сделать вывод, что созданный метаязык имеет инвариантную базу логико-семантического описания содержания текста в виде таблиц основного статического содержания, речевых регистров и рематических доминант, однако дальнейшие решения должны приниматься исследователем исходя из конкретных целей и материала исследования.

ЛИТЕРАТУРА

1. *Зубов, А. В.* Вероятностно-алгоритмическая модель порождения текста (семанτικο-синтаксический аспект) : дис. ... д-ра филол. наук / А. В. Зубов. – М., 1985. – 399 л.
2. *Золотова, Г. А.* Русский язык: От системы к тексту : учеб. пособие / Г. А. Золотова, Г. П. Дручинина, Н. К. Онипенко. – М. : Дрофа, 2002. – 318 с.
3. *Золотова, Г. А.* Коммуникативные аспекты русского синтаксиса / Г. А. Золотова. – 2-е изд., стереотип. – М. : Эдиториал УРСС, 2001. – 368 с.
4. *Abbott, H. P.* The Cambridge Introduction to narrative / H. P. Abbott. – University Press Cambridge, 2005 – 203 p.

ФОРМАЛИЗАЦИЯ ПРОЦЕССА ПОСТРОЕНИЯ СЕМАНТИЧЕСКОЙ КОНФИГУРАЦИИ ПРЕДЛОЖЕНИЯ

В рамках традиционной семиотики предложение рассматривается как знаковая единица языка, а точнее двусторонняя языковая единица. Это означает, что семантика предложения представляет собой связь между самим знаком, т.е. означающим, и тем, что он замещает, т.е. означаемым [1]. Специфика данной связи заключается в том, что в глубинной сфере значимых единиц содержится как референция к тому, что отражается (денотат), так и указание на то, как денотат отражен (сигнификат) [2]. Особенностью семантики предложения является то, что его сигнификат и денотат представляют собой неизоморфные структуры, которые взаимодействуют друг с другом в процессе построения семантической конфигурации данной синтаксической единицы [3]. Так, два предложения: пушкинское: *Горит восток зарею новой*, и предложение *Начинается утро*, описывают одну и ту же ситуацию, но их пропозициональные структуры отличны. В пушкинском предложении ситуация описывается как воздействие, а во втором предложении, она же описывается как состояние. Из этого следует, что семантика предложения является результатом определенного, вероятно разного для различных языков соотношения денотативной и сигнификативной структур. Тем не менее, любой говорящий на русском языке поймет, что предложение А. С. Пушкина представляет именно начало утра. Из этого следует, что в каждом языке, в каждой культуре имеются определенные закономерности координации денотата и сигнификата, и соответственно, определенные принципы построения семантической конфигурации, иначе вряд ли бы можно было понять не только новые метафоры, но и те, которые уже потеряли метафоричность. Мы все говорим, что часы идут, хотя прекрасно знаем, что вещь под названием часы никуда не перемещается, и даже используем этот факт в загадках: *Что всегда идет, но никуда не движется?*

Исходя из специфики подхода к анализу глубинной сферы предложения, а именно, рассмотрение ее сквозь призму взаимодействия денотата и сигнификата, результирующего в семантической конфигурации, представленные выше идеи, были обобщены нами под названием «композиционный синтаксис». Одной из задач этого направления семантического синтаксиса, является поиск закономерностей отражения денотата предложения, который мы назвали ситуацией, или положением дел посредством его сигнификативной структуры, которую мы назвали пропозицией. Таким образом, ключевыми понятиями в терминологическом аппарате композиционного синтаксиса являются следующие: «денотативная область», «ситуация», «сигнификативная структура», «пропозиция».

Изначально ситуации были названы денотативными областями. Под денотативной областью мы понимаем ту когнитивную структуру, которая образуется в результате восприятия человеком внешнего и внутреннего мира. Именно эта когнитивная структура является тем, что известно в современной

лингвистике как картина мира, и она является денотатом, то есть обозначаемым, для предложения. Та часть денотативной области, которая отражается в отдельно взятом предложении, получила название ситуации. До недавнего времени мы полагали, что денотативные области и ситуации не могут быть классифицированы в зависимости от отношений, которые связывают их конструктивные элементы. Однако поиск закономерностей отражения денотативных областей в целом и их фрагментов (т.е. ситуаций) в частности показал, что денотативные области и ситуации можно типологизировать и свести в некоторые классы, имеющие общее правило отражения в пропозиции.

Пропозиция в контексте композиционного синтаксиса обозначает логическую структуру предложения. Конституирующими именными элементами пропозиции являются аргументы. Строение пропозиций, т.е. число и функции их аргументов определяются характером связывающего их отношения, то есть предикатом. Характеристики предикатов, как показал анализ большого числа предложений, задаются двумя категориями: направленность (наблюдается иерархия аргументов или же ее нет) и динамичностью (наблюдается изменение отношений во времени или же оно остается постоянным). Пересечение значения указанных категорий дает в результате четыре класса пропозиций: векторная пропозиция – отражает динамическую связь, развивающуюся во времени и направленную от одного из аргументов к другим: *John built a house*; пропозиция развития – отражает динамическую, но не направленную связь: *John meets Mary = Mary meets John*; пропозиция отношения – отражает статическую (т.е. не развивающуюся во времени) направленную связь: *John has a car*; скалярная пропозиция – отражает статическую и ненаправленную связь: *John resembles Peter = Peter resembles John*.

Классы денотативных областей значительно разнообразнее классов пропозиций, во-первых потому, что в денотативной области может быть не одна связь, объединяющая ее компоненты, а несколько; во-вторых, денотативные области могут быть как простыми, так и сложными, т.е. включающими несколько структур, каждая из которых может рассматриваться как отдельная денотативная область, включенная в другую денотативную область, как например, денотативная область иллокутивного воздействия, и в-третьих, денотативные области могут пересекаться, что означает, что одна из денотативных областей включает в качестве своих компонентов результаты или некоторые иные элементы других денотативных областей в качестве своих обязательных элементов, как на пример денотативная область информирования [4].

Интегрирование ситуаций, то есть некоторых частей денотативных областей, в пропозицию предложения образует семантическую конфигурацию предложения или то, что традиционно известно как семантическая структура предложения [5]. А поскольку пропозиции и ситуации не изоморфны, в каждом языке имеются свои закономерности помещения имен компонентов денотативных областей в аргументные места.

В данной публикации будет дано краткое описание правил интегрирования простых денотативных областей, которые обнаружены в английском языке и, насколько мы полагаем, могут быть представлены в виде алгоритмов. В качестве примера рассмотрим отражение денотативных областей

созидания и разрушения в предложениях, пропозиция которых наиболее близка к структуре ситуаций. Исследование отражение многих денотативных областей предложений английского языка позволило сделать вывод о том, что в английском языке, синтаксис которого характеризуется тем, что синтаксические функции маркируются позиционно, в поверхностной структуре предложения имеются два места, специально предназначенные для интегрирования ситуации в пропозицию предложения. Первое место, или место подлежащего, которое мы назвали «центром эмпатии», вторым является место непосредственно после сказуемого, или само сказуемое, получившее имя «фокус интереса» [6]. Пропозициональные структуры предложений, сказуемыми которых являются глаголы созидания и разрушения, одинаковы. В них имеется направленный динамический предикат, а именно соответствующий глагол и четыре аргумента: «агенс», «инструменталис», «пациенс» и «результат». В пропозиции может так же появиться необязательный элемент, а именно «бенифициент». Структуры денотативных областей, если из них исключить характеристики связи, так же кажутся одинаковыми и включают некоторые действия, которые можно так же назвать динамическими направленными, и четырех участников: «деятеля», «орудие», «объект» и «результатирующее следствие».

Рассмотрим возможности заполнения центра эмпатии и фокуса интереса именами субстанций выполняющих различные функции в ситуациях созидания и разрушения. Наиболее частым оказывается помещение в позицию центра эмпатии деятеля: *John built the house; John burnt the house*. Тогда, в позиции инструмента оказываются орудия, которые чаще всего опускаются, а заполнение позиции пациенса зависит от денотативной области. Если предикат называет созидание, то в этой позиции размещается “результатирующее следствие” (‘house’). А если предикат называет разрушение, то в этой позиции оказывается “объект” (‘house’), поскольку результирующим следствием в этой ситуации являются угли и пепел). Попытка помещения в центр эмпатии инструмента для данных денотативных областей приводит к появлению сомнительных, с точки зрения правильности, предложений – **Axe and pile built the house; Fire burnt the house*. Важно то, что при отражении других ситуаций появление инструмента в позиции центра эмпатии исключает из пропозиции деятеля. В принципе он может появиться в предложении, но не в аргументной позиции. Наиболее интересной оказывается попытка помещения в центр эмпатии объекта. Для ситуации разрушения появление в агентивной позиции объекта может дать два варианта предложения, с пассивной структурой *The house is burnt* или же с так называемой инактивной конструкцией *The house burnt*. Жесткие правила существуют и для помещения компонентов ситуации в позицию фокуса интереса. Использование в этом месте деятеля автоматически приводит к появлению пассивной конструкции: *The house was built by John; The house was burnt by John*. В этой позиции могут быть и орудия, однако в предложениях отражающих созидание помещение орудия в эту позицию иногда требует построения более сложной структуры и использования глаголов со значением ‘use’, на пример *John employed axe and pile to build the house*. Довольно строгие правила обнаруживаются и при

сопоставлении направленности синтаксических и денотативных связей. Это правило можно сформулировать следующим образом: если направленность синтаксических и денотативных связей совпадают (например, *I have a book*), то предлог, как правило, не появляется; если синтаксические и денотативные связи противоположены (например, *The book belongs to me*), то предлог необходим [7].

Исследования ограничений выявленных при интеграции денотативных структур в пропозицию позволили сформулировать и более частное правило, связывающие компоненты ситуации с предикатом. Это правило можно сформулировать так: если характеристика субстантивного компонента денотативной области предопределяет специфику предикативной связи, то вариативное представление данного компонента в аргументных местах пропозиции невозможно. Сопоставим потенциальную возможность заполнения аргументных позиций для предложений с глаголами 'smell' и 'reek'. В предложениях с глаголом 'smell' в позицию центра эмпатии можно поместить как источник запаха: *The rose smells sweetly* так и экспириентора, то есть воспринимающего: *Jane smells the flower*. В предложениях же с глаголом 'reek' в позиции центра эмпатии может быть только источник запаха: *His dirty clothes reeked*. Причина ограничения заключается в том, что глагол 'reek' указывает не только на наличие запаха, также как и глагол 'smell', но и отмечает его специфическое свойство, зависящее от источника. Поэтому предложения: **John reeked Mike's dirty clothes* оказывается невозможным.

Представленные выше правила, регулирующие построение семантической конфигурации предложения, кажутся слишком обобщенными. Поэтому покажем их действие на примере построения семантической конфигурации предложения, отражающего денотативную область разрушения. Поскольку модели пропозиций и ситуаций включают символы и сокращения, необходимо дать их значения:

- Ag – агенс;
- Pt – пациенс;
- In – инструменталис;
- Fn – финал;
- – направление семантической связи.
- Значение символов ситуации:
- Dr – деятель (doer);
- Tl – орудие (tool);
- Ot – объект (object);
- Rt – результат (result);
-]→ – направление действия.

В модели пропозиции курсивом отмечены аргументы, задающие структуру предложения (центр эмпатии и структуру интереса).

1. Для отражения выбираем ситуацию денотативной области разрушения.

Модель ситуации: Dr]→ Tl]→ Ot]→ Rt.

2. Для именования действия (→) выбираем глагол 'break'. Этот же глагол является предикатом (→) пропозиции:

Ag → In → Pt → Fn.

3. Поставим имя Ot в фокус интереса и расставим остальных участников ситуации по аргументным позициям:

a) $Ag / Dr \rightarrow In / Tl \rightarrow Pt / Ot \rightarrow Fn / Rt$

При заполнении аргументных позиций словами получаем стандартное, для отражения данной ситуации, предложение:

John (Ag / Dr) broke ($\rightarrow$] $\rightarrow$) a glass (Pt / Ot) with a stone (In / Tl) into pieces (Fn / Rt).

Порядок членов предложения обычно не совпадает с порядком семантических элементов [8]. В английском языке, однако, наблюдается обязательное совпадение порядка центра эмпатии (первая позиция) и фокуса интереса (3 и 2 позиции);

б) $Ag / Tl \rightarrow Pt / Ot \rightarrow Fn / Rt$

В результате получаем возможное предложение, которое при заполнении позиций в семантической конфигурации имеет вид: A stone broke the glass into pieces.

4. Поставим имя Tl в фокус интереса. В этом случае имя Ot вытесняется из фокуса интереса и перемещается в центр эмпатии, вытесняя из него имя Dr. Предикат принимает пассивную форму:

$Ag / Ot \rightarrow \leftarrow [be Pt / Tl \rightarrow \leftarrow [Fn / Rt$

The glass (Ag / Ot) was broken with a stone (In / Tl) into pieces (Fn / Rt).

5. Поставим имя Dr в фокус интереса. Из этой позиции вытесняется имя Ot и занимает освободившееся место центра эмпатии. Предикат получает пассивную форму:

$Ag / Ot \rightarrow \leftarrow [be Ved \rightarrow \leftarrow [Pt / Dr \rightarrow \leftarrow [Fn / Rt.$

The glass (Ag / Ot) was broken by John (Pt / Dr) into pieces (Fn / Rt).

6. Поместим Rt в фокус интереса, Ot вытесняется в позицию центра эмпатии:

a) $Ag / Ot \rightarrow] \rightarrow Pt / Rt$

The glass (Ag / Ot) broke into pieces (Pt / Rt);

b) $Ag / Ot] \rightarrow \leftarrow Pt / Rt$

The glass (Ag / Ot) was broken into pieces (Pt / Rt).

7. Поместим предикат в фокус интереса. В этом случае в предложении из компонентов ситуации сохраняется тот, который был в фокусной позиции в стандартном отражении, т.е. Ot. Он и занимает позицию центра эмпатии:

a) $Ag / Ot] \rightarrow \leftarrow$

The glass (Ag / Ot) is broken;

b) $Ag / Ot] \rightarrow \leftarrow (Pt / Rt)$

The glass (Ag / Ot) broke into pieces (Pt / Rt).

При отражении ситуации созидания стандартное предложение имеет иную семантическую конфигурацию:

$Ag / Dr \rightarrow] \rightarrow In / Ot \rightarrow] \rightarrow Pt / Rt$

John (Ag / Dr) built his house (Pt / Rt) out of oak logs (In / Ot).

Появление в позиции фокуса интереса в стандартной конфигурации ограничивает вариативность поверхностной структуры в английском языке, поскольку в позиции центра эмпатии могут быть помещены только деятель (Ag / Dr) или объект (Ag / Ot). В первом случае предикат получает форму активного залога, а во втором – пассивного.

Сопоставление образования семантических конфигураций предложений с глаголами 'break' и 'to build' показало, что строение конфигураций стандартных предложений и, как результат, парадигма их варьирования полностью задается предикатом стандартных предложений. Глагол во всем комплексе своих значений задает тип и структуру пропозиции [9], а его конкретные значения указывают на денотативные области и ситуации в них, поскольку сказуемое является одновременно именем предиката позиции и именем отношения, связывающего компоненты денотативной области. Обе эти характеристики достаточно полно отражены в дефинициях имен предикатов, представленных в толковых словарях. Поэтому формализация синтаксического потенциала предикатов не вызывает особых трудностей.

Иначе обстоит дело с именами участников ситуаций. Толковые словари не дают полной картины синтактико-семантических функций имен компонентов денотативных областей. Более того, как показали исследования синтаксического потенциала терминов родства, эти термины очень редко встречаются в предложениях описывающих ту ситуацию, для именованной которой они предназначены. В то же время они широко используются при отражении ситуаций связанных с теми семами их лексического значения, которые в дефинициях не упоминаются, но которые образуют прессупозиционный фонд этих слов. Эти термины встречаются в основном в предложениях, описывающих речь, а так же передачу информации. Это означает, что термины родства в большинстве своих употреблений реализуют синтаксический потенциал семы 'человек'.

Поскольку в толковых словарях большинство сем семантической структуры существительных не указаны и более того, в исследованиях лексикологического характера, где можно их обнаружить, практически никогда не отмечается семантико-синтаксический потенциал семного состава слов, необходимо найти иные пути определения семантико-синтаксических характеристик существительных, поскольку без этого формализовать способы порождения семантической конфигурации предложения, а от нее перейти к формализации способа образования поверхностной структуры, не представляется возможным. Можно было бы использовать ассоциативные словари. Однако в них, как правило, представлены так называемые парадигматические отношения, а синтаксические связи включают стандартные, т.е. появляются в ассоциациях только элементы денотативных областей, для именованной участников которых и созданы данные существительные [10]. Поэтому для выявления семантико-синтаксического потенциала существительных, т.е. определения ролевых характеристик их референтов необходимо выявить их дистрибуцию, т.е. постараться собрать как можно больше предложений, в которых эти имена используются.

Сказанное выше позволяет нам заключить, что алгоритмизация процесса построения семантической конфигурации предложения возможна с позиций композиционного (интегрального) синтаксиса, а точнее с позиций обязательного разделения денотативной и сигнификативной сторон семантики предложения, структурирования данных аспектов и дальнейшего их интегрирования в единую семантическую конфигурацию на основе действия механизма

выбора фокуса интереса говорящего и механизма размещения центра эмпатии. Вместе с тем, стоит отметить, что полная формализация построения семантической конфигурации предложения возможна только при условии определения семантических и синтаксических потенций, по крайней мере, существительных, реализующих на поверхностном уровне именные компоненты семантической конфигурации. В настоящее время определение семантико-синтаксического потенциала имен существительных является основным проблемным полем исследований, проводимых в рамках композиционного синтаксиса.

ЛИТЕРАТУРА

1. Моррис, Ч. У. Знаки и действия. Значение и означивание / Ч. У. Моррис // Семиотика; сост., общ. ред. Ю. С. Степанова. – М., 1983. – С. 118–133.
2. Фреге, Г. Смысл и денотат / Г. Фреге // Семиотика и информатика : сб. науч. ст.; гл. ред. В. А. Успенский. – М., 1997. – Вып. 35. – С. 351–379.
3. Bogushevich, D. G. Outline of the Compositional Syntax / D. G. Bogushevich [et al.] // Proceedings of the 16th International Congress of Linguists, 20–25 July 1997 : Philosophy of Linguistics. – Oxford : Pergamon, 1997. – Paper № 0013.
4. Иванова-Мицевич, И. В. Денотативная область вербального информирования и особенности ее отражения в английском предложении : дис. ... канд. филол. наук : 10.02.04 / И. В. Иванова-Мицевич. – Воронеж, 2009. – 200 л.
5. Иванова-Мицевич, И. В. Интегральная модель семантики английского предложения / И. В. Иванова-Мицевич // Грамматика разноструктурных языков: сб. науч. ст. к юбилею проф. В. Ю. Копрова. – Воронеж : Наука-Юнипресс, 2011. – С. 151–159.
6. Иванова-Мицевич, И. В. Факторы структурирования семантики предложения / И. В. Иванова-Мицевич // Науч. Вестн. Воронеж. гос. архитектурно-строительного ун-та. Серия «Современные лингвистические и методико-дидактические исследования». – 2009. – Вып. 1 (11). – С. 124–132.
7. Коренько, Е. А. Формальные различия предложений с глаголами-конверсиями в современном английском языке : дис. ... канд. филол. наук : 10.02.04 / Е. А. Коренько. – Минск, 2006. – 123 л.
8. Лэм, С. М. Очерк стратификационной грамматики / С. М. Лэм. – Минск : Высш. шк., 1977. – 136 с.
9. Федорова, Н. Н. Синтаксико-семантический потенциал глаголов перемещения в английском языке : дис. ... канд. филол. наук : 10.02.04 / Н. Н. Федорова. – Минск, 1992. – 220 л.
10. Курилович, О. Ф. Взаимодействие ассоциативных структур при порождении высказывания (на материале английского языка) : дис. ... канд. филологических наук : 10.02.04 / О. Ф. Курилович. – Минск, 2008. – 144 л.

Yu. Verbynenko

DISCOURSE PARTICLES: PROBLEMS OF LINGUISTIC DESCRIPTION

Among not fully developed lexical units that control the process of communication (such as modal words, conjunctions, etc.) can be distinguished class of discourse particles. Such language constructs provides text coherence, transfer speaker's attitude to statements, focus one's attention to the context; discourse

particles make statements more clear, structured; regulate emotional coloring, make text clearer [1; 2]. Functions of markers in the context are very wide, amongst them: statement organizing, shifting from one topic to another, text macrostructure expressing, marking of individual point of view, expressing attitude to the statements, its characterizing etc. [3] Such focusing of discourse particles allows to qualify them as units, which provide certain functions of linguistic pragmatics.

At the same time, the number of such units, frequency of their usage and their formal grammatical structure are associated not only with the structure of a particular language but also with speaker's individual linguistic view of the world. All these cause difficulties during their translation. Since language and culture are inseparable, using of foreign languages in isolation from the culture is impossible; the difference between cultures usually has no clear recordings in dictionaries, so researchers point out that cross-language cultural barrier creates additional problems to the lingual communication [1]. All said above becomes especially important in the context of modern scientific communication.

Also it is difficult to describe discourse particles due to that fact that the theory of their semantic description and its lexicographical representation are not developed. There are some difficulties in the forming of dictionary definitions of the units of this class, as they are often integrated into the context. This makes it difficult to define system boundaries of units of this class and to build their correct classification.

Some authors (K. Holker, 1991, L. Brinton, 1996, A. Jucker & Y. Ziv, 1998) had attempted to highlight linguistic units with discourse markers, that are, in particular, referred to phonological, semantic, syntactic, functional, sociolinguistic and stylistic similarities and differences. Later on linguists tried to classify the discourse markers due to their belonging to certain lexical and grammatical classes (prepositions, conjunctions – coordinating and subordinating, interjections, parenthesis, modal particles some adverbs). French linguist Denis Paillard and his colleagues noted the uncertain status of discourse words, and indicated function words as the closest part of speech to which they could be related [4].

However, nowadays there is no even minimally closed list of discursive markers, as well as a complex of their system signs. In the works of foreign scientists various discourse markers and their groups are studied, mainly from the structural and semantic point of view. Thus, different linguists have different opinions concerning what lexical units have to be marked as discourse markers (Rouchota 1998; Schiffrin 1987; Dijk 1977; Blakemore 1992; Vishnevskaya, Lihareva 2000, Tyurina 2003) [5].

The problems of polysemy and polyfunctionality of discourse particles are also academically interesting. But attention to these problem was not given in dictionaries and grammars; and they did not get rightful place in discourse particles descriptions [6].

Thus, it is obvious that there is lack of theoretical and, especially, lexicographical works on linguopragmatics of discursive particles, and there are almost no works that presented Ukrainian language. In Ukrainian lexicography A. Luchyk studies similar problems ("Words equivalents in Ukrainian", "Russian-Ukrainian and Ukrainian-Russian dictionary of word equivalents").

E. Khachaturyan in [7] notes that many discourse particles also have non-discourse meaning. But it is difficult to find a correlation between them. Linguists' thoughts about these particles are differs greatly. Such scientists as B. Fraser think that there are homonyms relationships between two identical units, that "pragmatic meanings of discourse markers are not relate to the meanings of homonymous forms" (Fraser, p. 389). Others divide lexical units, which meanings are determined by meanings of their "semantic relatives", and lexical units, which meanings are not dependent from their homonymous forms meanings.

However, recent studies are more and more clear formulate an idea that there is a correlation between semantics of discourse particles and semantics of lexical words, which allows to talk about forming of discursive particles through partial desemantization.

For example:

1. Скажем (допустим) / Скажімо (припустимо)

Discourse meaning:

If the state of a system is changed, **as by** heating, the values of its state functions change too.

Если состояние системы изменится, **скажем**, при нагревании, значения его функций состояния тоже изменятся.

Якщо стан системи зміниться, **скажімо**, при нагріванні, значення його функцій стану теж зміняться.

Сказать (что-л.) / Сказати (що-н.)

Non-discourse meaning:

It seems impossible **tosay anything against** such a point of view.

Кажется, невозможно **сказать** ничего против такой точки зрения.

Здається, неможливо **сказати** нічого проти такої точки зору.

2. Допустим (предположим) / Припустимо

Discourse meaning:

Let us assume that the situation is represented by a triangle.

Допустим, что эту ситуацию можно представить в виде треугольника.

Припустимо, що цю ситуацію можна представити у вигляді трикутника.

Допускать (позволять)

Non-discourse meaning:

The X-ray tube current **should not be allowed to** exceed 80 μ A.

В рентгеновской трубке **нельзя допускать** превышение тока выше 80 μ A.

У рентгенівській трубці **не можна допускати** перевищення струму вище 80 μ A.

3. Другими словами / Иначе кажучи

Discourse meaning:

– **In other words**, we observe evidence of a bimodal relationship.

Другими словами, мы наблюдаем свидетельство бимодальных отношений.

Иначе кажучи, ми спостерігаємо підтвердження бімодальних відношень.

Non-discourse meaning:

He had said this **in other words**, not the way you has said this just now.

Он говорил это **другими словами**, не так, как ты сейчас пересказала.

Він говорив це **іншими словами**, не так, як ти зараз переказала.

4. В любом случае / У будь-якому разі

Discourse meaning:

– Such separation may, **in any case**, not be possible.

В любом случае, такое разделение не представляется возможным.

У будь-якому разі, такий поділ не представляється можливим.

Non-discourse meaning:

You know how to behave in **any possible case**.

Ты знаєш как себя вести **в любом случае** из возможных.

Ти знаєш як себе поводити **в будь-якому випадку** з можливих.

5. На самом деле (действительно) / Насправді (дійсно)

Discourse meaning:

Actually, neither silicon nor germanium crystals have been satisfactory for this application.

На самом деле, ни кристаллы кремния, ни германия не были удовлетворительными для этих целей.

Насправді, ні кристали кремнію, ні германію були задовільними для цих цілей.

Non-discourse meaningЖ

– A blackbody does not **really** exist in nature.

Черного тела **на самом деле** в природе не существует.

Чорного тіла **насправді** в природі не існує.

The last example represents special semantic condition – superposition, where discourse or non-discourse meaning depends from the position in the sentence [8].

All units that have lost their original meaning and transferred to discourse particles, get a number of formal features (mainly syntax and intonation ones) that help to distinguish discourse meaning and usage from non-discourse. E. Khachatryan identified the main features:

– discourse particle could not be an answer to a question when using in isolation;

– it do not use with negation (unless negation is a part of a discourse word);

– it usually is omitted with indirect speech;

– it can not be repeated in echo-question;

– unlike members of sentences, position of a discourse word (that has no syntactic function in the sentence) is not fixed, but is determined by a semantic criteria;

– usually, discourse particle or the entire construction with it in a statement is distinguished by lexical means (such pauses).

In this work we attempted to classify and describe Ukrainian discourse units and their English and Russian equivalents. The study was conducted on the dictionaries: "Russian-Ukrainian-English for Physicists" Yudina S., "Russian-English Translators Dictionary: A Guide to Scientific and Technical Usage" Zimmerman M., "Russian-English Chemical and Polytechnic Dictionary" Ignatiev Kallahen L.

The basis of our classification are the following five groups of discourse words, distinguished by Viktorova E.:

– organizational-structural (поскольку – оскільки – aslongas / because / since; однако – but / howewer);

– subjective-modal (фактически – фактично – actually / in fact / as a matter of fact; казаться – здаватися – appear as / seem to);

- emphasizing (следует отметить, что – слід відзначити, що – it should be observed that / it should be pointed out that);
- reflexives (следует отметить, что – слід відзначити, що – it should be observed that / it should be pointed out that);
- discourse words of direct addressing (а именно – а саме – namely / that is; с точки зрения – з точки зору – from a viewpoint of).

REFERENCES

1. *Викторова, Е. Ю.* Вспомогательные и коммуникативные единицы в политическом дискурсе : проблемы перевода (на материале русского и английского языков) / Е. Ю. Викторова // Изв. Саратов. ун-та. – Саратов, 2010. – Т. 10, № 4. – С. 40–47.
2. *Еремина, И. С.* Установление английских эквивалентов русского *как бы* : (экспериментально-сопоставительное исследование на материале английского и русского языков) : дис. ... канд. филол. наук : 10.02.20 / И. С. Еремина. – Пятигорск, 2005. – 133 с.
3. *Правикова, Л. В.* Дискурсивные маркеры: современное состояние проблемы / Л. В. Правикова // Вестник ПГЛУ, 2000. – № 4.
4. *Киселева & Пайар.* Дискурсивные слова русского языка / Киселева; под ред. Пайар // Опыт контекстно-семантического описания. – М. : Метатекст, 1998.
5. *Каменский, М. В.* Социолингвистическая парадигма дискурсивных маркеров (на материале английского языка) : дис. ... канд. филол. наук : 10.02.19 / М. В. Каменский. – Ставрополь, 2007. – 179 с.
6. *Нанпи, Т.* Правда прямо идет, а с нею не разминешься (о статусе дискурсивных слов и частиц) / Т. Нанпи.
7. *Хачатурян, Е. В.* Семантика и синтактика дискурсивных слов глагольного происхождения в современном итальянском языке : дис. ... канд. филол. наук : 10.02.05 / Е. В. Хачатурян. – М., 2000. – 171 л.
8. *Широков, В. А.* Комп'ютерна лексикографія : моногр. / В. А. Широков // НАН України. Укр. мов.-інформ. фонд. – Київ: Наук. думка, 2011. – 351 с.

А. И. Головня

СИСТЕМНОСТЬ ОМОНИМИИ В СЛОВОСЛОЖЕНИЯХ

Омонимия присуща всем языковым объектам. Первый основной момент рассуждения состоит в том, что омонимия является объектом-системой, так как содержит все системообразующие категории: множество первичных элементов (омонимов), множество отношений единства и противоположности (одна и разные части речи, а внутри части речи – одна и разные грамматические позиции, один и разные корни), множество законов композиции в омонимическом узле (слои и связки). Второй момент, связанный с первым, – омонимизация является процессом-системой, что гарантирует обнаружение действия части или всех законов преобразования объектов-систем, приводящее к омонимизации. В третьих, омонимичность является отношением-системой. Следуя этим трем положениям, мы с необходимостью получим системные результаты.

Мы логико-математическим путем доказали необходимость существования этого феномена для любого языка мира, использующего триаду понятий – звук–буква–фонема и построили типологию омонимии в виде семи разновидностей [1]. При системном подходе омонимии можно определить

как симметрию формы и асимметрию содержания (ср. *прибыли* – глагол и однокоренное существительное, омонимичное не только с глаголом, но и внутри собственной парадигмы).

В работе «Омонимия как системная категория языка», выполненной в русле работ НИЛ ТиПЛ [2], в деталях рассмотрены проблемы симметрично-асимметричного характера омонимии. Поэтому здесь мы рассмотрим лишь некоторые аспекты, которые не нашли отражения в монографии и публикациях и будут касаться словосложений и аббревиатур.

Среди них важнейшим является мерность межпарадигматической омонимии. В случае полисемии (многозначности) все формы сопоставляемых слов совпадают, именно поэтому иногда сложно отличить омонимию от многозначности (ср. *ключ* – источник и инструмент для открывания замка). При омонимии такие явления редки (*пион* – цветок и пи-мезон). Чаще всего находится несколько неомонимичных форм, что дает право говорить о том, что в разряде существительных максимум омонимичности достигает 24 форм при минимуме в 2 формы, как это видно из сопоставления двух парадигм, представленных в таблице 1.

Т а б л и ц а 1

Парадигматическая и межпарадигматическая омонимия

	Единственное число		Множественное число	
	М. р.	Ж. р.	М. р.	Ж. р.
И.	змеИ	змеЯ	змеИ	змеИ
Р.	змеЯ	змеИ	змеЕВ	змеЙ
Д.	змеЮ	змеЕ	змеЯМ	змеЯМ
В.	змеЯ	змеЮ	змеЕВ	змеЙ
Т.	змеЕМ	змеЕЙ	змеЯМИ	змеЯМИ
П.	змеЕ	змеЕ	змеЯХ	змеЯХ

Из 24 падежных форм неомонимичными являются только две, что говорит о высокой степени омонимичности и маркированности этих форм (омонимичные формы в примерах выделены заглавной буквой). Новым в знании об омонимии является понятие «омоузла» как системы своего рода. Суть понятия «омоузел» заключается в том, что он представляет внутри-классовую (омонимичные формы принадлежат к одной части речи) и меж-классовую омонимию. Эта суть станет достаточно понятной после рассмотрения примера одного омоузла. Он представлен на матрице.

М а т р и ц а

Омоузел как система своего рода

Исход	1	2	3	4	5	6	7	8	9	10
1. пол	полу	поле	полom							
2. полom			полom							
3. пола	полу	поле		полой						
4. полой				полой	полое					
5. поле		поле				полю	полей	полем		
6. полоть						полю		полем	полоть	
7. полоть									полоть	
8. полить							полей			
9. полое			полom		полое					полого
10. полый			полom							полого
11. полая				полой						
12. полого										полого

Нулевой столбец представляет исход в виде словарных форм представления (в словаре дается именительный падеж единственного числа и форма инфинитива для глаголов). Следующие столбцы представляют омонимичные связки в виде форм разных парадигм. Здесь мы имеем 12 слоев и 10 омоузлов мощностью от двух до четырех форм. Омонимия при таком представлении базируется на линейной симметрии начал слов столбца и асимметрией финалей. В следующих за исходом столбцах имеет место симметрия форм и асимметрия значений. Если столбец исхода представляет словообразование (разные слова), а строки представляют словоизменение, то целый омоузел дает совмещенное представление о двух главных процессах в языке, приводящих к омонимизации целостной системы.

Представлением омоузла как системы своего рода, а именно, системы узлов (изоузел, гомоузел и т.п.), мы хотим показать, что омонимия сложных слов качественно отличается от омонимии слов, созданных аффиксальным способом. Здесь не будет узлов вышеуказанного вида, а для двусложных слов будут веерообразные открытые системы, ориентированные влево или вправо относительно омонимичного компонента. Для трехкомпонентных будет не две, а семь подсистем.

Т а б л и ц а 2

Веерообразное представление омонимии

	-бат	проф-	
	-кор	вол-	
	-див	воен-	
	-банк	глав-	
	-партия	парт-	
ком-	-вуз	гор-	ком
	-бриг	испол-	
			

Омонимия сложных слов, связывается с их структурными особенностями, так как в обычных словах одна упаковка может быть омонимичной и неомонимичной, а сложные слова и аббревиатуры как минимум двухкомпонентны. Этот момент с необходимостью требует рассмотрения таких общесистемных параметров как левизна, правизна и лево-правизна [1].

Более того, освоение омонимией элементов сложных слов ведет к дальнейшей экспансии омонимии, она начинает проявляться на уровне «полнозначное слово» – «элемент сложного слова» или «полнозначное простое слово» – «полнозначное простое слово», что в еще большей мере усложняет процесс автоматической обработки слов.

Омонимия сложных слов и аббревиатур возникает в процессе сжатия двух и более корней в одно слово и характеризуется с содержательной стороны тем, что омонимия системы увеличивается, т.е. сложные слова вносят свою долю в ее омонимизацию. Омонимии двух типов (омонимия элементов сложных слов и омонимия целого слова) сопутствует и полисемия, представляющая разные способы расширения значения.

Анализ материала показывает, что в структурном плане омонимия элементов сложного слова связывается с позицией омонимичного элемента: он может занимать первую или вторую позицию в двусложных словах и все три позиции

в трехсложных, при этом двусложные и трехсложные слова могут содержать омонимичный элемент в одной и той же или в разных позициях. Например, *армвоенюрист* – *армглавконцерт*, в первом случае элемент *арм-* обозначает *армейский*, во втором – *армянский*; *Уралчермет* – *гидрометцентр*; *райнарсуд* – *минсудпром*; *грузкурортстрой* – *грузавто-транспорт*; *центризбирком* – *белнацком* – *губкомпарт* – *помкомвзвода* – *главком*. Рассмотрение элемента *ком-*, встречающийся в трех позициях (начало, середина и конец слова), что он объединяет самые разные значения: *командир*, *коммунистический коммунальный*, *комиссар*, *командующий*, *ком-мерческий*, *комиссариат*, *комитет*, *коммуникационный* и др. (ср. *комбат*, *командарм*, *компартия*, *горкомхоз*, *военком*, *райвоенкомат*, *партком*, *инком-банк* и т.п.). И эта омонимия увеличивается на глазах в пределах одного конкретного элемента и охватывает все большую часть слов.

Можно легко представить, что в аббревиатурах степень омонимичности будет еще большей, так как чисто комбинаторные возможности там возрастают.

Омонимия обычных слов служит базой для появления омонимии своеобразного «второго порядка», вторичной омонимии уже сложных слов. К ней присоединяется и специфическая омонимия, возникающая между неомонимичным элементом сложного слова и другим, возникающая в результате усечения части этого элемента. Так, слово *мел* и *мелодия* не являются омонимами. Но усечение слова *мелодия* до элемента сложного слова *мело-* создает омонимичные отношения.

Вот некоторые примеры первых элементов, являющихся омонимичными: *мело* – *мелодия*, *мел*; *пар* – *пар*, *пара*; *соц* – *социальный*, *социалистический*; *авиа* – *авиационный*, *воздушный*; *авто* – *автомобильный*, *автоматический*, *собственный*, *самодвижущийся*; *архи* – *старшинство*, *высшая степень чего-либо*; *био* – *относящийся к жизни*, *биологический*; *обще* – *общий*, *обязательный для всех*, *свойственный*, *охватывающий отрасли*, *распространяемый на всех*, *охватывающий что-либо целиком*; *пол* – *половина*, *середина*, *полномочный*; *само* – *действие направлено кем-н. на самого себя*, *действие совершается без посторонней помощи*; *сверх* – *превышение предела*, *высшую степень качества*; *спец* – *специальный*, *специалист*; *средне* – *средний период чего-н.*, *срединное положение*, *средняя норма чего-н.*; *суб* – *нахождение внизу*, *подчиненность*, *второстепенность*; *супер* – *расположение сверху*, *главенство*, *высшая степень качества*, *вторичность*; *теле* – *действующий на дальнее расстояние*, *телевизионный*; *транс* – *движение через какое-либо пространство*, *следование за чем-либо*, *сообщение через посредство чего-либо*; *фото* – *основанный на действии света*, *фотографический*; *лизо* – *часть от лизать (лизоблюд)* и вещество (*лизокефалин*).

Отмечается омонимия элементов сложных слов и в ряде других случаев, касающихся межъязыковой омонимии в заимствованиях. Например, *пара/плегия* – *пара/хронизм* – *пара/нефрит*, в первом слове элемент *пара* – обозначает *вдоль* (от гр. *para* – *вдоль*, *plexis* – *связывание* – *паралич* обеих конечностей), во втором – *вопреки* (от гр. *para* – *вопреки*, *chronos* – *время* – *хронологическая ошибка*), в третьем – *возле*, *около* (от гр. *para* – *возле*, *около*, *perhros* – *почка* – *воспаление околопочечной клетчатки*).

В ряде случаев отмечается омонимия вторых элементов. Компонент – *граф* вступает в омонимические отношения по семантическому признаку

(предмет и носитель профессионального признака). Отмечена омонимия по грамматическому признаку «одушевленность-неодушевленность»: *телеграф* и *фотограф*; *путеводитель* и *автоводитель*; *животновод* и *нитевод*. Отмечена лексико-грамматическая омонимия: *богомол* – *мукомол* (молиться и молоть); *треугольный* и *древесноугольный* (угол и уголь); *стократ* и *демократ* (кратность и «кратос» – власть); *вещдок* и *плавдок* (доказательство и док); *госдача* и *кроводача* (дача и давать); *животновод* и *нитевод* и *углевод* (разводить, проводить и водород – этот случай особо показательный, так как показывает, что сложное слово-калька «водород», введенное в систему, подвергается дальнейшему сжатию, что и приводит к омонимии – ср. ряд водород – углеводород – углевод); *аммонолиз*, *автолиз* и *блюдолиз* (*лиз* – саморастворение и усеченная основа от *лизать*); *аэроплан* и *генплан* (*план* от планировать в значении *плавно летать*, *парить* и *план* от *планировать*); *автократ* и *стократ* (*крат* – власть и *крат* от кратно). Отмечена омонимия между первым и вторым элементом: *бармен* и *миллибар* (*бар* – небольшой ресторан, *бар* – единица измерения давления).

Омонимия словосложений имеет симметрично-асимметричный характер, поэтому система омонимии в словосложениях может быть представлена в терминах симметрии/асимметрии, как комбинаторика трех признаков: формальное и смысловое единство первых корней (Э1), единство формы и значения опорных компонентов (Э2), идентичность слов в плане грамматического значения и его выражения (Гр.). Plusом обозначим совпадение каждого из перечисленных признаков, а также полное несовпадение. Очевидно, здесь требуется пояснение. Понятно, что если не только форма, но и значение компонентов в составе разных слов совпадает, то единство этих компонентов сохраняется (например, одинаковые первые корни в словах *правовед* – *правонарушение* или однокоренные опорные компоненты в словах *золотоискатель* – *богоискатель*). Но в словах с разными, не совпадающими формально, следовательно, и семантически, компонентами также сохраняется единство отношений – как умножение минуса (несовпадение форм) на минус (несовпадение значений) дает плюс (сохранение отношений).

Тогда минусом обозначим омонимичные отношения между словами или их составляющими элементами (единые формы +, но разные значения –). Первый минус будет означать совпадение формы первых компонентов при разном значении, второй минус – формальное совпадение семантически различных опорных элементов, третий – формально одинаковое выражение разных грамматических значений. Теоретически система состоит из восьми и только восьми подсистем.

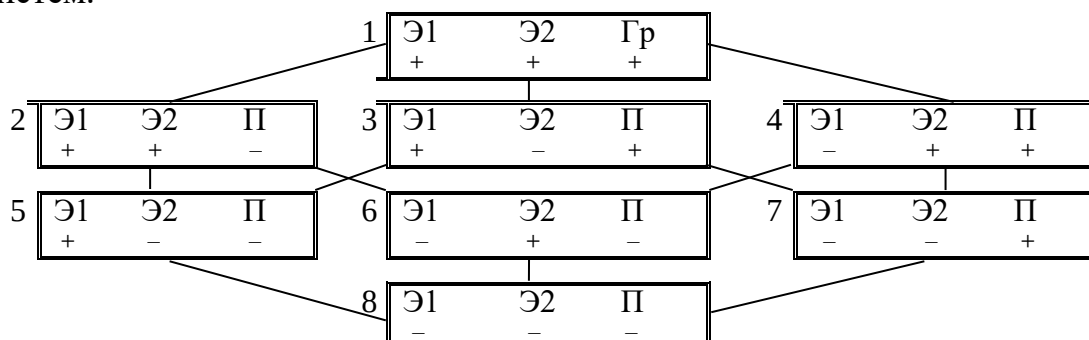


Рис. 1. Симметро-асимметрия словосложений

Наибольшее наполнение имеют системы, содержащие один омонимичный признак.

Подсистема 4 заполняется сложными словами, в которых первый элемент обладает лексической омонимией: 1.авто... 1) ‘автомобильный’: *автогараж, автогонки, автодеталь, автопробег*; 2) ‘автомобиль’: *автобензовоз, автобеномешалка, автозаправочный, автосамосвал*; авто...2 – ‘автоматический’ *автоподача, автосварка, автосцепка*; авто...3 – ‘самоходный, самодвижущийся’: *автогрейдер, автодрезина*; авто 4 – ‘свой, собственный, само...’: *автобиография, автогравюра, правописание, правонарушение, правобережье*.

Подсистема 3 включает слова с лексически омонимичными опорными элементами и разными или совпадающей предшествующей частью (*каучуконос сахаронос, электролит – аэролит*).

Подсистема 2 – наполняется словами, различающимися грамматическими характеристиками, с семантически одинаковыми опорными компонентами. Этим условиям отвечают слова с грамматической омонимией опорных компонентов (*льновод, зверовод, цветовод – групповод, экскурсовод, железосплав, ферросплав – самосплав, лесосплав, водонос (одуш.) – медонос (неодуш.)*).

Подсистемы следующего уровня имеют значительно меньшее наполнение.

Подсистема 7, представляющая собой развитие признаков 3-й и 4-й подсистем, непродуктивна: лексическая омонимия обеих частей представлена единственным примером *Минзаг* (Министерство заготовок) – *минзаг* (минное заграждение). Но ее могут заполнять и сложные слова-омонимы: *микрометр* – ‘инструмент с микротермическим особо точным винтом для измерения контактным способом’ и *микрометр* ‘дольная единица длины’, *костолом* (разг., неодуш.) – ‘болезнь, сопровождающаяся ломотой в костях’ и *костолом* (прост., одуш.) – ‘тот кто в драке ломает кости, наносит увечье противнику’, ‘минерал’, *гелиотрон* – ‘геодезический прибор’; *мухоловка* 1. Приспособление для ловли мух, 2. Насекомоядное растение семейства росянковых и *мухоловка* – 1. Небольшая птица отряда воробьиных, питающаяся насекомыми, 2. Членистоногое животное отряда губоногих, охотящееся за мухами и другими насекомыми; *душевнобольной* (прил.) и *душевно-больной* (сущ., м. р.).

Подсистема 6 представляет собой развитие признаков 2-й и 4-й подсистем; практического наполнения не имеет – это объясняется тем, что омонимия чисто грамматическая предполагает совпадение опорных компонентов (так как два последних признака объединяются общим планом выражения), а сочетание такого опорного компонента с омонимичной предшествующей частью затруднительно. Для иллюстрации придумаем омоним слову (звезда – источник радиоизлучения) *подиумзвезда, шоузвезда* **подиумзвезда* (звезда подиума) и **шоузвезда* (звезда шоу) аналогично *телезвезде*.

Подсистема 5 объединяет признаки 2-й и 3-й подсистем с лексикограмматической омонимией опорных элементов. Первые части слов в таких парах будут различаться: *стеклобой, мордобой, политбой* (существительные) – *вразнобой* (наречие); *дискобол* (одуш.), – *пионербол, мотобол*’ (неодуш.), *дискоболка* (одуш.) – *футболка*’ (неодуш.), *линкор* (неодуш.) – *юнкор* (одуш.), *лорд-мер* (одуш.), *секундомер* (неодуш.).

Подсистема 1 максимально наполнена. Она включает все сложные слова, в которых нет омонимичных элементов: *высокомерность, гематология, десятиборье, доброжелательность* и т.п. Наличие этой подсистемы подтверждает симметрично-асимметричный характер омонимии словосложений.

Подсистема 8, которая содержит три омонимичных признака, может быть наполнена немногочисленными примерами типа *полочка* ‘маленькая полка’ и *пол-очка* ‘половина очка’, *волком* (от волк) и *волком* ‘волостной комитет’, *ползала* (от ползать) и *ползала* ‘половина зала’ [2; 1].

Наличие трех типов омонимии в сфере сложных слов (лексической, грамматической, лексико-грамматической) показывает неодинаковую качественную и количественную наполненность. Омонимия присуща более всего сложным существительным, которые, как известно, составляют большинство сложных слов. Вместе с тем отмечаются случаи омонимизации компонентов других частей речи, не являющихся производными сложных существительных, например: *прямо/угОльный, тре/угОльный и пыле/угольный, каменно/ угОльный; паро/образовательный, слово/образовательный* (1. образовать – создать) и *обще/образовательный, само/образовательный* (2. образовать – обучить).

Использованный выбор классификационных признаков позволил рассмотреть омонимию в рамках композитов и их однопозиционных составляющих. Омонимичными могут быть и разнопозиционные компоненты, например: *пед/совет – мото/пед, бар/мен – кило/бар, моль-малютка – кило/моль, вод/хоз – ското/вод, глав/реж – власо/глав, губ/исполком – душе/губ, кол/договор, кол/хоз – дрово/кол, дыро/кол* и др. Это ставит новые задачи исследования лексической левизны, правизны и лево-правизны.

ЛИТЕРАТУРА

1. Карпов, В. А. Язык как система / В. А. Карпов. – Минск : БГУ, 1992.
2. Головня, А. И. Омонимия как системная категория языка / А. И. Головня. – Минск : БГУ, 2007.

Е. А. Гурбик

АНГЛИЙСКИЙ ЯЗЫК В СРЕДСТВАХ МАССОВОЙ ИНФОРМАЦИИ ПАКИСТАНА

Важной сферой функционирования английского языка в Пакистане являются средства массовой информации (СМИ). Первое место по значимости среди них занимает пресса, включающая печатные газеты и журналы. За период от основания Пакистана по сегодняшний день печатные издания претерпели многие изменения как по количеству наименований газет и журналов и их тиражу, так и по стилю изложения материала.

В настоящее время основными читателями газет и журналов являются жители городов. Население же отдаленных районов провинций зачастую вовсе не имеет доступа к каким-либо масс-медиа. Главная причина подобной ситуации заключается в особо критическом экономическом состоянии некоторых зон страны и недоступности любых СМИ их жителям.

Второй причиной непопулярности прессы среди многих слоев общества можно назвать низкий уровень грамотности в стране, особенно в малых городах и сельских районах, вследствие чего тираж газет и журналов, в первую очередь на английском языке, значительно мал.

На сегодняшний день в Пакистане газеты и журналы издаются в основном на урду и английском языке, с незначительным числом газет на таких местных региональных языках, как синди, пенджаби, сарайки, гуджрати, балочи и пашто. Наибольшее количество газет и журналов в Пакистане издается на урду.

За последние двадцать лет наблюдается развитие прессы и улучшение качества ее производства. Большинство газетных издательств принадлежит частным владельцам. В число наиболее крупных издательских компаний входят Jang, Dawn и Nawa i Waqt, печатные издания которых доступны в основных городах страны как на АЯ, так и на урду.

В число газет и журналов крупнейшей издательской компании Jang входят Daily Jang (ежедневная газета на урду), Al Awaam (ежевечерняя газета на урду), Akhbar e Jahan (журнал на урду); The News (ежедневная газета на английском языке), Mag (еженедельная газета на английском языке).

Dawn (ежедневная газета на английском языке), Herald (ежемесячная газета на английском языке) и Star (ежевечерняя газета на английском языке) публикуются второй по значимости в Пакистане издательской компанией Dawn.

The Nation (ежедневная газета на английском языке), Nawa i Waqt (ежедневная газета на урду), Nida e Millat (еженедельная газета на урду) и Family (журнал на урду) публикуются третьей крупной издательской компанией Nawa i Waqt.

В соответствии с отчетом Бюро статистики Пакистана [1], в 2007 году общий тираж ежедневных газет в Пакистане составлял 9 934 951 экземпляров. Тираж прессы на урду насчитывал 7 564 250, в то время как тираж газет и журналов на английском языке составлял 1 206 557 экземпляров. Приведенные статистические данные свидетельствуют об очевидном количественном доминировании прессы на урду.

По данным Бюро статистики Пакистана [2], в 2010 году в стране издавалось 952 наименования газет и журналов на урду и английском языке. Из них издания на урду составили 761 газету и журнал, тогда как количество изданий на английском языке составило 84. Наибольшее число периодических изданий на английском языке в провинциях страны публикуется в провинции Балочистан, где по состоянию на 2010 год насчитывалось 28 газет и журналов. Второе место по количеству прессы на английском языке занимала провинция Синд (26 газет и журналов). 20 англоязычных журналов и газет издавалось в провинции Пенджаб. Таким образом, на долю этих трех провинций Пакистана приходилось 74 издания на английском языке из 84 существовавших в то время в стране. Достоверных данных, проясняющих причины наибольшего количества СМИ на английском языке именно в провинциях Балочистан и Синд, нет, однако можно сделать предположение о стратегической значимости этих провинций и их географического расположения. Именно в провинциях Балочистан и Синд сконцентрировано значи-

тельное количество вооруженных сил Пакистана, которые не только всегда поддерживали негласную языковую политику по усилению позиции английского языка в Пакистане, но и сами являются инициаторами некоторых газет и журналов на этом языке.

По данным Газетного общества Пакистана (All Pakistan Newspaper Society), в 2015 году в Пакистане издавалось 58 газет и журналов на английском языке [3], что свидетельствует о критическом состоянии англоязычной прессы.

Необходимо отметить, что правительство Пакистана практикует двойственную политику по отношению к СМИ на английском языке и урду. Несмотря на наличие 19 статьи Конституции Пакистана 1973 года [4, р. 12] о свободе прессы в рамках закона, а также дополнительно принятого в 2002 году Постановления о свободе информации (Freedom of Information Ordinance) [5], власти Пакистана осуществляют негласный контроль и управление прессой. Однако политика сдерживания общественного мнения распространяется только на урдуязычную прессу.

Первая причина особой политики по отношению к прессе на урду заключается в численном преобладании читателей газет и журналов на этом языке, тогда как пресса на английском языке читается преимущественно элитой. Следовательно, стиль подачи информации и подбор материала контролируется правительством негласно.

Вторая причина двойственной политики правительства по отношению к прессе состоит в тематической направленности СМИ: газеты и журналы на урду дают более подробное и глубокое описание обстановки в стране. Именно эти издания определяют общественное мнение и настроение широких слоев населения в Пакистане.

Пресса на английском языке отличается большей либеральностью вследствие отсутствия контроля со стороны правительства касательно тематики и стиля изложения информации. В основном статьи в англоязычной прессе направлены на отражение главных новостей и событий в стране, связанных с политической, экономической, культурной и спортивной обстановкой в Пакистане.

Газеты и журналы на английском языке пользуются популярностью за пределами страны и издаются для иностранных граждан, а также для пакистанских иммигрантов, проживающих в разных странах мира.

Важной причиной общего снижения числа печатных изданий на английском языке и урду является распространение новых технологий и появление электронных газет, а также многочисленных информационных сайтов. Огромное количество информации на английском языке доступно в электронной форме. Все большее число населения Пакистана, читающее материалы на английском языке, предпочитает использовать электронные источники; сведения же на урду в Интернете не настолько разнообразны и тематически обширны.

Количественное преимущество газет и журналов на урду не позволяет прессе на английском языке доминировать в пакистанском обществе. Однако, поскольку английский язык в Пакистане реализует немаловажную функцию демонстрации власти, прессе на английском языке принадлежит элитный статус среди периодических изданий.

Вторым по значимости видом СМИ в Пакистане является телевидение, на котором полностью доминируют каналы на урду. Число передач на английском языке на пакистанских каналах и продолжительность их вещания незначительны. По данным Бюро статистики Пакистана (Pakistan Bureau of Statistics) [6], в 2010 году количество часов трансляции передач на английском языке значительно снизилось по сравнению с периодом с 2001 года по 2009 год.

В настоящее время в Пакистане единственными пакистанскими телевизионными каналами, ведущими трансляцию всех программ на английском языке, являются PTV World и Dawn TV. Телеканал Dawn TV транслирует исключительно новости 24 часа в сутки. PTV World представляет более разнообразную программу передач, включающую новости (региональные и международные, бизнес, спорт, погода); аналитические и обозревательские программы; развлекательные передачи (сериалы, реалити-шоу, культурные и религиозные программы, документальные фильмы) и музыкальные программы. Передачи этого канала транслируются не только в Пакистане, но и в странах Аравийского полуострова, Иране, Афганистане, Таджикистане и Бангладеш.

Все другие телевизионные каналы, ведущие трансляцию как в Пакистане, так и в некоторых других странах мира, предоставляют набор программ преимущественно на урду. Наряду с урдуязычными существуют каналы и на других региональных языках, зона трансляции которых охватывает лишь ограниченную территорию отдельных провинций.

Преобладание урдуязычных телевизионных каналов и передач объясняется двумя основными причинами, первая из которых связана с незначительным числом населения Пакистана, владеющего английским языком и способного воспринимать англоязычную речь.

Вторая причина заключается в наличии большого количества американских и британских каналов, ведущих трансляцию программ на английском языке для стран Южной Азии и Пакистана в частности. Наличие таких каналов, с одной стороны, дает возможность населению Пакистана смотреть разнообразные западные программы на английском языке и, с другой стороны, позволяет правительству государства поддерживать политику двуязычия в стране, не прикладывая особых усилий и затрат в этой области. К числу иностранных каналов на английском языке в Пакистане относятся BBC World News (Великобритания), CNN (США), DW-TV (Германия), NHK World (Япония), WB Channel (совместно Индия – США), Animal Planet (США), Discovery Channel (США) и Disney World (совместно Индия – США).

Отдельную группу иностранных телеканалов на английском языке составляют те из них, которые транслируются исключительно для Пакистана. Тематика передач этих каналов требует исключения некоторых передач и фильмов, не соответствующих заповедям Ислама как официальной религии государства. На сегодняшний день в стране транслируются телеканалы США, заключившие контракты с пакистанскими телекомпаниями. К группе телеканалов, созданных совместно Пакистаном и США, относятся HBO Pakistan, AXN, Cartoon Network и Nickelodeon.

Телевизионные каналы на английском языке не занимают нишу элитных СМИ. В силу того, что практически все англоязычные каналы являются иностранными и предлагают западные программы передач, телевидение носит нейтральный характер по отношению к политической и социально-экономической обстановке в стране.

И наконец, еще одним видом СМИ, заслуживающим внимания в языковой ситуации Пакистана, является радио. Радиовещательные каналы по-прежнему пользуются немалой популярностью среди населения Пакистана. Самой главной тому причиной является доступность радиовещания, поскольку это самый дешевый для пользователей вид СМИ. Легкий доступ к радиовещательным каналам, вне зависимости от местонахождения человека, позволяет этому виду СМИ занимать важное место в Пакистане и по охвату аудитории быть наравне с прессой и телевидением.

По данным Управления контроля электронных СМИ Пакистана (Pakistan Electronic Media Regulatory Authority) [7], в 2015 году лицензия была выдана 141 коммерческому и 45 некоммерческим радиостанциям по всему Пакистану. В этот список включены все функционирующие радиостанции страны, проводящие трансляцию радиопередач, как на различных региональных языках, так и на урду и английском языке.

По данным отчета исследовательской организации Gallup в Пакистане, по состоянию на 2002 год [8] число радиослушателей каналов FM составляло 39 % от общего населения страны.

Около 18 радиостанций, ведущих передачи на урду и английском языке, осуществляют трансляцию по всему Пакистану. Единственной коммерческой радиостанцией в стране, все радиопередачи которой проводятся на английском языке, является City FM 89.

Помимо коммерческих радиостанций, лицензии выдаются и университетам, радиовещание которых проводится преимущественно на английском языке. Передачи радиостанций университетов носят образовательный характер, и поскольку высшее образование в Пакистане дается на английском языке, трансляция радиопередач ведется именно на этом языке.

Прогнозируя будущее развитие СМИ в Пакистане, можно говорить о маловероятном изменении ситуации в пользу увеличения количества СМИ на английском языке. Это, прежде всего, связано с очень медленным прогрессом в сфере образования и грамотности: только при условии стопроцентной грамотности населения можно будет говорить о количественном увеличении и качественном расширении границ англоязычных СМИ. Второй фактор маловероятности перемен в состоянии СМИ на английском языке – удовлетворенность нынешним состоянием не только владельцев самих СМИ, но и правящей верхушки и элиты Пакистана. Поскольку только элита и правящие слои общества владеют английским языком, увеличение или сужение количества СМИ на этом языке не влияет на основную массу населения, которое английским языком не владеет. В настоящее время СМИ Пакистана на английском языке находятся под сильным влиянием глобализационных процессов и «информационного взрыва», которые определяют их развитие.

ЛИТЕРАТУРА

1. Newspapers and Periodicals [Electronic resource] / Average circulation of newspapers and periodicals by language/ type (Numbers). – Pakistan Bureau of Statistics. – Mode of access : <http://www.pbs.gov.pk/sites/default/files/other/yearbook-2008/Social%20&%20Cullture/19-3.pdf>. – Date of access : 23.01. 2016.
2. Newspapers and Periodicals [Electronic resource] / Newspapers and periodicals by language and province. – Pakistan Bureau of Statistics. – Mode of access : <http://www.pbs.gov.pk/sites/default/files/other/yearbook2011/Social%20&%20Cullture/19-2.pdf>. – Date of access : 23.01. 2016.
3. Periodicals in English [Electronic resource]: Member publications in the English language. All Pakistan Newspapers Society. – Mode of access : http://apns.com.pk/-member_publication/language_wise.php?language=english. – Date of access : 24.01.16.
4. The Constitution of the Islamic Republic of Pakistan [Electronic resource] / Article 19: Freedom of Speech etc. – National Assembly of Pakistan. – Mode of access : http://www.na.gov.pk/uploads/documents/1333523681_951.pdf. – Date of access : 24.01.16.
5. Ordenances [Electronic resource] / Freedom of Information Ordinance 2002. – Government of Pakistan Ministry of Information, Broadcasting and National Heritage. – Mode of access : http://infopak.gov.pk/Downloads/Ordenances/-Freedom_of_%20Information_Ordinance2002.pdf. – Date of access : 24.01. 2016.
6. Telecasting [Electronic resource] / Telecasting hours by language. – Pakistan Bureau of Statistics. – Mode of access : <http://www.pbs.gov.pk/sites/default/-files/other/yearbook2011/Transport%20&%20Comm/20-29.pdf>. – Date of access : 25.01. 2016.
7. Radio stations [Electronic resource] / List of commercial FM radio licenses. – Pakistan Electronic Media Regulatory Authority. – Mode of access : <http://www.pemra.gov.pk/>. – Date of access : 25.01. 2016.
8. Radio stations [Electronic resource] / Current level of radio listenership. – Gallup Pakistan. – Mode of access : <http://gallup.com.pk/gallup-pakistan-media-report/>. – Date of access : 26.01. 2016.

Н. П. Дарчук

КОМПЬЮТЕРНАЯ ГРАММАТИКА УКРАИНСКОГО ЯЗЫКА *АГАТ* И ПЕРСПЕКТИВЫ ЕЕ ИСПОЛЬЗОВАНИЯ

Потребность в лингвистическом обеспечении систем «человек – машина – человек» обусловлена необходимостью удобной кооперации человека и машины, опирающейся на естественный язык. В социальном плане значимость лингвистических проблем компьютеризации связана с такими основными направлениями индустрии обработки знаний, как сбор, хранение, систематизация, распространение, интерпретация информации, для чего создается специальное лингвистическое обеспечение.

Лингвистическое обеспечение автоматических систем – совокупность средств для осуществления компьютеризации языковой деятельности – необходимо практически для любой интеллектуальной деятельности человека. С технологической точки зрения речь идет о создании того или иного типа автоматической системы обработки информации, на входе и выходе которой текстовая информация на естественном языке. Типы систем разно-

образны и направлены на моделирование разных языковых задач, в частности, таких как диалоговое взаимодействие, сжатие информации, реферирование текста, логическая обработка содержания, перевод и т.д. [1; 2; 3]. Прикладные системы, которые создает лингвист в этой области, – это **лингвистически осмысленные метаязыки** – модели представления знаний, каждая из которых основывается на **теоретических положениях языкознания** и реализуется с помощью методов **структурно-математической лингвистики**. Воплощенная в прикладные задачи диалектическая трихотомия «традиционная лингвистика – структурно-математическая лингвистика – компьютерная лингвистика» способствует высокому уровню лингвистического обеспечения автоматических систем.

Для автоматического анализа украинскоязычного текста сотрудниками лаборатории компьютерной лингвистики Института филологии Киевского национального университета имени Тараса Шевченко создана компьютерная грамматика АГАТ. **Компьютерная грамматика** – это системное, строго упорядоченное, формализованное, лингвостатистическое, интегральное описание знаковых единиц определенного языка в виде структурных моделей с необходимой и достаточной аналитикой для выполнения задач искусственного интеллекта, которые воспроизводят и имитируют исследовательскую деятельность лингвиста [4]. Компьютерная грамматика АГАТ имеет такие особенности: **уровневый подход** (уровни взаимодействуют между собой от нижнего до верхнего, каждый следующий уровень использует результаты анализа предыдущего); **открытость** стратификационной структуры грамматики, что является принципиальным моментом, поскольку позволяет довольно точно расширять объем лингвистического обеспечения, усложнять словарное и модифицировать программное обеспечение без перестройки всей системы, которая является иерархическим комплексом компьютерных моделей: морфемно-словообразовательной, морфологической, синтаксической моделей, построенных на основе формальных, точных и однозначных правил. Эти модели можно считать исследовательскими, потому что заложенные в грамматики алгоритмические правила приводят к выявлению того или иного языкового явления (морфов, словоформ с их частеречными и категориальными характеристиками, словосочетаний, деревьев зависимостей предложений и т.п.). Алгоритмически симитирована деятельность лингвиста – а именно обеспечен переход от совокупности текстов к системе, которая лежит в их основе, выявлены элементарные единицы и классы элементарных единиц. Разработанные модели являются моделями анализа, индуктивными, несемантическими и детерминистскими (структурными).

Язык – относительно открытая система относительно открытых подсистем, каждую из которых можно моделировать и устанавливать определенные отношения между подсистемами. Как свидетельствует практика, большинство современных лингвистических процессоров модульного типа, адекватных поэтапному членению процесса анализа, отвечает уровням языка: фонемному (единица фонема/аллофон), морфемному (морфема/морф), лексическому (лексема/словоформа), синтаксическому (модели словосочетаний, модели предложений). Семантический модуль не имеет уровня

соответствия в языковой системе, но он на семном уровне завершает (на данном этапе разработки) смысловой анализ украинского текста. Все единицы модулей взаимодействуют (АГАТ отражает их последовательную работу), но морфологические, синтаксические, семантические характеристики обрабатываются разными алгоритмами и программами с разным словарным обеспечением. Системно адекватным этому является такое лингвистическое обеспечение, которое имеет открытый модульный характер, и в некоторых случаях предусматривается их совмещенность для пополнения данных и коррекции результатов (например, после автоматического морфологического анализа обоих этапов остается несколько процентов словоформ с неснятой омонимией, которая окончательно снимается на синтаксическом уровне). Система может выполнять отдельные конкретные задачи анализа (напр., только морфологического). При этом открытость является главным фактором эффективного функционирования АОТ, гарантией того, что для введения новой информации не потребуется перестраивать всю систему.

Словарное обеспечение является информационной основой модели и реализуется в виде иерархической, модульно совместимой и открытой системы. Компоненты, которые представляют лингвистическую модель АОТ, это лингвистические процессоры, которые последовательно, один за другим обрабатывают входной текст. Вход одного процессора – это выход из другого. В созданной системе выделяются такие модули:

- Морфологический анализ. Построение морфологической аннотации слов входного текста;
- Синтаксический анализ. Выделение словосочетаний. Построение дерева зависимостей всего предложения;
- Морфемный анализ. Членение входного текста на морфы;
- Семантический анализ. Построение тезауруса текста.

Для каждого уровня разработан метаязык его представления – константы и правила их комбинации. На морфологическом уровне – константами являются граммы (род, число, падеж, время, лицо, наклонение), на морфемном – тип морфа, на синтаксическом – тип словосочетания, тип связи, на семантическом – семантические категории. Важным фактором является поступательность в работе модулей: следующий анализатор улучшает результаты предыдущего уровня. Например, на синтаксическом уровне «доснимается» грамматическая и лексико-грамматическая омонимия, а семантический анализатор помогает достроить синтаксический граф.

Обязательной составляющей компьютерной грамматики является **автоматический морфологический анализ (АМА)** словоформ, потому что ни морфемный, ни синтаксический, ни семантический анализы не могут обойтись без определения для словоформы ее частеречной характеристики и словоизменяемых форм. В **задачи АГАТ-морфологии** входят: автоматическое определение для единиц текста грамматической информации о месте их в морфологической системе языка; автоматическая идентификация словоформ одной лексемы. Морфологическая аксиоматика была налажена на возможность алгоритмического оперирования грамматическими данными. В прикладном аспекте создан словарь квазиоснов объемом в 210 тыс. оди-

ниц и, соответственно, словарь словоформ, которые порождаются соединением информации, взятой из таблицы основ и вспомогательной таблицы, – приблизительно 3,2 млн. словоупотреблений, что обеспечивает приписывание морфологической информации словоформам практически на 97 %. Методологически АГАТ-морфология украинского языка является автоматическим формально-морфологическим процессором с элементами морфолого-синтаксического анализа. Особое внимание при создании АГАТ-морфологии уделено определению речевых условий, в которых реализуются актуализированные грамматические значения единицы-омонима. Был создан Грамматический словарь омоформ, а также определены речевые условия для реализации значений исследуемой словоформы, сформулированные с помощью лингвистического метода – контекстного анализа (КА). В основе КА лежит утверждение о том, что многозначные элементы языка функционируют в своих конкретных значениях в определенном лексико-грамматическом контексте. Реализация этой идеи нашла отображение в создании автоматического конкорданса, теоретической основой которого являются:

1) наличие таких определителей, по которым каждое значение словоформы (грамматическое, лексическое) детерминируется в контексте другими словоформами, их сочетаниями или другими текстовыми признаками;

2) текстоцентрический подход к его созданию: он составляется на определенном массиве текстов для определенной словоформы или лексемы. Такой словарь-конкорданс исчерпывающе иллюстрирует использование определенной лексики и всех ее ЛСВ с лексико-грамматическими значениями.

АГАТ-синтаксис компьютерной грамматики украинского языка создавался как лингвистический процессор, настроенный на моделирование синтаксической структуры входного предложения на уровне словосочетаний (1-й этап) и дерева зависимостей (2-й этап). Результат анализа – синтаксическая структура предложения, которая является совокупностью данных о синтаксических связях слов / словоформ в словосочетании – минимальной единице предложения.

АГАТ-синтаксис базируется на **формально-синтаксической теории** представления предложения. Это комплекс алгоритмических операций, которые выполняются над цепочками информации морфологического характера, представленными в исходном тексте, для установления синтаксических связей между текстовыми единицами. Практическая реализация теоретических положений осуществлялась путем взаимодействия двух структурных методов: для представления синтаксической структуры предложения в терминах словосочетаний применен **метод непосредственно составляющих**, а структуры целого предложения – **дерево зависимостей**. Алгоритмически и программно в синтаксическом модуле можно осуществлять **переход от непосредственно составляющих к дереву зависимостей**: корнем дерева является глагол-сказуемое, в узлах предложения находятся словоформы, каждая дуга дерева, которая связывает пару узлов, интерпретируется как подчинительная связь.

Значение АГАТ-синтаксиса заключается в том, что, опираясь на теоретический синтаксис в делении словосочетаний на **именные, адъективные,**

местоименные, глагольные и наречные, можно автоматически выявлять тип сочетаемости – **предикативный, подчинительный и сочинительный** – каждого полнозначного слова в тексте. В соответствии с концепцией АСА при выделении словосочетаний предусматривался предварительный этап – создание грамматики валентностей с подграмматиками для глагола, существительного, адъектива, а также словника фразеологизмов и коллокаций. Созданием такой информационной базы в виде подграмматик валентностей **были расширены возможности украинского теоретического синтаксиса в получении** из текстов разного стилевого и жанрового направления **информации** о конструктивных возможностях сочетаемости каждой части речи и типовых моделях для определенной части речи. Установление по правилам АГАТ-синтаксиса для каждого слова **подчинительных, предикативных и сочинительных** типов связей является воспроизведением общей системы отношений между компонентами описываемой ситуации в предложении. Перспективу АГАТ-синтаксиса мы видим в соединении его с семантикой, точнее – в соединении лексики и грамматики, поскольку влияние семантики на сочетаемость общепризнано и виды синтаксической связи между связанными словами являются производными от их семантики.

Разработка АГАТ-синтаксиса в пределах АОТ связана с общетеоретической необходимостью изучения сочетаемости лексических единиц, что открывает возможность в современной украинистике исследовать грамматическую и лексическую валентность слов, моделировать типовую сочетаемость классов слов, синонимию словосочетаний разных структурных типов, опираться на лексическую и грамматическую валентность как критерий синонимичности, изучать законы комбинаторики словосочетаний разных типов и разрядов или разграничение свободных и фразеологических словосочетаний, взаимодействие стойкости и идиоматичности и пр. Несмотря на то, что перечисленные проблемы так или иначе рассматриваются в теоретической грамматике, АГАТ-грамматика открывает новые перспективы в **исследовании живой лингвистической реальности, каковой являются тексты**. Применение АСА к Корпусу украинского языка дает возможность исследователям украинского языка установить синтаксическую и семантическую емкость такой единицы, как словосочетание, а в прикладном плане – разработанный автоматический синтаксический модуль анализа украинского текста – это механизм, с помощью которого становится реальностью составление по крайней мере двух словарей: частотного словаря словосочетаний и частотного словаря сочетаний простых предложений в сложном.

С помощью АГАТ-синтаксиса компьютер «поднимается» еще на одну ступеньку в процессе «понимания» содержания текста, приближаясь к разрешению конечной задачи АОТ – построению его семантического представления. Если АГАТ-морфология в терминах лексико-грамматических классов слов обеспечивает «понимание» компьютером денотативной информации, представленной в тексте, то АГАТ-синтаксис открывает путь к релятивной информации, т.е. к пониманию семантико-синтаксической структуры предложения. Синтаксические связи не существуют без семантических. И если не понятна синтаксическая структура предложения, не понятен и его смысл.

Задача **АГАТ-морфемики** заключается в возможности осуществлять лингвистические исследования морфемной и словообразовательной структуры, а именно: 1) составлять алфавитно-частотные словари всех типов морфов на базе текстов разных стилей и жанров; 2) объединять алломорфы в морфему; 3) устанавливать системные и функциональные характеристики морфем; 4) автоматически конструировать морфемно-словообразовательные гнезда. Информационной основой морфного сегментатора АГАТ-морфемики являются две базы данных – 170 тыс. слов и 3,5 млн. словоформ, в которых каждое слово и его словоформа представлены в виде морфной модели с информацией о типах морфов, их структурные отношения с другими морфами. В базовом словаре омонимичным корням (около 3100 единиц) и корневым алломорфам приписываются индексы из списка омонимических корней, а корневому алломорфу (около 2900) – инвариантная форма. Формализация морфных структур слов через описание их в терминах программных процедур морфной базы данных позволяет создавать на основе этой базы данных автоматизированную систему анализа, способную выполнять ряд таких **прикладных задач**: 1) группировать лексику по общеаффиксальным классам; 2) классифицировать лексику по морфным моделям; 3) создавать корневые и аффиксальные словари с учетом омонимии и алломорфии. Такое формализованное описание морфной структуры предусматривает моделирование структурных отношений морфов в двух планах организации слова как языкового знака: в плане выражения и плане содержания.

Работа морфемного модуля АГАТ-морфемики осуществляется самостоятельно и в связи со словообразовательным модулем с целью автоматического построения словаря морфемно-словообразовательных гнезд. Для этого: 1) группируется лексика в общекорневые выборки по процедуре идентификации корневых морфов, выделенных в словах морфемной базы; 2) на основе выборки всех общекорневых слов строятся словообразовательные гнезда как статьи электронного словообразовательного словаря на основе теоретических принципов словообразовательной производности. Формализованное описание морфных структур, предложенное при составлении электронного морфемного словаря, позволяет использовать морф как инструмент в проведении автоматического морфемного анализа других лексикографических систем.

Компьютерная грамматика вообще и АГАТ-грамматика в частности, является динамическим механизмом, который является системой правил оперирования грамматическими значениями с акцентом на формы их выражения, с помощью которых компьютеру открывается доступ к денотативной и релятивной информации. **Компьютерные словарь и грамматика** – два тесно связанных и согласованных компонента структуры языка. Их согласованность определяется общностью основных функций и сохранением в компьютерной памяти как языковых единиц, готовых к употреблению, так и грамматических правил, по которым в соответствии с заданием автоматически осуществляется анализ текста. Результаты работы АГАТ-грамматики выложены для общего пользования на портале www.mova.info.

ЛИТЕРАТУРА

1. Гончаренко, В. В. Автоматическая переработка текста / В. В. Гончаренко [и др.]. – Кишинев : Штиинца, 1978. – 93 с.
2. Зубов, А. В. Основы искусственного интеллекта для лингвистов / А. В. Зубов, И. И. Зубова. – М. : Логос, 2007. – 319 с.
3. Зубов, А. В. Информационные технологии в лингвистике / А. В. Зубов, И. И. Зубова. – М. : Academia, 2004. – 205 с.
4. Дарчук, Н. П. Комп'ютерне анотування українського тексту: результати і перспективи / Н. П. Дарчук. – Київ : Освіта України, 2013. – 543 с.

Р. В. Детскина

СТРУКТУРНЫЕ МОДЕЛИ АНГЛОЯЗЫЧНЫХ ИМЕННЫХ СЛОВСОЧЕТАНИЙ

Интерес к теоретическим и практическим исследованиям по проблемам словосочетаний в целом и именных словосочетаний в частности значительно возрос в последнее время. Это объясняется, прежде всего, новыми возможностями изучения именных словосочетаний, связанными с достижениями в области автоматической переработки текста.

Хотя основной лексической единицей языка является слово, но одно лишь слово не может раскрыть всей глубины языка, поэтому оно соединяется с другими словами, конструкциями, образуя словосочетания и предложения, из которых строится текст. Несмотря на то, что словосочетание является одной из основных единиц синтаксиса и изучение словосочетания, наравне с изучением предложения, составляет предмет синтаксиса, до настоящего времени не существует общепринятого определения этой синтаксической единицы, и имеются серьезные расхождения в ее трактовке в отечественной лингвистике и за рубежом [1; 2; 3]. В. В. Виноградов рассматривает словосочетания как грамматические единства, которые возникают из сочетания слов по законам или правилам данного языка, состоят не менее чем из двух полнозначных слов и являются строительным материалом для предложения, выражая в составе предложения единые, хотя и расчлененные значения. Он также говорит, что словосочетания входят в систему коммуникативных средств языка только в составе предложения и через предложение. По его словам, рассматриваемые вне предложения, как строительный материал для него, словосочетания, так же, как и слова, «относятся к области номинативных средств языка, средств обозначения предметов, явлений, процессов и т.п.» [4, с. 231].

В английском языке множество видов словосочетаний. Поэтому так сложно выявить конечное количество структур, чтобы охватить все возможные варианты сочетания слов в словосочетании, что необходимо для составления алгоритма их перевода. Однако, несмотря на многообразие структурных моделей именных словосочетаний, можно выявить их определенные структурные типы, отвечающие двум основным функциям словосочетаний – номинативной и коммуникативной.

Материалом для анализа послужили 400 англоязычных именных словосочетаний из различных художественных произведений таких авторов как О. Уайльд и Дж. Лондон.

Проведенный анализ англоязычных именных словосочетаний показал, что наиболее часто с существительным сочетаются прилагательные, причастия настоящего и прошедшего времени. Существительное может осложняться, образуя словосочетания, состоящие из четырех – пяти слов. Однако возможность построения сложных многословных словосочетаний отнюдь не свидетельствует о безграничности их объема. Усложнение простых словосочетаний ограничено рамками некоммуникативных объединений слов. Следовательно, объем словосочетаний ограничен его грамматической природой, его качественным отличием от предложения.

Результаты проведенного анализа по выявлению структурных моделей проанализированных англоязычных именных словосочетаний представлены в табл. 1.

Наиболее частотной структурой именного словосочетания является прилагательное и существительное, например: *sweet water, a little hand, next effort*. Тем не менее, для образования именных словосочетаний также используются такие структуры, как: причастие настоящего или прошедшего времени и существительное (*enduring bursts, tired faces*); существительное и существительное (*the lawnmower man*); числительное и существительное (*two hundred voices*); притяжательные конструкции (*Joseph's story*), а так же словосочетания, в которых главным словом является прилагательное (*almost opaque*), местоимение (*the mental ones*).

Т а б л и ц а 1

Структурные модели проанализированных англоязычных именных словосочетаний и их переводы на русский язык (фрагмент)

№ п/п	Англоязычное именное словосочетание	Структура именного словосочетания	Перевод
1	<i>Shy grace</i>	A ¹ +N	Робкое изящество
2	<i>Startled eyes</i>	PII+N	Испуганный взгляд
3	<i>Pretended indifference</i>	PII+N	Притворное равнодушие
4	<i>All the spiders</i>	D+N	Все пауки
5	<i>New ropes</i>	A+N	Новые сети
6	<i>Dark spaces</i>	A+N	Темный угол
7	<i>The late post</i>	A+N	Последняя почта
8	<i>Two letters</i>	Num+N	Два письма
9	<i>The hunting maser</i>	PI+N	Охотничий маузер
10	<i>Hard weather</i>	A+N	Плохая погода
11	<i>Several times</i>	D+N	Несколько раз
12	<i>Great a part</i>	A+N	Большая роль

¹ В таблице используются следующие обозначения:

- N – существительное
- A – прилагательное
- D – определяющее слово
- PII – причастие прошедшего времени
- PI – причастие настоящего времени
- Num – числительное
- G – притяжательный падеж
- Pron – местоимение
- Adv – наречие.

13	<i>Women's life</i>	G+N	Жизнь женщины
14	<i>Elaborate answers</i>	A+N	Точнейшие ответы
15	<i>Every possible observation</i>	D+A+N	Любое возможное возражение
16	<i>Long western arm</i>	A+A+N	Длинный западный склон
17	<i>The sad crying</i>	A+N	Грустные завывания
18	<i>Endless corridors</i>	A+N	Бесконечные коридоры
19.	<i>The old lady's nephew</i>	A+G+N	Племянник старой леди
20.	<i>The lowest button</i>	A+N	Нижняя пуговица
21.	<i>All poor people</i>	D+A+N	Все бедные люди
22.	<i>Gradual process</i>	A+N	Постепенно
23.	<i>Quick one</i>	A+Pron	Быстро
24.	<i>Shut eyes</i>	PII+N	Закрытые глаза
25.	<i>Acute mental vision</i>	A+A+N	Острое внутреннее зрение
26.	<i>The pale student</i>	A+N	Бледный студент
27.	<i>Unhallowed arts</i>	PII+N	Порочные науки
28.	<i>Six yards</i>	Num+N	Шесть ярдов
29.	<i>Rare smiles</i>	A+N	Редкая улыбка
30.	<i>Eternal reassurance</i>	A+N	Неизменная уверенность
31.	<i>Four or five times</i>	Num+or+Num+N	Четыре или пять раз
32.	<i>Such circumstances</i>	D+N	Подобные обстоятельства
33.	<i>A safe delivery</i>	A+N	Безопасные роды
34.	<i>The sorrowful task</i>	A+N	Печальная процедура
35.	<i>Considerable eminence</i>	A+N	Довольно высокий холм
36.	<i>The opposite side</i>	A+N	Противоположная сторона
37.	<i>Some abruptness</i>	D+N	Некоторая неожиданность
38.	<i>Pleasant honeyed words</i>	A+PII+N	Приятные сладкие слова
39.	<i>The same people</i>	D+N	Определенные люди
40.	<i>The same place</i>	D+N	Одно место
41.	<i>Written records</i>	PII+N	Письменные данные
42.	<i>Human life</i>	N+N	Жизнь людей
43.	<i>Any standard</i>	D+N	Никакие стандарты
44.	<i>A double life</i>	A+N	Двойная жизнь

Анализ отобранных англоязычных именных словосочетаний показал, что типичным количеством слов для словосочетания является два: *a plane crash, murderer fellows, the Lakers game*. Однако встречаются и очень распространенные словосочетания, состоящие из четырех и более слов: *two bullet-torn or moth-eaten cavalry pennants, some very particular business, a canary-yellow two-piece bathing suit*.

При переводе отобранных англоязычных именных словосочетаний мы отметили, что некоторые словосочетания переводятся на русский язык одним словом, например: *the lawnmower man* – 'газонокосильщик', *a French window* – 'терраса', *the little girl* – 'девочка'. Некоторые словосочетания переводятся как более сложные конструкции, например: *any satisfactory description* – 'описание, которое можно удовлетворить', *the lyre-end sofa* – 'диванчик, похожий на лиру', *two bullet-torn or moth-eaten cavalry pennants* – 'два кавалерийских флажка, продырявленные пулями'.

На основе анализа структуры отобранных англоязычных именных словосочетаний были получены структурные модели, которые сведены в табл. 2.

Т а б л и ц а 2

Синтаксические модели проанализированных англоязычных словосочетаний

№ п/п	Синтаксические модели	Абсолютная частота употреблений	Относительная частота употреблений
1	Двухсоставные базовые модели	322	0.805
	1. A+N	170	0.527
	2. Num+N	32	0.1
	3. N+N	32	0.1
	4. D+N	30	0.093
	5. PII+N	26	0.081
	6. PI+N	16	0.05
	7. G+N	10	0.031
	8. A+Pron	4	0.012
	9. Adv+A	2	0.006
2	Трехсоставные базовые модели	66	0.165
	1. A+A+N	18	0.28
	2. D+A+N	10	0.15
	3. A+PII+N	6	0.09
	4. Adv+A+N	6	0.09
	5. A+N+N	4	0.06
	6. Num+A+N	4	0.06
	7. D+N+N	4	0.06
	8. Adv+Adv+PII	2	0.03
	9. PI+N+N	2	0.03
	10. G+A+N	2	0.03
	11. Num+N+N	2	0.03
	12. A+G+N	2	0.03
	13. A+PI+N	2	0.03
	14. Num+Num+N	2	0.03
3	Четырехсоставные базовые модели	10	0.025
	1. D+Adv+A+N	4	0.4
	2. D+Num+Num+N	2	0.2
	3. A+N+PI+N	2	0.2
	4. A+A+A+N	2	0.2
4	Пятисоставные базовые модели	2	0.005
	1. Num+PII+PII+N+N	2	1

Анализ англоязычных именных словосочетаний показал, что большинство словосочетаний состоят из двух элементов. Меньше всего представлено пятисоставных именных словосочетаний. Промежуточное положение занимают трехсоставные и четырехсоставные словосочетания. Наибольшее разнообразие структур наблюдается у словосочетаний, состоящих из трех элементов. Наглядно полученные результаты можно представить на диаграмме (рис. 1).

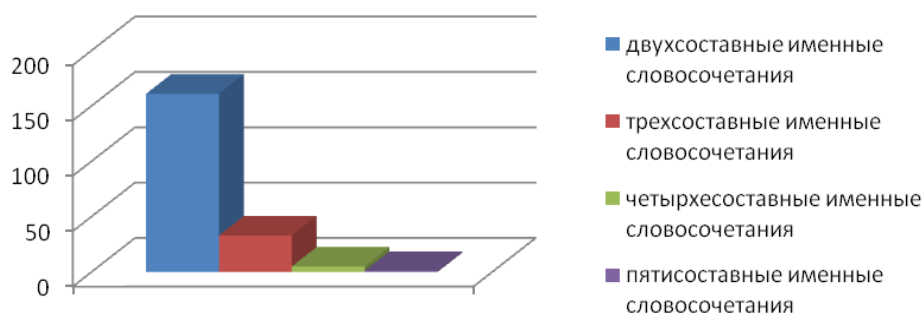


Рис. 1. Диаграмма соотношения синтаксических моделей в проанализированном массиве англоязычных словосочетаний

Среди двухсоставных именных словосочетаний наиболее часто встречаются словосочетания, имеющие структуру A+N. Словосочетания типа N+N, Num+N, D+N, PII+N встречаются реже. Наименее употребительной структурой двухсоставного именного словосочетания является Adv+A. Эти данные можно представить в виде диаграммы (рис. 2).

Двухсоставные именные словосочетания

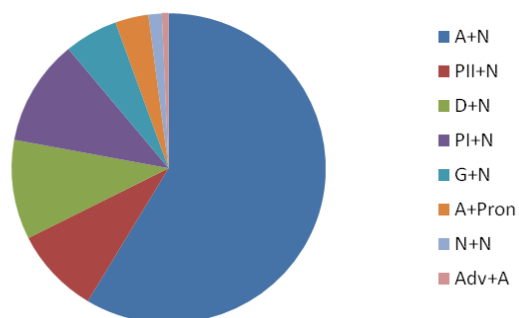


Рис. 2. Диаграмма представленности двухсоставных именных словосочетаний в проанализированном массиве

Среди трехсоставных именных словосочетаний наиболее часто встречается структура A1+A2+N. Промежуточное положение занимают словосочетания с такими структурами, как: D+A+N, A+PII+N, Adv+A+N, Num+A+N, D+N1+N2. Реже всего используются именные словосочетания, которые имеют структуры A+G+N, A+N1+N2, A+PI+N, A+N1+N2, Adv1+Adv2+PII, PI+N1+N2, G+A+N, A+N1+N2, Num+N1+N2. Эти данные можно наглядно представить на следующей диаграмме (рис. 3).

Трехсоставные именные словосочетания

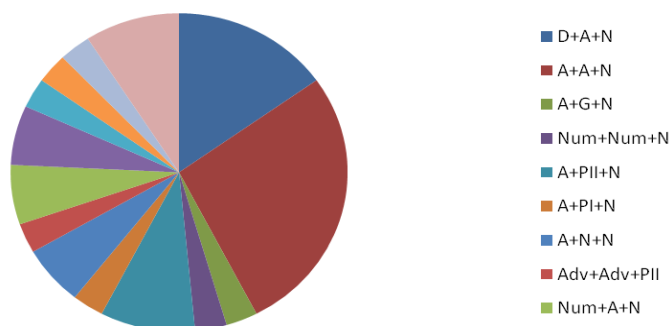


Рис. 3. Диаграмма представленности трехсоставных именных словосочетаний в проанализированном массиве

Среди четырехсоставных именных словосочетаний наиболее часто используются словосочетания, имеющие структуру D+Adv+A+N. Так же четырехсоставные именные словосочетания представлены в виде таких структур, как D+Num1+Num2+N, A+N1+PI+N2, A1+A2+A3+N. Эти данные можно представить на диаграмме (рис. 4).



Рис. 4. Диаграмма представленности четырехсоставных именных словосочетаний в проанализированном массиве

Пятисоставные именные словосочетания представлены в проанализированном массиве единственной структурой Num+PII+PII+N1+N2.

ЛИТЕРАТУРА

1. Долгов, Ю. С. Теория валентности и словосочетание / Ю. С. Долгов // Русский язык. – Минск, 1992. – С. 59–68.
2. Печников, А. Н. Член предложения или словосочетание? / А. Н. Печников // Филологические науки. – 2003. – № 5. – С. 49–57.
3. Смольянинова, Е. Н. Актуальные проблемы синтаксиса словосочетания в русистике / Е. Н. Смольянинова. – Л., 1980. – 60 с.
4. Виноградов, В. В. Из истории изучения русского синтаксиса (от Ломоносова до Потебни и Фортунатова) / В. В. Виноградов. – М., 1958. – 400 с.

Н. А. Копачева

ПРАГМАСТИЛИСТИЧЕСКИЕ ФУНКЦИИ СЛОВОПОРЯДКА В АНГЛИЙСКИХ СЛОЖНЫХ ПРЕДЛОЖЕНИЯХ

В настоящей статье ставится цель проанализировать особенности порядка размещения главного и придаточного предложений в сложном предложении, определить формы и функции порядка следования его частей в медийном дискурсе. Изучение проблемы порядка следования синтаксических сегментов в структуре сложного предложения связано с рассмотрением предложения как единицы коммуникации, как минимального речевого произведения. Основным средством, обеспечивающим возможность построения связных речевых произведений, является именно структура предло-

жения. «Порядок слов, формы самих слов и др. являются не чем иным, как элементами или составными частями структуры предложения» [1, с. 11]. Далее, как утверждает Л. С. Бархударов, «сами предложения употребляются не в отрыве друг от друга, но вступают в определенные связи, образуя группы предложений, именуемые традиционно сложными предложениями» [1, с. 13].

Развивая идею статуса сложного предложения как коммуникативной единицы В. А. Белошапкова подчеркивает, что «функциональная целостность является неотъемлемой и важной чертой сложного предложения» [2, с. 18].

В настоящем исследовании мы рассматриваем сложное предложение как структурное и семантическое единство, обладающее функциональной и коммуникативной целостностью. Структурная и семантическая целостность сложного предложения предполагает определенную организацию его составляющих и специфических для сложного предложения способов такой организации. Такими специфическими способами являются формы порядка частей сложного предложения. Анализ практического материала позволяет нам выделить три формы расположения главного предложения относительно придаточного: препозиция, интерпозиция и постпозиция. Однако для сложного предложения существуют определенные особенности употребления придаточной структуры в той или иной позиции. Словорасположение в сложном предложении имеет характерные особенности. Порядок слов в частях сложного предложения иногда отличается от порядка слов в аналогичных по синтаксическому строению и смыслу простых предложениях, ибо сложное предложение по своей структуре не является сложением простых предложений. Это объясняется тем, что словопорядок оформляет сложное предложение как единое коммуникативное целое, вследствие чего определенное словорасположение одной части требует определенного словорасположения в другой. Кроме того, в некоторых типах сложных предложений наблюдается жестко регламентированный порядок следования частей, другие же типы допускают свободу их расположения, что объясняется степенью гибкости или негибкости их структуры.

Находясь в постпозиции, придаточное предложение относится к отдельной части главного предложения, или в некоторых случаях ко всему главному предложению, выполняя определенную коммуникативную функцию. Анализ исследуемого материала позволяет нам выделить факторы, определяющие место придаточного предложения в структуре сложного. Так, постпозиция придаточного зависит от лексико-грамматической природы антецедента: в большинстве случаев постпозитивное придаточное следует за структурно обязательными членами предложения и занимает контактное положение по отношению своего антецедента. К структурно обязательным элементам предложения мы относим подлежащее и сказуемое, прямое и косвенное дополнение, обстоятельства при некоторых глаголах. Например:

“...I know that my chances of sleep are over.”

“An island tour takes us to the far eastern shore where we learn of Garifuna people.”

Расположение придаточного предложения в препозиции в большинстве случаев зависит от типа придаточного предложения. Анализ практического материала свидетельствует о том, что распределение функций между элемен-

тами главного и препозитивного придаточного в функции обстоятельства времени осуществляется следующим образом: функция придаточного заключается в выражении временного ориентира по отношению к действию главного предложения или обозначения предпосылок и обстоятельств событий, о которых говорится в главном предложении. Функция главной части – это комментарий, оценка, разъяснение событий, о которых идет речь в придаточном предложении. Например:

“When Feix Baumgartener steeped off the small ledge of a space capsule for a record 24’ mile plunge to earth Sunday, it represented a giant promotion leap for Red Bull...”

Следовательно, придаточные предложения в постпозиции описывают действие, сопутствующее главному. Придаточные предложения в препозиции отличаются двухсторонней связью главного и придаточного предложений и «характеризуются отнесенностью ко всему главному предложению и определяют его по характеру и условиям предикативного признака» [3, с. 81]:

“If we are to go on attracting foreign business, we need good flight connection”.

Таким образом, смысловую и логическую самостоятельность придаточного предложения и высокую степень его влияния на содержание главного предложения можно назвать отличительной чертой препозитивных конструкций.

Исследование показывает, что интерпозитивные конструкции отличаются от препозитивных и постпозитивных по структурному признаку и по семантической нагрузке. Как правило, придаточные интерпозитивные конструкции располагаются между структурно обязательными элементами главного предложения и семантически не относятся ни к одному элементу главного предложения. Например:

“The stunts, however big they might be, are widely seen as not landing as well”.

Располагаясь в интерпозиции по отношению к главному предложению, придаточное включается между обязательными элементами главного, не относясь синтаксически ни к одному из них, и выражает дополнительное сообщение. Например:

“Online games maker Zynga, which was reported to have a quarterly loss in April, reports a second-quarter earnings today”.

Как видно из приведенных примеров, в случае интерпозитивного придаточного компоненты главного предложения разъединены или, другими словами, расположены дистантно.

Таким образом, каждый тип порядка следования главного и придаточного обстоятельственного имеет свои особенности. Постпозитивное придаточное, относящееся к какому-либо члену предложения, реализует функцию обстоятельственного расширения и описания. Придаточное предложение в препозиции, соотносясь со всем предложением, выполняет функцию детерминирующего обстоятельства. Придаточное в интерпозиции, не относящееся ни к какому члену предложения, выражает дополнительное сообщение. Анализ порядка расположения придаточного предложения относительно главного показывает, что выбор зависит от целей говорящего, от того, какую информацию он хочет сообщить.

В настоящем исследовании мы проанализировали некоторые функции форм порядка частей сложного предложения. Анализ материала показал, что не все типы придаточных предложений обладают достаточной позиционной гибкостью, чтобы занимать любую из трех позиций по отношению к главному предложению. Такую позиционную гибкость больше проявляют придаточные обстоятельственные. Например, сложноподчиненное предложение:

“It might be adjusted by throwing charges and weighting them until the measure is adjusted to deliver the correct amount of powder” может быть трансформировано в *“Until the measure is adjusted to deliver the correct amount of powder it might be adjusted by throwing charges and weighting them”*.

Такое структурно-синтаксическое изменение формы порядка не меняет общий смысл предложения, но влечет за собой изменение тема-рематического членения предложения и, как результат, его прагматико-стилистической функции.

При порядке следования частей предложения «главное – придаточное» высказывание, как правило, разворачивается от темы к реме, и придаточное предложение несет большую коммуникативную нагрузку, выступая в роли ремы. При синтаксическом построении «придаточное – главное» может измениться соотношение темы и ремы в зависимости от прагматической установки говорящего, с одной стороны, и производимого на слушающего эффекта, с другой. Изменение порядка следования элементов отражает изменение коммуникативного намерения автора, что подтверждает тесную взаимосвязь коммуникативной и стилистической функций порядка следования элементов.

Таким образом, анализ позиционных характеристик частей сложного предложения в медийном дискурсе позволяет нам утверждать, что сложные предложения с придаточными обстоятельственными обладают определенной позиционной гибкостью: изменение порядка следования их частей влечет изменение прагматической, стилистической и коммуникативной функций, что свидетельствует о значимости порядка языковых элементов в системе языковых средств английского языка.

ЛИТЕРАТУРА

1. Бархударов, Л. С. Структура простого предложения современного английского языка / Л. С. Бархударов. – М. : Высш. шк., 1966. – 200 с.
2. Белошапкина, В. А. Сложное предложение в современном русском языке / В. А. Белошапкина. – М. : Просвещение, 1967. – 357 с.
3. Беданова, С. К. Позиция придаточного предложения как синтаксическая проблема / С. К. Беданова. – М. : Просвещение, 1985. – 189 с.

С. Е. Кунцевич

ОСНОВНЫЕ НАПРАВЛЕНИЯ ИССЛЕДОВАНИЯ ПРОБЛЕМЫ ЖАНРА В ЛИНГВИСТИКЕ

Целью предлагаемой статьи является краткий обзор существующих в лингвистической науке направлений исследования жанров и выделение наиболее перспективных из них. Для этого в первую очередь необходимо

пояснить базовые термины рассматриваемого проблемного поля – *жанр* (речевой / риторический / дискурсивный / литературный / литературно-художественный / виртуальный), а также установить соотношение понятий *жанр текста* и *тип текста*. Данное проблемное поле выступает объектом разноплановых исследований, включая работы, проводимые в рамках прикладной лингвистики.

Понятие речевого жанра, зародившееся в поэтике Аристотеля, развивалось в трудах многих ученых, использовавших термины *жанр речи/общения, языковой жанр*. Значительный вклад в разработку данного понятия внесла русскоязычная исследовательская традиция. С точки зрения автора лингвофилософской концепции жанра М. М. Бахтина [1], речевые жанры соответствуют типичным ситуациям общения, представляя собой типы высказываний, являющиеся относительно устойчивыми в тематическом, композиционном и стилистическом планах. Под высказыванием М. М. Бахтин понимал минимальную, отличающуюся смысловой завершенностью единицу общения, о границах которой сигнализирует смена коммуникантов. Речевой жанр передает однотипное предметно-смысловое содержание и предстает в высказывании как факт социального взаимодействия, как соотношение смысловых позиций участников общения. Признаками речевого жанра выступают: сфера и конкретная ситуация общения, предполагаемый адресат, адресант, авторский замысел, целостность, сходные композиционно-стилистические особенности, обусловленные смысловым содержанием высказывания и субъективным отношением адресанта к нему. Перечисленные признаки свойственны и целому тексту, поэтому к жанру, наряду с высказыванием/ первичным речевым жанром, можно отнести типовую разновидность текста/ сложное высказывание/ вторичный жанр как устойчивую модель построения речевого целого. Определив понятия первичного и вторичного речевых жанров, М. М. Бахтин обозначил ряд связанных с ними нерешенных наукой проблем. Дискуссионной в лингвистике до сих пор остается **проблема уяснения специфики** речевых жанров в их взаимосвязи с условиями общения, а также **проблема разграничения** речевых и языковых жанров.

В настоящее время, следуя той или иной научной традиции, ученые оперируют различными трактовками термина *жанр*. В самом общем виде жанр понимается либо как разновидность функционального стиля, определяемая особой композицией, характером и эмоциональной окраской представляемой информации – языковой жанр, либо как закрепленная традицией речевая форма реализации функции практического назначения сообщаемого содержания – жанр общения. Особо подчеркнем, что именно **концепция М. М. Бахтина** оказалась надежной теоретической базой для множества актуальных исследований, выполненных в рамках прагматики, риторики, стилистики, компьютерной лингвистики, лингвистики текста, теории дискурса, жанроведения (генристики/ генологии).

Так, в работах по жанроведению основной интерес проявляется к двум объектам изучения, восходящим к концепции М. М. Бахтина: исследование жанров с использованием языкового материала, отражающего специфику бытовой коммуникации; исследование жанров, обеспечивающих другие

разновидности коммуникации. *Коммуникативно-прагматическое жанроведение* как новое направление современной лингвистики дает возможность глубже описать прагматический план профессионально-ориентированного и статусно-ориентированного дискурсов, их жанровую организацию, оптимальную приспособленность к выполнению функциональных задач профессионального или статусного общения [2]. Существенная роль при выделении жанров отводится в рамках названного направления таким базисным категориям коммуникативной прагматики, как характер ситуации общения и адресно-направленная интенциональность, которая соотносится с функциями информирования, убеждения и воздействия.

При рассмотрении жанра с позиций риторики на первый план выдвигается анализ жанра на стадии его формирования с учетом ситуации, в которой он реализуется. Под ситуацией здесь подразумевается не только конкретная обстановка (место, время, тип общения, адресат), но и событие или повод, требующий от адресанта обращения к определенному типу речи. Риторический жанр трактуют как результат речевой деятельности отправителя текста, объединяющий в себе основные устойчивые риторические элементы, специфицированные риторическим каноном [3, с. 75].

О. Б. Сиротина усматривает различие между речевым и риторическим жанрами в том, что первый предполагает отсутствие сознательного планирования и построения речи, обслуживая типичные ситуации неофициального неподготовленного общения [4]. Между тем второй соотносится с официальным, специально подготовленным и продуманным высказыванием, с целенаправленным выбором конкретных языковых средств для его выражения. В ряде случаев один и тот же жанр, в зависимости от степени осознанности речевого поведения адресанта, выступает либо как риторический, либо как нериторический. Владению риторическими жанрами как неотъемлемой частью коммуникативной и профессиональной компетенции, а также способами оформления социально значимого взаимодействия людей в речевом общении можно и необходимо обучать [5].

Понятие дискурсивный жанр фокусирует внимание исследователя на поведении языковой личности в процессе коммуникации и определяет характер его мышления [6]. Дискурсивная теория нацелена на изучение функционирования языка в социальном контексте с учетом всех аспектов деятельности человека. Дискурсивный жанр оказывается своеобразной реакцией индивида на конкретную ситуацию общения и рассматривается как средство формализации социального взаимодействия, как типизированная коммуникативная/ речевая практика членов определенного языкового сообщества.

Принципиальным представляется вывод лингвистов о том, что *прототипические модели* речевых/ дискурсивных *жанров* неким образом присутствуют в сознании человека, в соответствии с ними он строит, воспринимает и оценивает конкретное речевое произведение. Анализ теоретической литературы показывает, что ученые выделяют различные, зачастую не совпадающие жанрообразующие признаки, придерживаясь многофакторного подхода к описанию жанров.

Термин *дискурсивный жанр* более других соответствует **когнитивно-дискурсивному подходу** к исследованию текстов. К ведущим признакам дискурсивного жанра относят: социальность, адресность, соотнесенность с определенной профессиональной деятельностью, целеустановку, тематическую и композиционно-смысловую устойчивость. Взятые в совокупности, перечисленные признаки составляют методологическую основу для создания типологической жанровой модели, поскольку с их помощью можно упорядочить и логически объяснить эмпирические данные, полученные относительно самых разных дискурсивных жанров.

Целям анализа дискурса как системы, включая его жанровую дифференциацию, служит такое направление когнитивно-дискурсивных исследований, как **когнитивная прагмалингвистика**. По убеждению лидера названной школы И. С. Шевченко, дискурс следует изучать как систему-гештальт, представленную совокупностью кодового употребления (языковой аспект), передачей/использованием идей и убеждений (когнитивный аспект), взаимодействием коммуникантов в определенных социально-культурных ситуациях и контекстах (социально-прагматический аспект) [7].

На роль эксплицитного критерия разграничения жанров в современных исследованиях по дискурс-анализу и генеристике с полным правом претендует понятие *жанровая лингвопрагматическая доминанта* (сходные термины – жанровая доминанта и текстовая доминанта – упоминали в своих работах Н. М. Разинкина и Л. Г. Бабенко соответственно), заслуживающее более глубокой теоретической и практической разработки. Это понятие было принято за операционную единицу анализа в кандидатской диссертации Е. В. Хоменко, посвященной исследованию жанрообразующего потенциала аргументативно-оценочных средств [8]. Лингвопрагматическая доминанта понимается Е. В. Хоменко как базовый компонент структурно-семантической организации текста, который обеспечивает его прагматическое соответствие ситуации общения. Такой подход дает возможность эмпирическим путем выявить признаки, необходимые и достаточные для жанровой дифференциации англоязычного научного текста по энергетике, и может быть экстраполирован на иные тематические выборки текстов и другие жанры.

Во избежание противоречий в употреблении неоднозначного термина *жанр*, некоторые лингвисты предлагают оставить его литературоведам (как синонимичный понятиям *литературный* / *литературно-художественный жанр*) и заменить термином *тип текста*. Теория типов текста, нацеленная на изучение функциональных и структурных особенностей текстовых экземпляров различного тематического содержания, активно развивается в постсоветской лингвистике, причем, действительно, понятие типа текста чаще используется при описании нехудожественных произведений.

По мнению многих исследователей, в коммуникативном пространстве тип текста соотносится с жанром текста, однако обращает на себя внимание неоднозначность трактовки обозначенных понятий. Под типом текста понимается совокупность текстовых произведений, используемых для достижения необходимых коммуникативных целей в типовых ситуациях общения. В свою очередь, жанр выступает как совокупность текстов, выделенных на основе идентичности структурной организации, границ вариатив-

ности и позволяющих достичь необходимых коммуникативных целей в заданных ситуациях общения. По мнению Т. И. Синдеевой, жанр текста следует рассматривать как инвариант по отношению к формам его существования – отдельным текстам [9]. В. А. Тырыгина отмечает, что понятия текстотипа и жанра наполняются неодинаковыми смыслами и не являются тождественными, поскольку тип текста как категория не включает в себя некоторых характерных признаков, которые свойственны жанру, делая его категорией более конкретной [10]. В. Е. Чернявская акцентирует тот факт, что **перечень критериев**, релевантных для классификации типов текста, остается малоизученной проблемой [11].

В традиционном представлении текст выступает как завершенная смысловая последовательность языковых и/ или неязыковых знаков, обладающая определенной внутренней организацией, имеющая формальное начало и формальный конец. Однако нельзя не признать, что в наши дни в коммуникации все чаще используются незавершенные, не имеющие четко обозначенных границ тексты. Примером тому могут служить гипертексты, которые беспрерывно создаются в глобальном Интернет-пространстве; функционирующие в Сети художественные произведения, параллельно пишущиеся различными авторами, находящимися вдалеке друг от друга; Интернет-чаты без видимых признаков завершенности сообщений.

Каждая эпоха приходит в жизнь социума с новыми вызовами, которые необходимо исследовать с тем, чтобы оперативно реагировать на них. Все возрастающая роль сети Интернет заставляет обратить пристальное внимание лингвистической науки на особенности использующихся в ней текстов. Благодаря развитию Сети, произошло резкое увеличение информационного поля, изменились способы подачи информации. Традиционные бумажные носители неуклонно уступают место электронным формам, основанным на интерактивности как тесном взаимодействии пользователя с компьютером. В технократическую эпоху человек все чаще черпает информацию, играет в литературные и иные игры, читает книги посредством Интернета. Книга в ее бумажном варианте, оставаясь непревзойденным образцом человеческого гения, уступает место новым жанрам литературного и псевдолитературного творчества.

Нелинейный гипертекст – это метод объединения электронных документов; основанная на гиперссылках форма организации текста, дающая читателю возможность по своему усмотрению осуществлять его линейную структуризацию. Возможность нелинейного чтения и/ или письма обеспечивается сегментацией и иерархической ассоциативной атрибуцией фрагментов гипертекста. Разветвляясь, он позволяет пользователю планировать собственный текстовый маршрут. Еще одной принципиально важной характеристикой гипертекста как интерактивной системы, опосредованной компьютерной средой, является мультимедийность, то есть одновременное использование различных медиа – анимированная компьютерная графика, видео, музыка.

Исследователи предлагают различные классификации гипертекстов. М. Визель специфицирует их по трем параметрам: *содержание* (художественные и нехудожественные гипертексты, такие как энциклопедии,

справочники, телефонные книги), *место функционирования* (например, изолированные на CD либо сетевые гиперроманы, постоянно обновляемые пользователями), *способ прочтения* (чтение, чтение с комментариями, чтение-письмо) [12]. А. Андреев выделяет три основных *способа организации* гиперлитературы: «елка с игрушками» (текст с выделенными ссылками), «сад расходящихся тропок» (книга с набором глав и своего рода путеводителем, позволяющим знакомиться с главами в любом порядке, создавая собственный читательский маршрут), «макrame» (это может быть роман, созданный с помощью специальной компьютерной программы) [13].

Из вышесказанного вытекает, что ***гипертекст как актуальный объект лингвистического анализа*** разрушает традиционное представление о тексте как о линейно упорядоченном, обладающем внутренней тематической связностью единстве. В данных обстоятельствах с новой силой встает вопрос о критериях текстуальности, в ином ракурсе высвечиваются проблемы исчезновения автора и степени значимости авторской интенции.

За последние десятилетия в Интернет-пространстве накопилось великое множество оригинальных литературообразных текстов. Одновременно стали появляться и синонимичные термины, характеризующие данный процесс и отражающие различные подходы к нему: *виртуальная/ сетевая литература/ сетература/ кибература/ цифровая/ гипертекстовая художественная литература*.

Гипертекст и сетература существуют в нелинейной среде, обитают в виртуальной реальности и противопоставлены записанным на бумаге или оцифрованным линейным текстам. Понятие сетературы выдвигает в центр внимания идею об Интернете как средстве создания интерактивной творческой среды. Эта ***новая форма литературы*** кардинально меняет отношение читателя и автора, превращая их в полноправных соавторов.

В широком смысле сетература включает весь объем имеющихся в Интернет-пространстве литературных текстов, независимо от формы их представления (текст, гипертекст, гипермедиа) и способа написания (индивидуальное или коллективное авторство, создание текста с помощью компьютерной программы) [14]. Активно разрабатываются пользователями новые литературные формы, жанры и типы текстов – гиперпоэзия, гиперроман, гиперрассказ, гипердрама и т.п. На повестке дня их инвентаризация, выявление основных жанровых особенностей сетевого креатива, описание преимуществ и недостатков цифровой литературы, установление соотносительности виртуальных жанров с традиционными литературными жанрами.

Цельных трудов, посвященных детальному исследованию гипертекста и выявлению специфики виртуальных жанров (не только литературно-художественных) в мире не так много; в нашей стране эта проблема частично затрагивается лишь в единицах научных работ. К примеру, в диссертационном исследовании К. А. Беловой описан Интернет-дискурс Беларуси с акцентом на его социолингвистическом аспекте [15]. Исходной классификационной единицей в этой работе оказывается дискурсивный жанр, из которого выводится понятие субжанра и осуществляется разноуровневое описание *коммуникативно значимых субжанров* белорусского Интернет-пространства.

В заключение подчеркнем, что виртуальные жанры стремительно развиваются и проникают во все сферы жизни человека, резко между собой различаясь. Виртуальная литература представляет собой скорее сетевое, нежели литературно-художественное явление и в большинстве случаев воспринимается как одна из коммуникационных возможностей Интернета. Однако, несмотря на это, сетература выступает неотъемлемой частью современного литературного процесса. Имея свои положительные и отрицательные стороны, виртуальные жанры связаны с традиционными жанрами литературы, при этом явно наблюдается влияние одного феномена на другой.

Из вышеизложенного со всей очевидностью следует вывод о том, что проблема жанра в лингвистике не исчерпала своей актуальности, в перспективе она может быть разрешена с позиций современной лингвистической науки, тяготеющей к междисциплинарности.

ЛИТЕРАТУРА

1. *Бахтин, М. М.* Проблема речевых жанров / М. М. Бахтин // Эстетика словесного творчества: сб. избр. тр. – М., 1979. – С. 237–280.
2. *Гусейнова, И. А.* Коммуникативно-прагматические основания жанровой системы в маркетинговом дискурсе : автореф. дис. ... д-ра филол. наук, 10.02.19 / И. А. Гусейнова // ГОУ «Московский гос. лингв. ун-т». – М., 2010. – 41 с.
3. *Михайлова, Е. В.* Интертекстуальность в научном дискурсе (на материале статей) : дис. ... канд. филол. наук : 10.02.19 [Электронный ресурс] / Е. В. Михайлова. – Режим доступа : // www.diss.rsl.ru/diss.aspx?orig=102/0005020_5081.pdf. – Дата доступа : 10.07.2007.
4. *Сиротинина, О. Б.* Некоторые размышления по поводу терминов «речевой жанр» и «риторический жанр» / О. Б. Сиротинина // Жанры речи : сб. науч. ст. – Саратов, 1999. – С. 26–31.
5. *Седов, К. Ф.* Жанр и коммуникативная компетенция / К. Ф. Седов // Хорошая речь / под ред. М. А. Кормилицыной [и др.]. – Саратов, 2001. – С. 107–118.
6. *Плеханова, Т. Ф.* Дискурсивные жанры: проблема классификации и методологии исследования / Т. Ф. Плеханова // Постклассическая стилистика и дискурсивный анализ текста : материалы межкаф. кругл. стола, Минск, 14 дек. 2006 г. / Мин. гос. лингвист. ун-т, каф. стилист. англ. языка; редкол. : С. Е. Кунцевич (отв. ред.) [и др.]. – Минск, 2008. – С. 55–60.
7. *Шевченко, И. С.* Когнитивно-коммуникативная специфика электорального дискурса (на материале президентских кампаний США) / И. С. Шевченко // Коммуникация в поликодовом пространстве: языковые, культурологические и дидактические аспекты (КПП'11) : тез. докл. межд. науч.-практич. конф., 11–13 мая 2011 г. – СПб. : Изд-во Санкт-Петербургского государственного политехнического университета. – С. 29–31.
8. *Хоменко, Е. В.* Жанрообразующий потенциал аргументативно-оценочных средств (на материале англоязычных научных текстов по энергетике) : дис. ... канд. филол. наук : 10.02.04 / Е. В. Хоменко; УО «Минск. гос. лингв. ун-т». – Минск, 2014. – 324 с.
9. *Синдеева, Т. И.* Речевой жанр «Газетная рецензия» и его лингво-текстовые характеристики (на материале английского языка) : автореф. дис. ... канд. филол. наук : 10.02.04 / Т. И. Синдеева; Моск. пед. ин-т ин. языков им. М. Тореза – М., 1984. – 25 с.
10. *Тырыгина, В. А.* Проблема жанра в массово-информационном дискурсе : автореф. дисс. ... д-ра филол. наук, 10.02.04 / В. А. Тырыгина; Моск. гос. пед. ун-т. – Москва, 2008. – 46 с.

11. Чернявская, В. Е. Лингвистика текста. Лингвистика дискурса : учеб. пособ. / В. Е. Чернявская. – М. : Флинта : Наука, 2013. – 208 с.
12. Визель, М. Электронная литература [Электронный ресурс] / М. Визель // «НЛО». – Режим доступа : <http://magazines.russ.ru/nlo/2011/110/vi36-pr.htm>. – Дата доступа : 08.12.2014.
13. Андреев, А. Литература без бумаги [Электронный ресурс] / А. Андреев // Атмосфера – Авторская телекоммуникационная мультимедийная образовательная среда. – Режим доступа : <http://www.autor-edu.ru/ru/node/16>. – Дата доступа : 17.01.2017.
14. Алексрома, А. Сетера и Литера [Электронный ресурс] / А. Алексрома // Сетевая словесность. – Режим доступа : <http://www.litera.ru/slova/teoriya/cet-lit.htm>. – Дата доступа : 02.02.2015.
15. Белова, К. А. Интернет-дискурс Беларуси в социолингвистическом аспекте : дис. ... канд. филол. наук : 10.02.19 / К. А. Белова; УО «Минск. гос. лингв. ун-т». – Минск, 2015. – 176 с.

Г. В. Лосик

АНТРОПОЛОГИЧЕСКАЯ ИНФОРМАЦИЯ О ВАРИАТИВНОСТИ СООБЩЕНИЯ

На стыке нескольких наук часто рождались ответы на вопросы, которые специалистам одной науки не удавалось решить, а их поиск решения превращался в путь «толочь воду в ступе». В одной науке, в информатике статьями А. Я. Фридланда [1], К. К. Колина [2] показана критическая ситуация в дальнейшем изучении «голой» информации, канала ее передачи, исчисления объемов данных без исчисления «содержания» информационного сообщения, ставится под сомнение корректность понятий объективного и субъективного смысла в сообщении. В другой науке, психологии, статьями В. М. Аллахвердова [3] очерчена критическая ситуация в дальнейшем изучении психологии сознания и бессознательного, ставится вопрос причины возникновения сознания в филогенезе, роли его как когнитивного и адаптационного инструмента в жизни человека. Для выхода из критической ситуации в одной из вышеуказанных статей предложена гипотеза о двух компонентах информации и дается новое определение ее понятию [1, с. 77]. Суть гипотезы в предположении, что информационное сообщение содержит две компоненты, в информации имеются две составляющие: *содержание* и *данные*, из которых вторая объявляется новой, не классической. Смысл сообщению придает сознание человека-индивида. Сознание является генератором смысла создаваемого индивидом сообщения и детектором смысла получаемого им сообщения. В контексте данной гипотезы становится актуальным вопрос, каким же образом в мозге кодируется смысл. Чтобы ответить на этот вопрос, в данной статье с позиции теории кодирования рассматривается явление изоморфизма в анатомическом строении сенсомоторной системы у человека как вида.

В большинстве случаев сообщение передаётся от человека к человеку в дискретном коде (словами) и приёмнику достаточно знать правило декодирования сообщения, т.е. *алгоритм*. Человеку достаточно знать информацию о языке сообщения, об алгоритме декодирования и не обязательно быть

физически тождественным передатчику [4]. Однако мы считаем, возможен частный случай, случай человека, когда физическое строение передатчика и приемника как носителей информации – тождественно. Именно в этом случае возникает дополнительная информация о принимаемом сообщении.

Суть появления дополнительной информации в следующем. Если приёмник сообщения имеет сходное строение с передатчиком сообщения, то приёмник имеет возможность узнать дополнительную информацию о вариативности сообщения. Именно эту информацию о вариативности он может узнать не от источника сообщения, а без него, сам, из своего физического строения. Строго говоря, дополнительная информация возникает, не о самом принимаемом сообщении, а лишь о законах его вариаций при многократной отправке его передатчиком. Поначалу в приемнике в его исходном состоянии этой информации о вариативности в готовом виде нет. Для возникновения дополнительной информации нужен специальный эксперимент, в самом уже приемнике, в виде специальной последовательности искажения того сообщения, которое ранее принято как эталон. Почему такое возможно? Потому, что варианты искажения в приемнике в случае тождественности всецело повторяют варианты физической изменчивости передатчика.

Поэтому приемник, приняв поначалу эталон, может указанным экспериментом не воспользоваться, а может и воспользоваться. В первом из таких случаев приемник набирает статистику об эталоне распознаваемого сообщения и о зоне его вариативности, но набирает ее, анализируя лишь поступающие от передатчика сообщения. Во втором случае от передатчика приемник получает информацию только об эталоне сообщения. Зоны его вариативности он отказывается узнавать. В приемнике формируется сенсорный эталон принимаемого сообщения. Это и есть начальная первая фаза эксперимента по вскрытию дополнительной информации. Далее на второй фазе приемник выучивается сам копировать то сообщение, эталон которого запомнил. Для этого у приемника должна существовать система подражания, копирования сообщений, сходных с теми которые передавались передатчиком и которые приемник выучился распознавать. Далее на третьей фазе приемник производит синтез (уже на своем сигнале-дубле), на самом себе набор вариаций исходного эталона. Приемник синтезирует эти вариации один раз, он поочередно синтезирует те векторы вариаций эталона, физически потенциально возможные у приемника. А значит и у передатчика. Эти вариации, заметим, совершаются уже в моторной, а не сенсорной системе координат запоминания эталона. Этот эксперимент по симуляции неточностей совершается один раз. Синтезированные в моторной системе неточные варианты эталона приемник тут же принимает своей сенсорной системой и обогащает сенсорные эталоны моторной дополнительной информацией. Получается, что в итоге такого эксперимента в сенсорный эталон приемник сам добавляет в векторном коде информацию о его потенциальной вариативности. Дополнительная информация о передатчике появляется в приемнике, как бы, не из передатчика.

Проиллюстрируем важность топологического сходства передатчика и приемника как источника дополнительной информации на примере изоморфизма строения у человека слуховой и артикуляционной систем.

Рассмотрим полезность топологического сходства строения передатчика и приемника при кодировании информации на примере усвоения ребенком устной речи [5; 6; 7]. Исследования по детской речи [8, с. 83–84] показали, что фонетические эталоны слогов устной речи (ба, ва, гу, бо) ребенок осваивает в четыре этапа, формируя эталон звучания слога, дублируя его запоминание в двух местах, в двух признаковых пространствах. На первом этапе ребенок получает информацию о звучании слова, воспринимая его акустический сигнал, произнесенный взрослым. Ребенок узнает информацию о звучании слова, но информацию о допустимых неточностях и трансформациях звучания слова узнать не может. Моторная программа артикуляции слова, хранящаяся в памяти взрослого, недоступна ребенку для воздействий на нее. Поэтому, воспринимая речь взрослых, ребенок лишен возможности услышать, например, два варианта произнесения одного и того же слова со строгим отличием программы его артикуляции. Однако устную речь ребенок может не только воспринимать, но и воспроизводить. Для этого у него имеется артикуляторный аппарат, антропоморфный по строению у ребенка и взрослого.

На этом первом этапе в возрасте от 8 до 11 месяцев ребенок не владеет навыками артикуляции целого слова; не умеет осуществлять нужные артикуляторные движения для фонации слова, хотя лепетная речь у него уже развита. Однако в этом возрасте ребенок уже умеет узнавать простые слова на слух. На втором этапе, в последующие сроки ребенок учится их произносить, повторять. У него постепенно формируется моторная программа произношения слов. При этом структура программы одной и той же речевой единицы у ребенка и у взрослого оказывается одинаковой вследствие анатомо-физиологической изоморфности строения их артикуляторного аппарата у первого и второго. И поэтому ребенок, чтобы узнать артикуляторные неточности произношения взрослого, может их моделировать, воспроизводя неточности в собственной программе той же речевой единицы. Это становится следующим, третьим этапом овладения словом [8, с. 89–90]. На этом этапе ребенок совершает плановые элементарные отклонения поочередно в разных звеньях указанной собственной программы. При этом он всякий раз воспринимает на слух одну за другой пару реализаций слова, произнесенных им самим. На четвертом этапе, «вычитая» первую реализацию из второй, ребенок получает и запечатляет в сенсорику сведения о свойственных программе трансформациях, происходящих в его моторике.

Как показало проведенное исследование [8, с. 93–98], итерации лепетной речи ребенка служат для него тем специфическим источником, из которого он приобретает информацию о закономерностях модификации речевых сигналов взрослым. Лепетные итерации типа «ба-ба-ба», «ва-ва-ва» являются акустическими сигналами, которыми ребенок имитирует в своей речевоспроизводящей системе возможные неточности речепроизводства у взрослого. Другими словами, итерации лепета ребенка являются своеобразным видом перцептивных действий, которые слуховая система человека использует при изучении нестабильности работы рече-двигательной системы.

Важно отметить, что для предлагаемого варианта перцептивных действий субъект восприятия должен обязательно иметь в придачу систему син-

теза тех сигналов, восприятие которых он производит. Вероятно, по принципу, описанному для слухоречевого восприятия, формируется образное восприятие у человека мимики и пантомимики окружающих людей, коммуникативных и выразительных жестов другого человека. Всякое подражание, которое совершает в действиях один человек (ученик) по отношению к другому (учителю), согласно предлагаемой модели, обязательно сопровождается формированием у ученика умения не только подражать явлению, но и моделировать его неточности, чтобы константно в будущем его воспринимать.

Заметим, для порядка, что указанные условия необходимы только в том случае, когда в образ предмета поступает информация о нестабильности его формы. Минимальный же образ предмета может формироваться и без этой информации. В последнем случае восприятие предмета тоже будет возможным, но оно будет менее надежным, ибо будет нарушаться в случае, когда на вход анализатора поступает искаженный по форме предмет.

Мы видим, что благодаря изоморфизму строения артикуляторной системы передатчика информации (взрослого) и приемника (ребенка) – последний может не принимать от первого информацию о допустимых вариациях эталона. Ребенок узнает этот последний компонент нужной информации об эталоне из своей собственной артикуляционной системы.

ЛИТЕРАТУРА

1. Фридланд, А. Я. О сущности информации: два подхода / А. Я. Фридланд. – Информационные технологии. – 2008. – № 5. – С. 75–84.
2. Колин, К. К. Критика некоторых методологических подходов в информатике и информационное образование / К. К. Колин, Е. В. Трошин // Открытое образование. – 2005. – № 2. – С. 81–89.
3. Аллахвердов, В. М. Сознание в логике познания / В. М. Аллахвердов // Матер. V Межд. конф. по когнитивной науке; Калининград 18–24 июня 2012 г. – С. 216.
4. Шеннон, К. Работы по теории информации и кибернетике : пер. с англ. / К. Шеннон – М. : Иностранная литература, 1963. – 830 с.
5. Бельтюков, В. И. Взаимодействие анализаторов в процессе восприятия и усвоения устной речи / В. И. Бельтюков. – М., 1977.
6. Бельтюков, В. И. Программа овладения детьми произношением звуков речи / В. И. Бельтюков // Вопросы психологии. – 1979. – № 4. – С. 66–78.
7. Бельтюков, В. И. Об усвоении ребенком звуковой (фонемной) системы языка / В. И. Бельтюков, А. Я. Салахова // Вопросы психологии. – 1975. – № 4. – С. 71–80.
8. Лосик, Г. В. Перцептивные действия в восприятии речи / Г. В. Лосик. – Минск : Ин-т техн. кибернетики НАН Беларуси, 2000. – 168 с.

П. А. Маракулина

ПОНЯТИЕ РЕЧЕВОГО КОНФЛИКТА

В словаре психологической терминологии понятие «конфликт» определено как «трудно разрешимое противоречие, связанное с острыми эмоциональными переживаниями» [1, с. 161]. А. Я. Анцупов и А. И. Шипилов предлагают следующее определение конфликта: «под конфликтом понимается

наиболее острый способ разрешения значимых противоречий, возникающий в процессе взаимодействия, заключающийся в противодействии субъектов конфликта и обычно сопровождающийся негативными эмоциями» [2, с. 6]. Н. Ф. Вишнякова определяет конфликт как «возникновение трудноразрешимых противоречий, столкновение противоположных интересов на почве соперничества, отсутствие взаимопонимания по различным вопросам, связанным с острыми эмоциональными переживаниями» [3, с. 7]. Н. В. Гришина отмечает, что «конфликт выступает как биполярное явление – противостояние двух начал, проявляющее себя в активности сторон, направленной на преодоление противоречия, причем стороны конфликта представлены активным субъектом (субъектами)» [4, с. 17]. Таким образом, многие авторы сходятся во мнении относительно лежащего в основе конфликта столкновения некоторых объектов и процессов. Любой конфликт, независимо от его характера, конкретного содержания и вида, обязательно содержит в себе момент противостояния, противоборства.

Для предотвращения отрицательных последствий и использования конфликта в конструктивных целях необходимо установить его истинные внутренние и внешние причины, понять и спрогнозировать ход его дальнейшего развития. Н. Ф. Вишнякова считает, что в основе любого конфликта лежит ситуация, включающая либо противоположные позиции сторон по какому-то вопросу, либо противоположные цели или средства их достижения в данных обстоятельствах, либо несовпадение интересов, желаний, влечений оппонентов и т.п. Конфликтная ситуация, таким образом, обязательно включает в себя объекты и субъекты конфликта. Однако чтобы конфликт начал развиваться, необходим инцидент, когда одна из сторон начинает действовать, ущемляя интересы другой. Инцидент может возникнуть, как по инициативе субъектов конфликта (оппонентов), так и независимо от их воли и желания – вследствие объективных обстоятельств или случайности [3, с. 8]. Поэтому можно выделить предмет конфликта – противоречие, которое противоборствующие стороны пытаются разрешить; объект конфликта – проблемная ситуация, возникающая на основе пересекающихся интересов оппонентов, которые желают контролировать ее или управлять ею; субъект конфликта – конфликтная личность или проблемы, способные спровоцировать конфликт; оппоненты – участники конфликта. Ими в конфликтной ситуации могут выступать отдельные лица, группы людей, организации. Субъект конфликта и оппоненты в одной и той же конфликтной ситуации связаны определенными отношениями, при этом каждый из оппонентов часто претендует на единоличное манипулирование субъектом. Манипулирование субъектом – это скрытые или явные действия конфликтующей личности, направленные на субъект отношений с целью достижения определенного результата, значимого для этой личности. При этом выделяют деструктивное манипулирование как коммуникативное средство воздействия оппонента, направленное на достижение своих целей в ущерб другой конфликтующей личности, и конструктивное манипулирование как коммуникативную (вербальную или невербальную) систему воздействий и взаимодействий, направленную на достижение общих целей без ущерба интересов других людей. Следовательно, отношения между субъектом и оппонентами зависят,

в первую очередь, от степени предрасположенности субъекта конфликта к манипулированию. Но, чтобы конфликт произошел, со стороны оппонента должны последовать направленные на субъект действия. Эти действия называют инцидентом. Инцидент – это ситуативная провокация конфликтной ситуации, возникшая в результате стечения обстоятельств и явившаяся поводом для конфликта [3, с. 9–10].

Один из исследователей в области конфликтологии В. П. Шейнов приводит три формулы конфликта [5, с. 98]. Первая формула отражает зависимость конфликта (КФ) от конфликтогенов (КФГ) – слов, действий (или отсутствия действий), которые могут привести к конфликту. Механизм развития конфликта по первой формуле основывается на отрицательном восприятии и негативной реакции личности, против которой применен конфликтоген. Более конкретно первую формулу конфликта схематично можно выразить следующим образом: $КФГ_1 \leadsto КФГ_2, КФГ_3 \leadsto \dots \leadsto КФ$, где $КФГ_1$ – первый конфликтоген; $КФГ_2$ – второй конфликтоген, ответный на первый; $КФГ_3$ – третий конфликтоген, ответный на второй, и т.д. При этом важно иметь в виду, что каждый ответный конфликтоген является более сильным, чем тот, на который он отвечает [5, с. 98–100]. Вторая формула отражает зависимость конфликта от конфликтной ситуации (КС) и инцидента (И) и выражается следующим образом: $КС + И = КФ$. Третья формула отражает зависимость конфликта от нескольких конфликтных ситуаций. Ее можно выразить следующим образом: $КС_1 + КС_2 + \dots + КС_n = КФ$. Таким образом, сумма двух и более конфликтных ситуаций приводит к конфликту.

Всякий реальный конфликт представляет собой не единовременный акт, а процесс, нередко весьма длительный, состоящий из определенных этапов. Н. Ф. Вишнякова выделяет следующие этапы в развитии конфликта: постепенное нарастание напряженности в среде участников конфликта; увеличение количества проблемных ситуаций и углубление первичной проблемной ситуации; повышение конфликтной активности участников, изменение характера конфликта в сторону его ужесточения, вовлечение в конфликт новых лиц; нарастание эмоциональной напряженности, которая сопровождает конфликтные взаимодействия и может оказать как мобилизующее, так и дезорганизирующее влияние на поведение участников конфликта; изменение отношения участников конфликта к возникновению проблемной ситуации и к конфликту в целом [3, с. 19]. В. П. Ратников и другие исследователи выделяют следующие стадии в общей схеме динамики конфликта: латентную стадию (предконфликтную ситуацию); стадию открытого конфликта: инцидент (начало конфликта), эскалацию (развитие) конфликта, разрешение (завершение) конфликта; послеконфликтную стадию [6, с. 31]. А. Я. Анцупов и А. И. Шипилов соотносят основные периоды и этапы динамики конфликта следующим образом: предконфликт (возникновение объективной проблемной ситуации, осознание объективной проблемной ситуации, попытки решить проблему неконфликтными способами, предконфликтная ситуация); конфликт (инцидент, эскалация, сбалансированное противодействие, завершение конфликта); послеконфликтная ситуация (частичная нормализация отношений, полная нормализация отношений) [2, с. 47].

Проблема конфликта как жизненного феномена находится на пересечении интересов ученых разных научных направлений. Для лингвистов большой интерес представляет речевой конфликт как разновидность речевого взаимодействия. Так, В. С. Третьякова отмечает, что конфликт «подразумевает столкновение сторон, состояние противоборства партнеров в процессе коммуникации по поводу несовпадающих интересов, мнений и взглядов, коммуникативных намерений, которые выявляются в ситуации общения» [7, с. 58]. Дополнив данное толкование конфликта речевым аспектом, речевой конфликт можно трактовать как «коммуникативную ситуацию, в которой происходит столкновение двух сторон (участников конфликта) по поводу разногласия интересов, целей, взглядов и т.п., в результате чего одна из сторон (или обе стороны) предпринимают речевые действия, направленные против интересов партнера по коммуникации» [8, с. 20].

В процессе изучения речевого конфликта следует выделить различия между речевым конфликтом и схожими с ним коммуникативными явлениями, такими как спор, дискуссия и полемика. Спор – это борьба двух противоположных мнений и рассматривается как более общее родовое понятие, в которое можно включать дискуссию и полемику. Спор представляет собой «публичное обсуждение проблем, интересующих участников обсуждения, вызванное желанием как можно глубже разобраться в обсуждаемых вопросах» [9, с. 149]. Дискуссией называется публичный спор, целью которого являются выяснение и сопоставления разных точек зрения, поиск, выявление истинного мнения, нахождение правильного решения спорного вопроса. Дискуссия считается эффективным способом убеждения, так как ее участники сами приходят к тому или иному выводу [10, с. 82]. Полемика – это не просто спор, а конфронтация, противостояние, противоборство сторон, идей и речей. Полемику можно определить как борьбу принципиально противоположных мнений по тому или иному вопросу, публичный спор с целью защитить, отстоять свою точку зрения и опровергнуть мнение оппонента. Полемика – это всегда борьба, страстный эмоциональный спор по важным мировоззренческим, идеологическим, нравственным и научным вопросам [10, с. 82]. Все приведенные выше виды спора отличаются от речевого конфликта своим институциональным характером, регламентированностью протекания и (обычно) коллективной формой проведения.

Одним из важных вопросов речевой конфликтологии является проблема типологии конфликтов. По мнению В. С. Третьяковой, единая типология конфликтов отсутствует, что, в свою очередь, обусловлено расплывчатостью и многообразием существующих критериев, лежащих в основе построения данных классификаций [11, с. 124]. Современные представления о типологии конфликтов могут быть систематизированы следующим образом:

1) по количеству участников различают одиночные и групповые конфликты [12, с. 390];

2) в зависимости от характера статусных отношений между коммуникантами выделяют горизонтальные (с участием лиц, не находящихся в отношениях подчинения), вертикальные (с участием лиц, находящихся в отношениях подчинения) и смешанные (сочетающие горизонтальные и вертикальные отношения) конфликты [13, с. 10–11];

3) по остроте противодействия конфликтующих сторон различают конфликты низкой степени интенсивности, средней степени интенсивности, высокой степени интенсивности [14, с. 90];

4) по контролируемости протекания выделяют управляемые конфликты и неуправляемые конфликты [8, с. 21];

5) по характеру конфликтующих сторон или по субъектам-участникам конфликта различают внутриличностные, межличностные, между личностью и группой, межгрупповые, межгосударственные конфликты [8, с. 22];

6) по типу диалогического взаимодействия выделяют обоюдные конфликты, когда оба коммуниканта предпринимают агрессивные речевые действия по отношению друг к другу; односторонние конфликты, когда один из коммуникантов, на которого направлены конфликтные речевые действия, устранился от конфликтного противодействия, не предпринимая ответных агрессивных речевых шагов; гармонизирующие конфликты, характеризующиеся тем, что один из участников конфликта агрессивно активен в своем речевом поведении, в то время как его партнер по коммуникации демонстрирует доброжелательное речевое поведение и также проявляет активность, но направленную на снятие возникшего напряжения и на погашение конфликта [8, с. 22].

В основе речевого поведения участников конфликта лежат определенные речевые стратегии. В. С. Третьякова отмечает, что типология стратегий может быть построена на разных основаниях. Ученый предлагает типологию, в основе которой лежит тип диалогического взаимодействия по результату (исходу, последствиям) коммуникативного события – гармония или конфликт. Если собеседники осуществили свои коммуникативные намерения и при этом сохранили «баланс отношений», значит, общение строилось на основе стратегий кооперации. Взаимодействие партнеров коммуникации в этом случае представляет собой нарастающее подтверждение взаимных ролевых ожиданий, быстрое формирование у них общей картины ситуации и возникновение эмпатической связи друг с другом. Напротив, если коммуникативная цель не достигнута, а общение не способствует проявлению положительных личностных качеств субъектов речи, то коммуникативное событие регулируется стратегиями конфронтации. При осуществлении такого варианта взаимодействия происходит одностороннее или обоюдное неподтверждение ролевых ожиданий, расхождение партнеров в понимании или оценке ситуации и возникновение антипатии друг к другу [7]. К конфронтационным стратегиям относятся инвективная стратегия, стратегии агрессии, насилия, дискредитации, подчинения, принуждения, разоблачения и др., реализация которых, в свою очередь, вносит дискомфорт в ситуацию общения и создает речевые конфликты. Выбор той или иной речевой стратегии обуславливает набор направленных на ее реализацию приемов, т.е. речевых тактик. Как правило, стратегии конфронтации связаны с такими конфронтационными тактиками как угроза, запугивание, упрек, обвинение, издевка, колкость, оскорбление, провокация и др. Существуют также двузначные тактики, которые могут быть как кооперативными, так и конфликтными в зависимости от того, в рамках какой стратегии используется данная тактика. К таким тактикам относится, например, тактика лжи. Она выполняет коопе-

ративную функцию при реализации стратегии вежливости, цель которой «не навредить» партнеру, «поднять» своего собеседника. В то же время данная тактика может быть конфликтным средством при использовании ее в рамках стратегий конфронтации, например стратегии дискредитации. К двузначным тактикам относятся также тактики иронии, лести, подкупа, замечания, просьбы, смены темы и т.п. [7].

Отмеченные выше стратегии и тактики речевого конфликта характеризуются определенными сигналами или маркерами, роль которых выполняют языковые единицы разных уровней, обладающие конфликтными свойствами. Кроме того, речевой конфликт часто усиливается такими невербальными компонентами, как мимика, жесты, позы и различные действия коммуникантов.

ЛИТЕРАТУРА

1. Давыдов, В. В. Психологический словарь / В. В. Давыдов [и др.] ; под общ. ред. В. В. Давыдова. – М. : Политиздат, 1983. – 494 с.
2. Анцупов, А. Я. Конфликтология: теория, история, библиография / А. Я. Анцупов, А. И. Шипилов. – М. : Просвещение, 1996. – 268 с.
3. Вишнякова, Н. Ф. Конфликтология: учеб. пособ. / Н. Ф. Вишнякова. – М. : РИОР, 2002. – 194 с.
4. Гришина, Н. В. Психология конфликта / Н. В. Гришина. – СПб. : Питер, 2004. – 464 с.
5. Шейнов, В. П. Конфликты в нашей жизни и их разрешение / В. П. Шейнов. – Минск : Амалфея, 1997. – 288 с.
6. Ратников, В. П. Конфликтология: учебник для вузов / В. П. Ратников [и др.] ; под ред. проф. В. П. Ратникова. – М. : ЮНИТИ-ДАНА, 2001. – 512 с.
7. Третьякова, В. С. Конфликт как феномен языка и речи / В. С. Третьякова // Проблемы образования, науки и культуры : Извест. Уральского гос. ун-та. – Екатеринбург : Изд-во Уральского гос. ун-та, 2003. – № 27. – Вып. 14. – С. 143–152.
8. Чиркун, А. Б. Вербальные и невербальные средства нейтрализации речевого конфликта (на материале испанского и русского языков) : дис. ... канд. филол. наук : 10.02.19 / А. Б. Чиркун ; Минск. гос. лингв. ун-т. – Минск, 2009. – 121 л.
9. Культура русской речи : Энциклопедический словарь-справочник; под ред. : Л. Ю. Иванова, А. П. Сковородникова, Е. Н. Ширяева. – М. : Флинта : Наука, 2003. – 840 с.
10. Дьячкова, Е. Н. Формы реализации и построения речевого конфликта: спор, дискуссия, полемика / Е. Н. Дьячкова, Н. Р. Швец // Проблемы высшего образования. – 2011. – № 1. – С. 82–83.
11. Третьякова, В. С. Конфликт глазами лингвиста / В. С. Третьякова // Юрислингвистика-2. – Барнаул, 2000. – С. 127–140.
12. Шкатова, Л. А. Речеповеденческие стратегии и тактики в конфликтных ситуациях / Л. А. Шкатова // Философские и лингвокультурологические проблемы толерантности : коллект. монография / Л. А. Шкатова; отв. ред. Н. А. Купина, М. Б. Хомяков. – М. : ОЛМА-ПРЕСС, 2005. – С. 389–402.
13. Козлов, А. С. Регулирование и разрешение конфликтов: стратегии, подходы и индивидуальные стили / А. С. Козлов // Прикладная психология. – 1999. – № 5. – С. 1–23.
14. Латынов, В. Конфликт: протекание, способы разрешения, поведение конфликтующих сторон. Обзор зарубежных исследований по психологии конфликта / В. Латынов // Иностранная психология. – 1993. – № 2. – С. 87–92.

АНАЛИЗ МЕТОДОВ АВТОМАТИЧЕСКОГО ВЫДЕЛЕНИЯ ТЕРМИНОВ ИЗ НАУЧНО-ТЕХНИЧЕСКИХ ТЕКСТОВ

В условиях активной информатизации всех сфер общественной жизни и необходимостью оперативной обработки значительных объемов научно-технической информации, поступающей на иностранных языках, исследования по автоматическому выделению терминов в настоящее время представляются особенно актуальными. Выделение терминов и терминологических словосочетаний из научно-технических текстов при решении задач обработки естественного языка является объектом изучения многих направлений прикладной лингвистики, например, в лексикографии при создании терминологических словарей, автоматического реферирования, аннотирования и индексирования текстов, рубрикации текстов и их тематической структуризации, перевода текстов, а также извлечения знаний из текстовых источников.

Практика перевода и составления терминологических словарей, а также последние теоретические исследования показывают, что одной из наиболее важных и сложных проблем, как при решении традиционных лингвистических задач, так и при создании информационно-поисковых словарей и тезаурусов является отбор терминов.

Вопрос об отборе терминов тесным образом связан с актуальными и дискуссионными проблемами терминоведения, а именно с проблемой определения термина и четких границ, отличающих термин от нетермина, а также методов выделения терминов из научно-технических текстов.

В нашем понимании термин представляет собой элемент общелитературного языка, специфика которого заключается в функции выражения научного понятия, занимающего определенное место в системе понятий некоторой области знания и в сфере его специального распространения [1; 2]. Функциональный подход к различению термина от нетермина заключается в соотносении термина с профессиональным, а нетермина с бытовым понятием. Следует отметить, что к терминам относятся не только слова, но и словосочетания, обозначающие единое, но расчлененное научное или техническое понятие [3; 4].

В специальной литературе предлагаются различные критерии определения терминологичности лексических единиц, например, дефинитивный критерий, критерий концептуальной целостности, семантический критерий, статистический критерий, критерий воспроизводимости, психолингвистический критерий.

На обязательность дефинированности терминов указывали В. П. Даниленко [5, с. 15], Т. Л. Канделаки [6, с. 3–11], Н. З. Котелова [7, с. 125], Э. Ф. Скорородько [8, с. 15] и другие исследователи. Под дефиницией понимается логическое определение понятия, при котором устанавливается содержание понятия (признаки предметов, отражаемых в понятии), его отличительные признаки [3, с. 62]. Дефиниция должна быть емкой, то есть содержать все необходимые и достаточные признаки обозначаемого понятия и отражать признаки, по которому можно отличить одно понятие от другого, потому что «внутренняя форма слова-термина является сложной и много-гранной категорией» [9, с. 17].

Согласно критерию концептуальной целостности, предложенному В. М. Овчаренко, в качестве термина должны признаваться синтетические или семантически целостные образования, выражающие специальные понятия, а не свободные, или аналитические сочетания знаков, которые объединяются по действующим языковым моделям для выражения таких понятий [10, с. 95].

В основе семантического критерия лежит методика логико-семантического соотнесения профессионального понятия и соответствующего ему языкового выражения с опорой на некоторые методические способы определения терминированности лексических единиц [11; 12].

Статистический критерий заключается в формальном выделении терминов без обращения к смыслу и с применением статистических методов. Данный способ может быть осуществлен при наличии статистических описаний по текстам исследуемого подъязыка и любого подъязыка, с которым производится сравнение. Эти описания представлены в виде алфавитно-частотных или частотно-алфавитных словарей [13, с. 7]. Мерой терминологичности языковой единицы в одном подъязыке относительно другого будет выступать величина разности номеров, или рангов, этой единицы в двух подъязиках. Незначительная разность номеров и сходные статистические характеристики в художественных и научно-технических текстах позволяют утверждать, что эти единицы являются общеупотребительными. Если разность для словоформ в научно-технических текстах невелика, но она заметна в художественных текстах, то эти слова или словоформы называются общенаучными терминами. Те лексические единицы, которые характеризуются большой разностью по отношению как к художественному, так и к научно-техническому тексту, называются специальными терминами [12, с. 250–255]. На основе данного подхода была также разработана система определения маркеров специальных областей деятельности [14].

Критерий воспроизводимости основывается на положении, что терминологические словосочетания в речи не производятся, а воспроизводятся [15].

В основе психолингвистического критерия лежит логико-интуитивный подход, который позволяет четко членить текст на непредикативные сегменты, однако затрудняет членение непредикативных терминологических сегментов, выражающих составное специальное понятие [16].

Н. Ю. Зайцева выделяет три группы критериев терминологичности, которые объединяют как признаки терминов, так и желательные или обязательные требования [17, с. 6–7]:

- 1) синтаксические (с точки зрения организации означающего терминологического знака);
- 2) семантические (с точки зрения организации означаемого терминологического знака);
- 3) прагматические или функциональные (с точки зрения употребления терминологического знака в соответствии с задачами коммуникации).

К синтаксическим критериям относятся: краткость, которая подразделяется на лексическую краткость или нетавтологичность – фиксация в форме термина минимального количества отличительных признаков, отсутствие элементов, не несущих смысловой нагрузки, и формальную краткость,

которая часто достигается путем создания эллиптических конструкций и аббревиатур; способность к деривации как важнейшему средству развития и обогащения терминологических систем; соответствие нормам языка, которое достигается устранением профессиональных жаргонизмов, отклонений от фонетических и грамматических норм, а также замещением несвойственных литературному языку форм на нормативные формы [17, с. 58–66].

Семантическая группа включает такие критерии, как обозначение понятия специальной области знания, непротиворечивость отражения терминов концептов понятия, системность (принадлежность обозначаемого термином понятия определенному месту в системе понятий данной предметной области), дефинированность, однозначность, точность значения, независимость смысла термина от контекста, стилистическая нейтральность, конвенциональность и целенаправленный характер появления термина, неизменчивость значения устоявшихся терминов, полноточность, отсутствие синонимов, номинативность [17, с. 7–58].

К прагматическим критериям относятся: внедренность (общепринятость или употребительность термина или ТС), современность (соответствие терминологического знака уровню развития области знания), благозвучность и эзотеричность (намеренная недоступность, обусловленная желанием оградить профессиональное общение от его восприятия неспециалистами) [17, с. 67–70].

По мнению Н. Ю. Зайцевой наиболее важными являются семантические критерии, поскольку специфика терминов лежит не столько в плане выражения, сколько в плане содержания, в характере его значения. Исследователь выделяет три базовых критерия терминологичности: обозначение понятия специальной области знания, системность и дефинированность.

В целях решения задачи автоматического выделения новых терминов и терминологических словосочетаний из научно-технических текстов, а также для формализации такой процедуры для существующих терминов, специалистами разработано и предлагается достаточное количество методов.

Существующие методы классифицируются по следующим критериям: адаптивность, количество слов в извлекаемых терминах, используемые лингвистические ресурсы, применяемые количественные методы, способ фильтрации слов общей лексики [18].

Адаптивные методы автоматического выделения терминов основываются на текстах, которые, с одной стороны, служат для определения правил извлечения терминов, а с другой, являются базой для выделения терминов, что гарантирует достаточно высокие результаты только при работе с данными текстами.

Как известно, статистические методы достаточно эффективны при выделении однословных терминов и терминологических словосочетаний, состоящих из двух компонентов. Определенную сложность представляет собой выделение терминов произвольной длины. Это обусловлено проблемой определения границ терминов. В некоторых случаях для решения данной проблемы, в частности, для выделения именных терминологических словосочетаний может быть использован синтаксический анализатор, главная функция которого заключается в как можно быстрой и точной идентификации основных составляющих (и возможно синтаксических функций) в исходном тексте [19; 20].

Решение этой задачи может осуществляться несколькими путями.

1. Использование маркеров границ термина. На этом этапе текст разбивается на именные словосочетания максимальной длины при помощи лексико-семантических средств, определяющих границы этого типа словосочетаний (глаголы, местоимения, предлоги, определители).

2. Определение следования слов определенных частей речи, в которых может употребляться термин. Конструкции, которые совпадают с заранее заданными словосочетаниями, признаются терминологическими словосочетаниями.

3. Использование синтаксического анализатора для получения информации, необходимой для выделения словосочетаний. По результатам экспериментов данные, полученные анализатором, не являются высокоточными.

Кроме того, для выделения терминологических словосочетаний предлагается процедура определения цепочки слов на основании списка слов и знаков, которые не могут входить в состав терминологического словосочетания [21]. Далее происходит вычисление веса термина-кандидата в зависимости от частоты его самостоятельного употребления и употребления в составе словосочетаний. При этом, чем выше частота употребления в составе других словосочетаний, тем ниже его вес. Такой метод позволяет выделить многокомпонентные терминологические словосочетания.

Автоматическое выделение терминов может также быть основано на данных лингвистического анализа [18]. Выделяют четыре основных уровня данного анализа:

1) графематический анализ, который, например, позволяет выделить адреса электронной почты, даты, ссылки на электронные документы;

2) морфологический анализ, данные которого могут послужить основой для выделения шаблонов терминов, например, именных или глагольных групп;

3) синтаксический анализ, который является основой для поиска лексико-синтаксических шаблонов и морфосинтаксических связей;

4) семантический анализ, который предполагает использование справочной литературы в качестве словарной базы анализа.

Наряду с рассмотренным выше статистическим критерием определения терминологичности лексических единиц, специалисты также предлагают метод рекурсивного вхождения терминов в более длинные термины [21].

Для автоматического выделения терминов и терминологических словосочетаний применяются также статистические и семантические фильтры [22; 23; 24].

Интерес к разработке и использованию статистических фильтров был вызван тем фактом, что частота употребления терминов в специальном контексте отличается от частоты употребления нетерминов. Статистические характеристики могут варьироваться от самых простых (показатель частотности) до сложных (получение данных при помощи формул). Следует отметить, что в условиях небольшого объема корпуса и в результате невысокой частотности употребления единиц, данные сложных фильтров являются приблизительными. Результаты статистических фильтров зависят от того, насколько точно подобран определенный фильтр к данному тексту.

Идея семантических фильтров была предложена исследователями недавно, но уже получила широкое распространение. В основе разработки таких фильтров лежит положение о том, что термины выражают и формируют специальные понятия и, таким образом, имеют определенные семантические характеристики, отличающие их от нетерминов. Эксперимент показал, что применение семантического фильтра увеличила точность полученных данных на 5–9 %.

Анализ существующих методов автоматического выделения терминов из научно-технических текстов показывает, что, несмотря на все положительные стороны и возможности рассмотренных методов, некоторые проблемы остаются нерешенными, и, как следствие, пользователи получают не всегда точный и полный результат. При этом, одним из способов проверки работы того или иного метода является сопоставление полученных результатов и мнения эксперта в исследуемой специальной области знания. В результате, в настоящее время для успешного выделения терминов из научно-технических текстов наряду с использованием современных методов следует учитывать и мнение лингвиста и специалиста определенной сферы деятельности.

ЛИТЕРАТУРА

1. *Винокур, Г. О.* О некоторых явлениях словообразования в русской технической терминологии / Г. О. Винокур // «Труды МИФЛИ» : сб. ст. по языковедению. – М., 1939. – С. 3–54.
2. *Головин, Б. Н.* О некоторых проблемах изучения терминов / Б. Н. Головин // Научный симпозиум «Семиотические проблемы языков науки, терминологии и информатики» : матер. симпозиума; Москва, 1971 г.: в 2 ч.; редкол. : А. Г. Волков (отв. за вып.) [и др.]. – М. : Изд-во Моск. ун-та, 1971. – Ч. 1. – С. 64–67.
3. *Головин, Б. Н.* Лингвистические основы учения о терминах / Б. Н. Головин, Р. Ю. Кобрин. – М. : Высш. шк., 1987. – 104 с.
4. *Хроменков, П.* Современные системы машинного перевода: общие черты и различия / П. Хроменков // Компьютерная лингвистика и обучение языкам : сб. науч. ст. / МГЛУ; редкол.: А. В. Зубов (отв. ред.) [и др.]. – Минск : МГЛУ, 2000. – С. 154–159.
5. *Даниленко, В. П.* Русская терминология: Опыт лингвистического описания / В. П. Даниленко; АН СССР, Ин-т рус. яз. – М. : Наука, 1977. – 246 с.
6. *Канделаки, Т. Л.* Значение терминов и системы значений научно-технических терминологий / Т. Л. Канделаки // Проблемы языка науки и техники. Логические, лингвистические и историко-научные аспекты терминологии : сб. ст.; отв. ред. чл.-кор. АН СССР С. Г. Бархударов. – М. : Наука, 1970. – С. 3–39.
7. *Котелова, Н. З.* К вопросу о специфике термина / Н. З. Котелова // Лингвистические проблемы научно-технической терминологии: матер. совещания, АН СССР г. Ленинград 30 мая–2 июня 1967 г.; редкол.: С. Г. Бархударов (отв. ред.) [и др.]. – М. : Наука, 1970. – С. 122–126.
8. *Скороходько, Э. Ф.* Вариативность и скрытая полисемия в терминологии / Э. Ф. Скороходько // Научно-техническая информация. Сер. 2, Информационные процессы и системы. – М., 2003. – № 4. – С. 15–19.
9. *Кияк, Т. Р.* Лингвистические аспекты терминоведения / Т. Р. Кияк. – Киев : УМКВО, 1989. – 103 с.

10. Овчаренко, В. М. Термин, аналитическое наименование и номинативное определение / В. М. Овчаренко // Современные проблемы терминологии в науке и технике : сб. ст.; отв. ред. В. С. Кулебакин. – М. : Наука, 1969. – С. 91–121.
11. Головин, Б. Н. О некоторых доказательствах терминированности словосочетаний / Б. Н. Головин // Лексика, терминология, стили : межвуз. науч. сб.; Горьков. гос. ун-т им. Н. И. Лобачевского; редкол. : Б. Н. Головин (отв. ред.) [и др.]. – Горький : ГГУ, 1973. – Вып. 2. – С. 57–65.
12. Пиотровский, Р. Г. Статистическое опознание термина / Р. Г. Пиотровский, С. В. Ястребова // Статистика текста : сб. ст.; редкол.: А. И. Киселевский (гл. ред.) [и др.]. – Минск : БГУ, 1969. – С. 249–259.
13. Зубов, А. В. Основы лингвистической информатики : учеб. пособие : в 3 ч. / А. В. Зубов, И. И. Зубова / Минск. гос. пед. ин-т ин. яз.; редкол. : А. Д. Борисевич, С. А. Истомина. – Минск. : МГПИИЯ, 1992. – Ч. 2. – 138 с.
14. Le An Ha. A Practical Comparison of Different Filters Used in Automatic Term Extraction / An Ha Le // Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, 26–28 May 2004. – Lisbon, 2004. – P. 511–514.
15. Кобрин, Р. Ю. О сопоставительном анализе свободных фразеологических и терминологических сочетаний / Р. Ю. Кобрин // Ученые записки Горьковского гос. ун-та. Серия лингвистическая : матер. и исследов. по рус. грамматике и лексикологии / Горьк. гос. ун-т им. Н.И. Лобачевского; редкол.: Н. Д. Русинов (отв. ред.) [и др.]. – Горький : ГГУ, 1970. – Вып. 99. – С. 15–28.
16. Кобрин, Р. Ю. Психолингвистический эксперимент для оценки терминологичности элементов текста / Р. Ю. Кобрин // Автоматическая переработка текста методами прикладной лингвистики. – Кишинев, 1977. – С. 11–12.
17. Зайцева, Н. Ю. Информационно-семиотическая природа термина и типология языков / Н. Ю. Зайцева. – СПб. : Изд-во РГПУ им. А. И. Герцена, 2002. – 84 с.
18. Табарча, А. И. Анализ и сравнение методов автоматического извлечения терминов из текста / А. И. Табарча // Аспирант и соискатель. – 2010. – № 6. – С. 133–137.
19. Hulth, A. Improved Automatic Keyword Extraction Given More Linguistic Knowledge / A. Hulth // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03), Sapporo, 11–12 July, 2003. – Sapporo, 2003. – P. 216–223.
20. Kinyon, A. A language-Independent Shallow Parser Compiler / A. Kinyon // Proceedings of the Annual Meeting of the Association for Computational Linguistics, Toulouse, 6–11 July, 2001. – Toulouse, 2001. – P. 322–329.
21. Браславский, П. Сравнение пяти методов извлечения терминов произвольной длины / П. Браславский, Е. Соколов // Матер. Межд. конф. Диалог'2008 «Компьютерная лингвистика и интеллектуальные технологии». – М. : РГГУ, 2008. – С. 67–74.
22. Le An Ha. A Practical Comparison of Different Filters Used in Automatic Term Extraction / An Ha Le // Proceedings of the Fourth International Conference on Language Resources and Evaluation, Lisbon, 26–28 May 2004. – Lisbon, 2004. – P. 511–514.
23. Maynard, D. Identifying Contextual Information for Multi-Word Term Extraction / D. Maynard, S. Ananiadou // Proceedings of the Fifth International Congress on Terminology and Knowledge Engineering TKE-99, Innsbruck, 25–27 August, 1999. – Vienna : TermNet – Verlag, 1999. – P. 212–221.
24. Paice, C. D. A Three-Prolonged Approach to the Extraction of Key Terms and Semantic Roles / C. D. Paice, W. J. Black // Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'03), Sapporo, 11–12 July, 2003. – Sapporo, 2003. – P. 357–363.

ОСНОВНОЕ СОДЕРЖАНИЕ ТЕКСТА И ЕГО АВТОМАТИЧЕСКОЕ ОПРЕДЕЛЕНИЕ

В настоящее время при обработке текста рядом автоматических систем возникает задача построения его информационного портрета, который в компактной форме характеризует основное содержание текста, то есть описанные в нем события и их участников. Наличие информационного портрета позволяет производить сравнение документов по содержанию (осуществлять автоматическую классификацию, категоризацию и рубрицирование документов); находить документы, близкие по содержанию к заданному; компактно представлять содержание документа в виде списка ключевых тем, затронутых в тексте, или же в форме реферата, то есть в виде набора репрезентативных предложений текста, содержащих упоминания о ключевых темах. Рассмотрим подробнее суть понятия «основное содержание текста» и подходы к его автоматическому выявлению.

Ученые, исследующие проблемы семантики текста, характеризуют его содержание как информацию, заключенную в тексте в целом. Однако, по мнению Л. Латышева, содержание текста находится вне самого текста, а именно в голове отправителя (того, кто создает и передает текст) и в голове адресата (того, кому текст предназначен, кто его воспринимает). Оно складывается из значений, непосредственно представленных в тексте языковыми знаками. Однако содержание не является простой суммой механического сложения значений. Их взаимодействие в тексте носит гораздо более сложный характер. В частности, взаимодействуя друг с другом, широкие, а иногда и расплывчатые значения языковых единиц конкретизируются. Такая конкретизация значения в тексте называется актуализацией значения [1, с. 101].

Известно, что любому тексту свойственно внутреннее и внешнее содержание. По определению А. И. Новикова, внутреннее содержание текста – это «мыслительное образование, которое формируется в интеллекте человека и соотносится с внешней формой не поэлементно, а в целом соответствует всей совокупности данных языковых средств» [2, с. 5]. Внутреннее содержание или семантика текста существует в виде той информации, которая возбуждается под воздействием совокупности языковых средств, поэтому, говоря о семантике текста, имеют в виду «содержание текста в отношении к средствам его выражения» [2, с. 33]. Такое понимание семантики текста является характерным для информационных подходов к ее изучению.

В зависимости от отношения языкового знака к его другим частям различают следующие типы содержания текста [1, с. 102–106].

1. Сигнификативное содержание, образуемое отношением знак – сигнификат. Оно включает в себя как объективные свойства денотата (денотат – предмет или явление, обозначаемое языком в конкретном речевом произведении), так и коннотации (коннотация – любой содержательный компонент, который, не относясь к денотативному содержанию, сопровождает и дополняет его). К сигнификативным коннотациям относятся не все дополнительные

компоненты содержания, а лишь те, которые являются фактом коллективного языкового сознания и как таковые зафиксированы в парадигматическом значении языкового знака (то есть до того, как он употреблен в речи).

2. Содержание на уровне интерпретатора является результатом отношения знак – интерпретатор. В процессе общения люди употребляют языковые знаки не только в соответствии с их семантическим значением, но и в их несобственных значениях, особым образом интерпретируя их. Автор может построить текст с расчетом на то, что реципиент сам дополнит его подразумеваемой информацией. В этих случаях содержание текста выходит за рамки суммы семантических значений образующих его знаков.



Рис. 1. Внутриязыковое содержание текста

3. Внутриязыковое содержание возникает в результате отношения знак – знак. Языковые знаки находятся в разнообразных отношениях сходства и различия, родства и взаимозависимости. Обычно эти отношения не играют самостоятельной роли в тексте и остаются незаметными для участников коммуникации. Но текст можно построить таким образом, что это отношение производности перестает быть незаметным и приобретает смысловую роль. Достигается это путем противопоставления языковых знаков на основе их сходства или различия. Все вышесказанное можно представить схематически следующим образом.

Говоря о смысле текста, обычно имеют в виду ту часть содержания высказывания, которая представляется наиболее важной, главной. Смысл имеет двойное значение: с одной стороны, это то, что автор хочет вложить в свое высказывание, с другой – то, что читатель / слушатель извлекает из него. Смысл получает выражение в тексте при помощи отбора конкретных слов и их распределения по тексту. Так как слово может быть полисемантично, то определение конкретного значения – содержания, непосредственно выраженного совокупностью знаковых единиц – происходит в тексте, когда слова взаимно обуславливают друг друга, в результате чего значение слова меняется в зависимости от его окружения. В тексте раскрывается система предметных отношений – денотатов, закодированных в словесных значениях [3, с. 104–105].

Смысл текста представляет собой воспринятую получателем структуру денотатов текста, на основе которой строится соответствующая модель, не в полной мере соответствующая модели автора текста. Реципиент восприни-

мает текст последовательно. Отдельные фрагменты текста формируют в его сознании совокупность частных моделей, образующих в конечном итоге общую, интегральную, окончательную модель, соответствующую смыслу текста для данного реципиента. Однако знания, получаемые реципиентом при восприятии текста, могут отличаться от знаний, заложенных в текст автором. Другими словами, реципиент может понимать не все и не так, как задумывал автор текста. Понимание текста обусловлено целым рядом экстралингвистических факторов (социокультурным и ситуативным контекстом, эмоциональным состоянием реципиента, его фоновыми знаниями). В смысловой цельности текста отражаются те связи и зависимости, которые имеются в самой действительности (общественные события, явления природы, человек, его внешний облик и внутренний мир, предметы неживой природы и т. д.). Естественно, что явления действительности могут оцениваться людьми по-разному. Если смысл для разных реципиентов различен, то это означает, что понимание ими данного текста различно, то есть в результате восприятия этого текста ими построены разные модели. Если реципиент не может построить модель текста, то смысл для него отсутствует, он не понимает текст. Таким образом, каждый текст имеет только одно содержание, определяемое автором, и может иметь для разных реципиентов различные смыслы.

Как отмечалось выше, понятие «содержание» связано с категорией информативности речи и присуще только тексту. Оно сообщает читателю индивидуально-авторское понимание отношений между явлениями, их значимости во всех сферах жизни. При этом формируются два вида информации: содержательно-фактуальная информация и содержательно-концептуальная информация. Две разновидности информации называют, соответственно, темой и основной мыслью текста. Не только тема, но и основная мысль объединяют предложения текста и придают ему смысловую цельность.

Многие исследователи считают, что следует различать понятия «тема текста» и «содержание текста». При формировании содержания текста автор выделяет элементы описываемой ситуации, приписывая им определенные для данной ситуации значения. В итоге содержание текста определяется его денотативной структурой, формируемой в сознании человека; оно реализуется в тексте путем отбора конкретных слов и их распределением по тексту. Причем подобный отбор осуществляется таким образом, чтобы потенциальный получатель текста мог сформулировать в своем сознании то же представление об описываемом фрагменте ситуации, которое вложил в текст автор. Содержание также определяется выделением существенных связей в тексте – смысловых опорных пунктов. Они являются носителями обобщенного смысла отдельных частей текста и образуют внутриречевую схему, в которую переводится содержание текста в процессе его понимания. Необходимо отметить, что содержание всего текста складывается из содержания входящих в него абзацев.

Тема является одним из важных структурных компонентов текста. Существуют разные трактовки этого понятия. Под темой понимают: 1) содержательный компонент коммуникации; 2) объединяющее начало текста; 3) основной тезис текста, подлежащий дальнейшему развертыванию;

4) смысловое ядро текста, конденсированное и обобщенное содержание текста; 5) основной компонент смыслового начала текста; 6) содержательный компонент ситуации [4, с. 60]. Понятие «тема текста» может рассматриваться в широком и узком смыслах. Широкое понимание темы включает понятие сообщаемой в тексте информации, а узкое – название объекта, о котором идет речь. Но в обоих случаях общей объективной чертой темы является предмет изложения [4, с. 62–63]. Следовательно, тему текста можно представить как отражение в сознании человека наиболее существенных составляющих ситуации, описанной в тексте. Тема текста – это тот инвариант, который остается одним и тем же, как для автора текста, так и для его читателя. Именно тема определяет предметно-тематическую область словаря, где должен осуществляться поиск указанных слов содержания. Тема устанавливает совокупность определенных слов и словосочетаний, отражающих наиболее существенные составляющие описываемой в тексте ситуации и наиболее важные отношения между ними. Т. А. ван Дейк и В. Кинч рассматривают тематическую структуру текста как иерархическую структуру, поскольку тема всего текста может быть описана с помощью его более мелких подтем, которые в свою очередь членятся на еще более мелкие микротемы. Каждое предложение связного текста связано с той или иной подтемой основной темы текста [5].

При автоматическом выделении основного содержания текста центральной является проблема критериев, используемых для выбора наиболее информативных слов и предложений из первичного документа. Чаще всего для этих целей используются статистические, позиционные и лингвосемантические методы [6]. Рассмотрим более подробно статистические методы, позволяющие достаточно быстро и эффективно определять основное содержание текста. Они основаны на использовании статистических параметров для оценки информативности различных элементов текста и учитывают, прежде всего, показатель частоты встречаемости слов в тексте. В результате ранжирования лексики в том или ином документе определяют слова с высоким рангом и их сочетаемость в различных фразах [7, с. 144]. Именно статистические методы используются в наши дни в промышленных компьютерных системах для решения таких задач, как аннотирование и реферирование текстов, их тематическая классификация и кластеризация, смысловой поиск и т.д., которые можно рассматривать в комплексе как задачу тематического анализа. Статистическая информация об отдельных лексических единицах легко извлекается компьютером из текста, и есть все основания полагать, что она адекватно отражает его основное содержание.

В настоящее время существует целая группа статистических методов. Так, согласно одному из них процедура выделения ключевых слов текста сводится к выполнению следующих этапов [8, с. 125].

1. Компьютер составляет по каждому абзацу текста алфавитно-частотный словарь словоформ.

2. Далее все алфавитно-частотные словари абзацев объединяются в единый распределительный алфавитно-частотный словарь словоформ текста. В нем указывается общая частота употребления словоформы в тексте, число абзацев, в которых она встретилась, а также частота употребления в конкретных абзацах.

3. Затем производится сокращение распределительного алфавитно-частотного словаря словоформ текста до словаря потенциальных ключевых (опорных) слов. Эта процедура включает в себя следующие операции:

- из распределительного словаря удаляется вся служебная и общеупотребительная лексика;

- в оставшейся части словаря суммируются частоты всех грамматических форм одного и того же слова;

- в этом же словаре суммируются частоты синонимов (в том числе и контекстуальных);

- из оставшейся части распределительного словаря удаляются слова, которые встретились только в одном абзаце.

4. На последнем этапе словарь потенциальных опорных слов текста делится на две части. В первую часть входят главные опорные слова (ГОС), а во вторую – второстепенные опорные слова (ВОС). Данная процедура осуществляется с учетом коэффициента важности слова, который вычисляется по формуле $K_v = F \cdot m / N \cdot n$, где F – абсолютная частота употребления слова с учетом всех грамматических форм и синонимов; m – количество абзацев, в которых встретилось слово; N – общее число словоупотреблений в тексте; n – общее число абзацев в тексте. Предварительно, в зависимости от длины текста (в словах и абзацах) по специальной формуле определяется средняя (пороговая) величина коэффициента важности [9, с. 45]. Коэффициент важности каждого слова словаря сравнивается с этой величиной. Если коэффициент важности слова выше пороговой величины, то оно относится к числу главных опорных слов, если меньше пороговой величины – к числу второстепенных опорных слов текста. Слова, входящие в каждую из выделенных групп, не однородны по своему содержанию. Они обозначают основные компоненты описанной в тексте ситуации: адресата и адресанта (субъекты), совершаемые ими действия в определенное время и в определенном месте, а также объекты действительности, о которых идет речь в тексте.

Необходимо заметить, что для адекватного формализованного представления основного содержания текста необходимо оперировать не отдельными словами, а определенным набором ключевых слов, который целесообразно представить в виде таблицы основного статического содержания текста (ТОСС). Каждая ТОСС и представляет собой информационный портрет текста.

ЛИТЕРАТУРА

1. *Латышев, Л. К.* Технология перевода : учеб. пособ. для студентов вузов / Л. К. Латышев. – М. : Академия, 2005. – 320 с.
2. *Новиков, А. И.* Семантика текста и ее формализация / А. И. Новиков. – М. : Наука, 1983. – 97 с.
3. *Соколов, А. Н.* Семантика, логика и интуиции в мыслительной деятельности человека : психологические исследования / А. Н. Соколов, Л. Л. Гурова, Н. И. Жинкин. – М. : Педагогика, 1979. – 184 с.
4. *Вишнякова, С. А.* Теоретические основы обучения моделированию научного текста (Русский как иностранный, основной этап обучения) / С. А. Вишнякова. – СПб. : Европейский дом, 2001. – 258 с.

5. Дейк, Т. А. ван. Стратегии понимания связного текста / Т. А. ван Дейк, В. Кинч // Новое в зарубежной лингвистике. – М. : Прогресс, 1988. – Вып. 23. – С. 153–211.
6. Горяник, Л. В. Тематический фильтр текстов / Л. В. Горяник, Г. В. Дорохина // Искусственный интеллект. – Донецк, 2004. – № 4. – С. 580–586.
7. Блюменау, Д. И. Информационный анализ/синтез для формирования вторичных документов : учеб.-практич. пособ. / Д. И. Блюменау. – СПб. : ПРОФЕССИЯ, 2002. – 240 с.
8. Zubov, A. V. Компьютерная лингвистика / А. В. Zubov, И. И. Zubova // Основы лингвистической информатики : в 2 ч. – Минск : МГПИИЯ, 1992. – Ч. 2. – 203 с.
9. Zubov, A. V. Вероятностно-алгоритмическая модель порождения текста / А. В. Zubov // Проблемы порождения текста : в 4 ч. – Минск : МГПИИЯ, 1989. – Ч. 2. – 68 с.

Р. Шимуля

МЕТАФОРИЧЕСКИЕ НАЗВАНИЯ ЖАНРОВ ТЕАТРАЛИЗОВАННЫХ ПРЕДСТАВЛЕНИЙ В СОВРЕМЕННОМ РУССКОМ ПОЛИТИЧЕСКОМ ДИСКУРСЕ

Характерной чертой современного русского политического газетного дискурса является большая выразительность и экспрессивность. Экспрессивная функция данного типа дискурса проявляется, между прочим, в интенсивном употреблении лексем с вторичным метафорическим значением. Политическая метафорика помогает, с одной стороны, заинтересовать читателя, привлечь его внимание, а с другой – вызвать в нем эмоции, создать определенное отношение к рассматриваемым реалиям, преобразовать устоявшуюся в сознании читателя языковую картину мира. Использование в политическом дискурсе лексем с вторичным метафорическим значением является эффективным средством своего рода языковой манипуляции (понимаемой, вслед за О. Н. Быковой, как вид языкового воздействия, используемый для скрытого внедрения в психику адресата целей, желаний, намерений, отношений или установок, не совпадающих с теми, которые имеются у адресата в данный момент [1, с. 99]), так они могут навязать адресату определенное представление о политической действительности, сформировать отношение к ней, вызвать конкретную эмоциональную реакцию.

Метафорический перенос играет важную роль в образовании вторичных значений лексем, используемых в современном русском политическом дискурсе. Очень частотной сферой-источником, на базе которой происходит данное семантическое преобразование, является сфера-источник «Театр». Театральная метафорика изучалась многими учеными (А. Н. Баранов и Ю. Н. Караулов [2; 3], О. Ж. Григорьева [4], А. А. Каслова [5], Дж. Лакофф и М. Джонсон [6], Ю. Б. Феденева [7], А. П. Чудинов [8; 9; 10], Е. И. Шейгал [11], Н. Г. Шехтман [12] и др.), однако ее изучение является все время перспективным, так как конкретные метафорические модели могут подвергаться изменениям под влиянием экстралингвистических факторов. Целью настоящего исследования является подтверждение того, что лексем с вторичным метафорическим значением, образованным на базе сферы-источника «Театр», использованные в русском газетном политическом дискурсе, про-

должают быть продуктивным средством эмоционального воздействия на адресатов. Исследование проводилось на основе новейшего газетного материала (статьи из 4 газет: «Аргументы и факты», «Московский комсомолец», «Известия», «Комсомольская правда», номера за годы 2011–2015).

Проанализированный нами лексический материал можно разделить на несколько тематических групп. В настоящей работе сосредоточим внимание на одной из них – на названиях жанров театрализованных представлений. Примером лексемы, принадлежащей к данному блоку, может служить существительное **комедия** («драматическое представление со смешным, забавным или ироническим сюжетом» [13, с. 444]). Использование в политическом дискурсе данной лексической единицы в ее вторичном метафорическом значении может вызывать у адресата ощущение, что политическая жизнь – это что-то несерьезное, вызывающее смех, а даже лицемерное («Разыгрывать (ломать, играть и т.п.) комедию. Вести себя лицемерно, притворяться» [13, с. 444]). Приведем примеры:

«Я бы не голосовал сам за себя. Сам бы себя назвал клоуном. Для меня выглядело бы это несерьезным, как выглядят все остальные кандидаты, кроме двух правящих. Понятно, что с Немцовым это тоже была комедия» [14].

«Кроме того, известно, что 23 сентября, Госдума рассмотрит проект постановления о снятии его с должности руководителя комитета по информационной политике. В «Справедливой России» заявили, что никакие контакты с депутатом Митрофановым не поддерживаются. «Это какая-то комедия, потому что всем участникам процесса давно известно, что Митрофанов с мая 2012 года не является членом нашей фракции, никаких контактов с ним не поддерживается», – сказал РИА Новости первый зампред фракции СР Михаил Емельянов» [15].

«Но не успел Владимир Путин досрочно покинуть саммит – что долго поминать-то прежний мировой порядок, – не дожидаясь официального закрытия и совместного коммюнике, как Барак Обама тут же организовал встречу с лидерами ЕС для обсуждения ситуации с Россией и Украиной. Ну, вот тут-то все стало на свои места согласно дешевым билетам на дешевую комедию. Что же мешало организовать нечто подобное в те часы, когда Владимир Путин находился на «дружелюбном» и «конструктивном» саммите? Почему бы уж тогда не пригласить российского лидера на эту встречу, а не обсуждать за спиной России ее, России, кровные проблемы?!» [16].

Похожим примером является существительное **фарс**. Его значение – это «1. Вид средневекового западноевропейского народного театра и литературы бытового комедийно-сатирического характера. 2. В театре XIX–XX в.: комедия или водевиль легкого, игрового содержания с чисто внешними комическими приемами» [13, с. 1417]). Данная лексическая единица, использованная в политическом дискурсе во вторичном метафорическом значении, также может отрицательно ассоциироваться с чем-то несерьезным, с шутовством, но также с чем-то лицемерным, циничным и лживым («превращать что-то в фарс»). Приведем примеры:

«В то же время эксперт признает, что сами по себе эти выборы были диким фарсом на фоне событий на юго-востоке страны, «прошли с грубейшими нарушениями норм демократии». Вместо того чтобы провести

сначала референдум о госустройстве страны, принять изменения в Конституцию о статусе русского языка и самостоятельности регионов, Киев стремился поскорее узаконить результаты госпереворота» [17].

«Когда я вижу, что большие фракции коалиции не прислушиваются к мнению других фракций коалиции, когда происходят заговоры, когда, фактически, крупные фракции коалиции голосуют в сговоре с олигархическими группами, в частности Еремеева, Хомутынника, то есть де-факто коалиция существует, а по принципиальным вопросам, там, где нет общего языка, нет общей позиции, договариваются с остатками банды Януковича. Я в этом фарсе участвовать не собираюсь», – сказал он, добавив, что парламентское большинство «обманывает людей и возрождает преступные коррупционные схемы» в прокуратуре и судах» [18].

«Это влечет сразу очень печальные последствия. ЦИК не работает, значит, любые выборы превращаются в фарс. И у нас получается парламент не то что место для дискуссий, а вообще сборище людей, которые никого не представляют» [19].

Особенно сильные отрицательные ассоциации может вызывать использование в политическом дискурсе лексической единицы **клоунада** («цирковой жанр, шуточные или сатирические сценки, разыгрываемые клоунами» [13, с. 433]) во вторичном метафорическом значении. Сравнение политической жизни с цирковым жанром (а в связи с этим мира политики с цирком, а политиков с клоунами) может производить ощущение, что деятельность политиков – это шутовство, паясничанье, дурачество, а политики – это шуты, а не серьезные субъекты политического процесса, заботящиеся о своих избирателях. Приведем примеры:

«Ни одна из парламентских фракций законопроекты не опротестовала, вотума недоверия кабинету министров не вынесла, и документы вступили в силу. Без малейшего публичного обсуждения. **Клоунада для народа**» [20].

«Губернаторские выборы у нас уже превратились в клоунаду. Мало того, что возникло множество фильтров, так еще и возможность любой здоровой конкуренции перечеркнута хитроумной комбинацией, по якобы досрочному сложению полномочий, – подчеркнул политик» [21].

«Вы чувствуете, что политический процесс превращается в некую клоунаду?» [22].

В политическом дискурсе встречаем также словосочетание **мыльная опера** («многосерийный телевизионный сериал» [23, с. 405]). В данном случае политическая жизнь сравнивается с жанром телесериала, состоящего из огромного числа эпизодов, которые снимают и транслируются десятилетиями (у сюжета мыльной оперы почти нет начала и нет конца, она вынуждена сохранять медленный темп повествования). Жанр мыльной оперы ассоциируется с мелодраматическим и эмоциональным характером сюжета, а также с невысокими художественными достоинствами, поэтому его сравнение с политикой может вызывать отрицательные эмоциональные коннотации. Приведем примеры:

«Уже зимой республиканцам и демократам придётся снова сесть за стол переговоров и начать всё сначала. Как вы думаете, будет ли продолжаться эта «мыльная опера», которую они сейчас устроили, или они быстро достигнут соглашения, и все стабилизируется?» [24].

«Наверное, мало кто в Сербии хорошо разбирается в многовековых хитросплетениях отношений России и Украины. Вместе с тем здесь прекрасно помнят события девяностых годов. Поэтому любой международный конфликт, в котором американцы, так или иначе, занимают жесткую позицию, здесь рассматривается как очередная серия мыльной оперы «США устанавливают мировой порядок» [25].

«Российский спектакль на убийстве Немцова превращается в мыльную оперу» [26].

Политическая жизнь сравнивается также с другими жанрами театрализованных представлений, такими как *мелодрама* («1. Музыкально-драматическое произведение, в котором монологи и диалоги действующих лиц сопровождаются музыкой; 2. Драматургический жанр, произведения которого насыщены преувеличенными драматическими эффектами и имеют отчетливо выраженный морально-поучительный характер, а герои отличаются излишней эмоциональностью, сентиментальностью и резко противопоставляются по своему доброму и злему началу; произведения этого жанра» [13, с. 532]), *трагикомедия* («1. Драматическое произведение, в котором сочетаются элементы трагедии и комедии; 2. Печальное и одновременно смешное происшествие» [13, с. 1337]) или *оперетта* («1. Жанр музыкально-комедийного искусства, сочетающий инструментальную (оркестровую) музыку с вокальной (сольной или хоровой), с танцем и текстом; 2. Произведение такого жанра; 3. Сценическая постановка такого произведения» [13, с. 716]), что может создавать впечатление, что политика – это что-то искусственное, напояз, преувеличенное, а деятельность политиков является одновременно печальной и вызывает смех. Приведем примеры:

«Видимо, в июле мы станем свидетелями мелодрамы, во время которой, проникнувшись симпатией к журналисту, депутаты и главы городов, невзирая на свой политический окрас, окажут ему поддержку, дабы независимый кандидат прошел муниципальный фильтр. И все это без денег и партийной поддержки» [27].

«Жители, наблюдающие за этой политической трагикомедией, разочарованы как в тех, кого они избрали, так и в исполнительной власти города. Не удивительно, что терпение активной части горожан лопнуло, и на 28 февраля они заявили митинг против кризиса власти в Протвино, но администрация митинг не согласовала, потому что заявители неверно указали год» [28].

«Все это было бы смешно, когда бы не было так грустно. Трагикомедия с законопроектами-близнецами началась несколько дней назад, когда их рассматривали на заседании фракции «ЕдРа». На прямой вопрос: «Чем ваш законопроект отличается от законопроекта коммунистов?» – глава думских «медведей» Андрей Метельский не нашел ничего более оригинального, чем ответить: «Тем, что он написан от сердца» [29].

«Те режиссеры, которые пытались из химкинских выборов сделать оперетту «День выборов», своего достигли. Например, мы знаем, что подписи господину Пауку собирали те же люди, которые собирали их Шахову. Нам не дали их проверить, но интересно, те люди, которые подписались и за Паука, и за Шахова понимают, что делают?» [30].

Собранный нами материал показал, что сфера-источник «Театр» остается все время одной из продуктивных сфер-источников для возникновения метафорических вторичных значений лексем, использованных в газетном политическом дискурсе. Лексические единицы, составляющие проанализированную группу названий жанров театрализованных представлений, эмоционально описывают отношения в мире политики и могут вызывать у читателей отрицательные ассоциации с чем-то неестественным, несерьезным, далеким от реальной жизни. Данные лексемы могут быть эффективным средством языковой манипуляции сознанием адресатов, так как могут навязывать читателям определенное представление о политическом мире, формировать нужное отношение к нему, вызывать эмоциональную реакцию.

Театральность политического дискурса связана со спецификой его адресата (массовый наблюдатель, а не прямой адресат). Следует, однако, подчеркнуть также роль, которую играют в политическом дискурсе средства массовой информации. СМИ не только передают информацию, но также воздействуют на читателей. Оказание воздействия на адресатов может быть реализовано при помощи экспрессивности высказываний. К числу языковых средств, благодаря которым журналисты достигают экспрессивности печатных материалов, принадлежат также лексические единицы с вторичным метафорическим значением.

ЛИТЕРАТУРА

1. *Быкова, О. Н.* Языковое манипулирование : матер. к энциклопедическому словарю «Культура русской речи» / О. Н. Быкова // Теоретические и прикладные аспекты речевого общения : Вестн. Рос. риторической ассоциации. – Красноярск : Краснояр. гос. ун-т, 1999. – Вып. 1 (8). – С. 99–103.
2. *Баранов, А. Н.* Русская политическая метафора : матер. к словарю / А. Н. Баранов, Ю. Н. Караулов. – М. : Ин-т русского языка АН СССР, 1991. – 193 с.
3. *Баранов, А. Н.* Словарь русских политических метафор / А. Н. Баранов, Ю. Н. Караулов. – М. : Помовский и партнеры, 1994. – 396 с.
4. *Григорьева, О. Ж.* Политический театр современной России (взгляд филолога) / О. Ж. Григорьева // Интернет-журнал «Полемика». – 2001. – № 9.
5. *Каслова, А. А.* Метафорическое моделирование президентских выборов в России и США (2000 г.) : дис. .. канд. филол. наук / А. А. Каслова. – Екатеринбург : Урал. гос. пед. ун-т., 2003. – 208 с.
6. *Лакофф, Дж.* Метафоры, которыми мы живем / Дж. Лакофф, М. Джонсон // Теория метафоры. – М. : Прогресс, 1990. – С. 387–415.
7. *Феденева, Ю. Б.* Моделирующая функция метафоры в агитационно-политических текстах 90-х г. XX века : дис. ... канд. филол. наук / Ю. Б. Феденева. – Екатеринбург : Урал. гос. пед. ун-т., 1997. – 181 с.
8. *Чудинов, А. П.* Россия в метафорическом зеркале: Когнитивное исследование политической метафоры (1991–2000) / А. П. Чудинов. – Екатеринбург : Урал. гос. пед. ун-т., 2001. – 238 с.
9. *Чудинов, А. П.* Метафорическая мозаика в современной политической коммуникации : монография / А. П. Чудинов. – Екатеринбург : Урал. гос. пед. ун-т., 2003. – 248 с.

10. Чудинов, А. П. Политическая лингвистика / А. П. Чудинов. – М. : Флинта : Наука, 2006. – 254 с.
11. Шейгал, Е. И. Театральность политического дискурса / Е. И. Шейгал // Единицы языка и их функционирование : межвуз. сб. науч. тр. – Саратов : СГАП, 2000. – Вып. 6. – С. 92–96.
12. Шехтман, Н. Г. Сопоставительное исследование театральной и спортивной метафоры в российском и американском политическом дискурсе : дис. ... канд. филол. наук / Н. Г. Шехтман. – Екатеринбург : Урал. гос. пед. ун-т, 2003. – 227 с.
13. Большой толковый словарь русского языка / сост. и гл. ред. С. А. Кузнецов. – СПб. : Норинт, 1998. – 1536 с.
14. Аргументы и факты [Электронный ресурс]. – Режим доступа : <http://www.aif.ru/politics/world/27705>. – Дата доступа : 17.04.2015.
15. Московский комсомолец [Электронный ресурс]. – Режим доступа : <http://www.mk.ru/politics/2014/09/19/bastrykin-rekomenduet-gosdume-pomoch-sledstviyu-v-poiske-mitrofanova.html>. – Дата доступа : 17.04.2015.
16. Известия [Электронный ресурс]. – Режим доступа : <http://izvestia.ru/news/579495>. – Дата доступа : 17.04.2015.
17. Аргументы и факты [Электронный ресурс]. – Режим доступа : <http://www.aif.ru/euromaidan/prediction/1177159>. – Дата доступа : 21.05.2015.
18. Московский комсомолец [Электронный ресурс]. – Режим доступа : <http://www.mk.ru/politics/2015/04/22/timoshenko-mozhet-zanyat-mesto-lyashko-v-rade.html>. – Дата доступа : 21.05.2015.
19. Комсомольская правда [Электронный ресурс]. – Режим доступа : <http://www.msk.kp.ru/radio/stenography/28689/>. – Дата доступа : 21.05.2015.
20. Московский комсомолец [Электронный ресурс]. – Режим доступа : <http://mk.kn.md/articles/2014/08/07/gde-iskat-milliard-ionice.html>. – Дата доступа : 17.04.2015.
21. Известия [Электронный ресурс]. – Режим доступа : <http://izvestia.ru/news/571549>. – Дата доступа : 17.04.2015.
22. Комсомольская правда [Электронный ресурс]. – Режим доступа : <http://www.msk.kp.ru/radio/stenography/44085/>. – Дата доступа : 17.04.2015.
23. Толковый словарь русского языка конца XX века. Языковые изменения / под ред. Г. Н. Складчиковой. – СПб. : Рос. акад. наук, Ин-т лингвист. исслед., 1998. – 700 с.
24. Аргументы и факты [Электронный ресурс]. – Режим доступа : <http://www.aif.ru/onlineconf/1392848>. – Дата доступа : 17.04.2015.
25. Московский комсомолец [Электронный ресурс]. – Режим доступа : <http://www.mk.ru/politics/article/2014/03/05/994226-v-belgrade-proshel-miting-v-znak-podderzhki-rossii.html>. – Дата доступа : 17.04.2015.
26. Комсомольская правда [Электронный ресурс]. – Режим доступа : <http://www.kp.ru/online/news/1996645/>. – Дата доступа : 17.04.2015.
27. Московский комсомолец [Электронный ресурс]. – Режим доступа : <http://murmansk.mk.ru/articles/2014/07/09/velikolepnaya-pyatyorka-i-vratar.html>. – Дата доступа : 17.04.2015.
28. Московский комсомолец [Электронный ресурс]. – Режим доступа : <http://serp.mk.ru/articles/2015/02/25/bezvlastie-bezvolie-ili-bezumie.html>. – Дата доступа : 17.04.2015.
29. Московский комсомолец [Электронный ресурс]. – Режим доступа : <http://www.mk.ru/politics/2011/10/31/638279-v-mosgordume-ne-podelili-znamya-pobedy.html>. – Дата доступа : 17.04.2015.
30. Аргументы и факты [Электронный ресурс]. – Режим доступа : <http://www.aif.ru/onlineconf/1392717>. – Дата доступа : 17.04.2015.

РЕФЕРИРОВАНИЕ И КОММЕНТИРОВАНИЕ В СЦЕНАРИСТИКЕ КАК ТЕКСТОВЫЕ ТРАНСФОРМАЦИИ

Исследование свертывания при реферировании текста как последовательного устранения деталей, инициированное более полувека назад И. П. Севбо [1, с. 51], представляется как процедура редуцирования, свертывания текста, обратная хорошо известной в риторике амплификации, наращивания подробностей. Соответственно, обратной задачей реферирования можно рассматривать расширение текста комментированием. Особую разновидность таких задач, более сложных, чем реферирование и комментирование научно-технических текстов, из-за своей неопределенности и множественности интерпретаций, является переработки художественного текста в сценарий. В. И. Немирович-Данченко удачно описал практику подготовки роли артистами как задачу «наговаривать себе внутренние монологи, раскрывающие суть, зерно образа» [2, с. 96]. Более того, такая задача построения собственного комментария к роли рассматривалась как обязательная: «Никак нельзя, чтобы актер жил на сцене только теми словами, какие он сейчас произносит ... каждая фигура несет с собой что-то недосказанное ...» [цит. по: 3, с. 237]. Создается контрапункт внутреннего монолога из реплик «в сторону», произносимых и дополняющих текст роли. Такой контрапункт является дериватом записанного, произносимого текста, результатом его преобразования.

Ключевую роль в построении трансформационной грамматики для порождения комментария из текста получает предикация – приписывание признаков описываемым в тексте предметам. С позиций функциональных представлений подразумевается требование сохранения актуальных предикатов – рем как инварианта текстовых преобразований. Для определения такого инварианта по ремам значима так называемая предикабельность – само свойство понятия в роли предиката обозначать признак в отвлечении от предмета и приписывать его иному предмету. Соответственно, понятия, обладающие этим свойством или предикабилии, по Л. П. Гокиели, несут ответ на вопрос, «как можно чему-либо приписать нечто отличное от него» [4, с. 148]. Потенциальные предикаты в языке (в противоположность речи) представлены глагольной системой. Далее понятие предикабельности обобщается в разработанном Н. Ю. Шведовой понятии дейксиса и дейктических глаголов – предельно обобщенных, передающих свою семантику дополнениям и обстоятельствам.

Со своей стороны, требование инвариантности предикации обусловлено ее значимостью для представления описываемой в тексте ситуации, в которой выявляется конфликт, противоречие, лежащее в основе решаемой в тексте проблемы. Описание такого противоречия осуществляется через противопоставленные признаки, то есть при посредстве предикации. Приписывание различающихся или противоположных признаков предметам позволяет описать мотивы, лежащие в основе ситуации и определяющие антитезы проблематики текста. А. В. Бондарко, обосновывая понятие категориальных

ситуаций, отличал их от общих ситуаций как «нефокусированных» [5, с. 24] и предполагающих внутренние различия, доводимые до контрастов и конфликтов. Задачи представления текста как постановки и решения некоторой проблемы вызывает вопрос о степени полноты и определенности текста, в частности, о подтексте и, шире, латентном содержании, для раскрытия которого требуется комментарий. Максимальная определенность и полнота задается пропозициональной структурой и предполагает сохранение ее при преобразовании. При таком условии основными дериватами текста являются переводы высказываний в вопросительную и отрицательную формы.

Основу такого преобразования составляет известная техника общих и частных диктальных и модальных вопросов. Общие диктальные вопросы относятся к описываемой ситуации, а модальные вопросы предполагают вскрытие интенции, стоящей за высказыванием, истинности намерений. Диктальный вопрос ставится априори, модальный относится к уже высказанному апостериори. Соответственно, вопросы строятся направлены на актуальные предикаты – ремы сообщения и зависят от того, какова актуализация предикатов в перспективе повествования. Тут возникает вопрос о полипредикативности высказывания, обнаруживаемой уже в распределении рем, изменяемой с условиями актуализации (не говоря уже о различии и переменности субъектов модуса и диктума), и соответственно, о множественности ситуаций – полиситуативности [6, с. 250]. Типичным представителем такой переменности являются пословицы, истолкование которых определяет место ремы, а потому и смысл целого.

Так называемые сценические пословицы, построенные через развертывание соответствующего афоризма, наглядно демонстрируют преобразования пропозициональной структуры. Таковы преобразования пословицы *«Свято место пусто не бывает»* $\leftrightarrow$ *если место свято, то оно не пусто*. Диктальные частные вопросы: $\leftrightarrow$ – *Каково место? – Весьма свято! – Значит, пустовать не будет.* $\leftrightarrow$ *Каково наше святое место? – Да пока пусто! – Ну, долго пустовать не будет, ибо свято*. Модальные частные вопросы: $\leftrightarrow$ *Да так ил уж свято это место? – Очевидно, таки свято, ибо не пусто* $\leftrightarrow$ *Занято ли это место? – Нет, пока пусто, хоть и свято*. Альтернативные дериваты: $\leftrightarrow$ *Пустое место не свято* $\leftrightarrow$ *Место пусто, хотя и свято*. Ситуация пословицы описывается конфликтом СВЯТОСТЬ vs. ПУСТОТА. Такое противопоставление демонстрирует участок лексического ряда совместимости (на основе ситуативной частичной синонимии или антонимии). Приведенный пример позволяет сформулировать последовательность шагов преобразования высказывания в вопросы: определить рему, заменить ее многоточием (для диктального вопроса) или антонимом (для модального вопроса), задать вопросительное слово, изменить порядок слов.

Такие трансформации позволяют развернуть текст в диалогическую структуру типа катехизиса – чередования вопросов и ответов, альтернатив и возражений. В свою очередь, построение катехизиса не исчерпывает раскрытия латентной информации текста. Помимо производных «логического квадрата», получаемых в катехизисе из вопросов и отрицаний, требуется привлечение опыта, предполагаемого у адресата текста, раскрытие пресуппозиции, построение возможных импликаций из высказываний. Такие процеду-

ры предполагает интерпретация лирического стихотворения, предполагающая переработку его в солилоквию – разговор с этим текстом как с высказыванием Иного. Особенность лирики заключается в тотальной предикации: «Предложение в составе такого предельно тесного единства, каким является поэтический текст, утрачивает в принципе одно из главных коммуникативных свойств предложения–высказывания – способность члениться на тему и рему ... Лирическое стихотворение развивает и варьирует одну тему. Поэтому оно представляет сплошную рему» [7, с. 151]. Помимо превалирования предикации, лирика характеризуется преобладанием созерцательности и, соответственно, перцептивного аспекта, так что все высказывания могут быть сведены к страдательному залогу. Эти два обстоятельства – сквозная предикативность и перцептивность – влекут за собой неполноту и неопределенность лирической абстракции, откуда проистекает, в частности, специфический именной стиль изложения, построенного как последовательность назывных предложений, выводящий за пределы пропозициональных структур. В силу своей неполноты лирика представляет собой лакунарный текст, требующий дополнения воображением читателя и вызывающий его на спор.

Построение полемического контрапункта к поэтическому тексту можно продемонстрировать интерпретацией К. И. Чуковского, засвидетельствованной воспоминаниями М. Алигер. Так, читая «Признание» Е. А. Баратынского, он перемежал строки собственными комментариями: «*“Грущу я, но и грусть минует ...”* <...> Он уже давно не грустит, смею вас уверить <...> *“Подругу без любви – кто знает? – изберу я”* <...> Он то уж отлично знает и давно уже избрал ее, и отнюдь не без любви, старая лиса!» [8, с. 278]. Такому раскрытию поэтического подтекста у К. И. Чуковского служили и экспериментальные преобразования стиха в прозу, разрушавшие идиоматичность.

Переход от пропозициональной структуры к словосочетаниям как самостоятельным единицам текста, связанный с неопределенностью и неполнотой лирики, осуществляется в частности, через превращение участков текста обособлением. Такие обособленные члены утрачивают определенность по некоторым категориям, в частности, лица, места, времени, что выражается в предложенном В. В. Виноградовым понятии полупредикативности. Вместе с тем, иные категории предикации, такие как модус и функция ремы, сохраняются, что позволяет говорить о предикативности вне пропозициональности и называть ее носителей «второстепенными сказуемыми» [9, с. 307]. Полупредикативность особенно ярко проявляется в минимальных изотопных цепочках, ограниченных парой лексических единиц – ситуативных частичных синонимах или антонимах, риторической фигуре *гендиадес*, где обнаруживаются признаки зависимости и управления. Обособленные словосочетания с полупредикативными признаками предстают как потенциальные распространители предложения с признаками неполноты.

Именно в силу частичной предикативности перечислительные конструкции не сводятся к линейным последовательностям, обнаруживая иерархические отношения между словосочетаниями. Таков образец именного стиля в стихотворении Игоря Северянина: *«Они живут политикой, раздорами и войнами, / Нарядами и картами, обжорством и питьем, / Интригами и*

сплетнями, заразными и гнойными, / Нахальством, злобой, завистью, развратом и нытьем» (Игорь Северянин, «Чем они живут», 20.01.1923) [10, с. 212]. Полупредикативность здесь создает иерархию благодаря построению перечисления как ряда фигур гендиадес, выделенных союзами *и*. Примечателен прием корреспондирующего стиха (*versus rapportari*) в третьей строке. Ситуация обличения определяет подбор мотивов, среди которых выделяются кульминационной позицией эпитеты ГНОЙ & ЗАРАЗА. Основной конфликт определяется позицией наблюдателя, для которого неприемлемо описываемое состояние мира. Аспект состояния и модус антипатии позволяют построить такой вариант построчного комментария: <→ бытие войной за счет иных жизней – людоедство – братоубийство; → роскошь из горя других – прожигание жизни; → ложь и обман – основа их духа, она болезненна и смертоносна; → живущие за счет других ненавидят ими же ограбленных>. Использование полупредикативности позволяет выстраивать комментарий к лирическому стихотворению в виде цепочки реплик «в сторону», составленных из словосочетаний, выводимых из комментируемого текста. По такому принципу предлагаются и следующие сценарные комментарии к цитатам.

«Любовница – дивное русское слово, / И как бы ты смел на него напасть, / Когда оно – жизни основа / И в нем сочетались любовь и страсть?! / В этом слове есть что-то неверное, / Драматическое что-то есть, / Что-то трогательное и нервное, / Есть оправдываемая месть» (И. Северянин, «Любовница», 1929) [10, с. 59]. Две ситуации – защита образа от возможного порицания и раскрытие потаенного смысла – определяют мотивы жизнетворчества и ранимости жизни. Конфликт ситуации определяется как апологией образа, так и раскрытием ее внутреннего драматизма: жизнеутверждающее начало всегда стоит рядом с угрозой гибели. Аспект перехода к иному состоянию определяется указанными ситуациями. Возможные сопровождающие построчные реплики: <→ диво – очарование слова; → порицание порицает самого хулителя; → образ равноценен дару жизни; → неустойчивость, экзистенциальная ситуация риска страстного бытия – противоречием обозначено само сочетание; → драма – конфликт, присущий образу, и его театральность, видимость; → любовь ранима – хрупкость бытия; → месть – любовь и гибель рядом, Эрос – Танатос>. Проблема ОПРАВДАНИЕ БЫТИЯ (очерченная автором через «оправдание мести») вырисовывается за деталями образа.

Ситуация общения над разделенным временем бытием дана в одном стихотворении литературоведа Ивана Новикова, посвященного мысленной встрече с покойной женой. Тут всплывает давний мотив тихого звука, слышного только избранным (известный, в частности, из стиха Фридриха Шлегеля *Durch alle Töne tönet / Im bunten Erdenraum / Ein leiser Ton gezogen / Für den, der heimlich lauschet* ‘Сквозь все звуки пестрого земного сна тянется тихий звук для того, кто тайно вслушивается’): «Между звуков обычных / Для воскресшего дня / От земных всех отличный / Звук коснулся меня / Он другими не слышим, / Лишь моею душой» (Иван Новиков, «Звук особый», 24.07.1955) [11, с. 455]. Аспект тут определяется однократностью события, равнозначного чуду. Предлагаемые возможные реплики-комментарии раскрывают полемику с обыденностью: <→ значит, имеется необычное звуча-

ние – настороженность для его встречи: → день воскресает как Христос – наше бытие в дневной яви открыто Воскресению; → прикосновение звука, а не просто вслушивание, есть знамение». Конфликт стихотворения выражен мотивами СОКРОВЕННОЕ vs. ПОВСЕДНЕВНОЕ.

Наконец, резко конфликтная ситуация противостояния суеты и созерцательного бытия представлена в стихотворении А. Блока: «Ветер стих, и слава зоревая / Облекла вон те пруды. / Вон и схимник. Книгу закрывая, / Он смиренно ждет звезды. / Но бежит шоссе, / Убегает вбок ...» (А. Блок. «Ветер стих ...», 08.1914) [12, с. 271]. Мотив ОБЛАЧЕНИЕ ЗАРЕЙ, снизошедшее на монастырские пруды, контрастирует мотиву ПУТЬ, который в заключительной строке направляется «Мимо инока, прудов и звезд». СМИРЕНИЕ инока противостоит БОКУ, в который ведет шоссе. ОЖИДАНИЕ тут – антоним БЕГУ. Аспект эпизода подчеркивает несовместимость причастности вечности и беготни, которая всегда бесцельна, всегда «мимо», как указано в финале. Возможные построчные реплики дополняют такое раскрытие смысла: <→ подразумеваемый вечер, слава дня Божьего – упомянутый стихийный ветер (подсказка аллитерацией); → обзор – с прудов на обитель схимника, на книгу (точка зрения наблюдателя); → взор схимника обращен к небу; → дорога крива, сворачивает, уводит от осмысленного бытия». В конечном счете такие комментарии призваны воссоздать характеристику подразумеваемых действующих лиц, в том числе и безличных сил, таких, как описанная дорога, участвующих в ситуации.

Итак, переработка текста в сценарий представляется как его авто-дескрипция, где комментарии порождаются развертыванием подтекста. Лирическое стихотворение включается в солилоквию – внутренний монолог интерпретатора и дополняется комментариями в виде реплик «в сторону», чередуемых со стихами. Конечная задача таких реплик – выявить конфликт, лежащий в основе текста, определяющий его проблемность. Задачи, поставленные когда-то В. И. Немировичем-Данченко, представляется возможным рассматривать и в плоскости риторических процедур преобразования текста в комментарий и реплики.

ЛИТЕРАТУРА

1. Севбо, И. П. Структура связного текста и автоматизация реферирования. – М. : Наука, 1969 – 136 с.
2. Кнебель, М. О. Школа режиссуры Немировича-Данченко / М. О. Кнебель. – М. : Искусство, 1966 – 168 с.
3. Виленкин, В. Я. Вл. И. Немирович-Данченко. Очерк творчества / В. Я. Виленкин. – Издание Музыкального театра им. Народного артиста СССР Вл. И. Немировича-Данченко, 1941. – 326 с.
4. Гокиели, Л. П. Логика / Л. П. Гокиели. – Тбилиси : Мецниереба, 1967 – 272 с.
5. Бондарко, А. В. Категориальные ситуации. К теории функциональной грамматики / А. В. Бондарко // Вопросы языкознания. – 1983. – № 2. – С. 20–32.
6. Золотова, Г. А. Коммуникативная грамматика русского языка / Г. А. Золотова, Н. К. Онипенко, М. Ю. Сидорова // РАН. Ин-т рус. яз. им. В. В. Виноградова; Моск. гос. ун-т им. М. В. Ломоносова. – М., 1998. – 528 с.

7. *Ковтунова И. И.* Поэтический синтаксис / И. И. Ковтунова – М.: Наука, 1986 – 210 С.
8. *Алигер, М.* Долгие прогулки / М. Алигер // Воспоминания о Корнее Чуковском; изд. 2-е. – М. : Советский писатель, 1983 – С. 274–294.
9. *Фурашов, В. И.* Современный русский синтаксис / В. И. Фурашов – Владимир : ВГГУ, 2010 – 370 с.
10. *Северянин, Игорь.* Классические розы. Медальоны / Игорь Северянин. – М. : Художественная литература, 1991 – 224 с.
11. *Новиков, И. А.* Собр. соч. : в 4 т. / И. А. Новиков. – М. : Художественная литература, 1967 – Т. 4. – 560 с.
12. *Блок, А.* Собр. соч. : в 8 т. / А. Блок. – М., Л. : Гослитиздат, 1960. – Т. 3. – 716 с.

A. A. Boldak-Brial

L'ASPECT DIALOGIQUE DES PROCESSUS COGNITIFS DANS L'ÉCRIT DE VULGARISATION SCIENTIFIQUE

La vulgarisation scientifique a une grande importance dans le développement de la science aujourd'hui, car elle représente une forme de diffusion pédagogique des connaissances scientifiques qui cherche à mettre le savoir à portée de tous et chacun. L'explication au grand public des bases et des avancées d'un domaine scientifique peut être réalisée non seulement par le biais de la communication verbale directe mais aussi à travers la communication écrite. La connaissance et les opinions sur le monde et son évolution sont formées et modulées à travers le prisme des informations auxquelles le public a accès, la plupart venant de la presse écrite.

L'écrit scientifique de popularisation d'essence est sociocognitif car il nécessite l'interaction des individus (plus précisément, les scientifiques et le lectorat) dans le but d'accroître leur degré d'information.

En premier lieu, la diffusion des savoirs au public est «une pratique inscrite dans le champ social, agie par les acteurs sociaux et surdéterminée par leurs enjeux» [1, p. 14]. Ainsi, le discours écrit de popularisation scientifique peut être vu comme un texte de nature sociale et communicationnelle car il est souvent caractérisé par la présence des énoncés qui sont formellement dialogiques. D'après Bakhtine, le dialogue peut facilement être observé dans l'espace textuelle. Selon Bakhtine dans 'Le Marxisme et la philosophie du langage' (1977), tout texte est fondamentalement dialogique car «à la base de la division du discours en parties, dénommées paragraphes dans leur forme écrite, on trouve l'ajustement aux réactions prévues de l'auditeur et du lecteur» [2, p. 158]. Sans doute, les éléments dialogiques donnent au discours écrit une nouvelle dimension: «Le dialogue peut émerger au sein d'un texte non-dialogué, sous une forme esquissée et naître comme spontanément. C'est le cas lorsque l'auteur met en scène des formes d'adversité polémique controversiale, en répartissant les voix en une alternance d'objections et réponses... C'est le cas également lorsqu'il se fait des objections à lui-même. Ces émergences dialogiques sont localisées et surtout subordonnées étroitement au plan d'énonciation assumé par le repère énonciatif-source» [3, p. 40].

Les textes de divulgation scientifique sont différenciés par leur caractère social et communicationnel car ils sont destinés à faciliter le dialogue entre

vulgarisateur et lecteur. Dans le cadre d'un texte pareil, le dialogisme constitutif se déploie sous la forme d'une interaction inter-discursive mais aussi interlocutive. Daniel Jacobi dans 'Diffusion et vulgarisation: itinéraires du texte scientifique' (1986) constate que «toutes les multiples formes du dialogisme (réflexion sur son propre discours, discours sur une partie du discours, utilisation du discours d'autrui, discours dans le discours et énonciation sur l'énonciation) se retrouvent aisément dans le discours scientifique et le discours de vulgarisation scientifique» [4, p. 78].

Le dialogisme inter-discursif présuppose que «un énoncé est rempli des échos et des rappels d'autres énoncés, auxquels il est relié à l'intérieur d'une sphère commune de l'échange verbal. (...) Un énoncé est tourné non seulement vers son objet mais aussi vers le discours d'autrui portant sur cet objet» [5, p. 302]. Certains linguistes, comme Jacqueline Authier-Revuz dans '*Le fait autonymique: langage, langue, discours – quelques repères*' (2003), proposent de nommer ce type de dialogisme comme 'autodialogisme' ou 'dialogisme constitutif' [6]. Selon cette forme de dialogisme, le texte est traversé par une multiplicité des voix et des types de discours. Julia Kristeva, pour sa part, propose la notion de l'intertextualité pour parler d'une interaction textuelle au sein d'un texte [7]. Concernant le contexte, voici quelques exemples des énoncés avec les éléments inter-discursifs présents dans les textes anglophones sur la cosmologie: «*Feynman proclaimed that each electron that makes it through to the phosphorescent screen actually goes through both slits. It sounds crazy, but hang on: Things get even more wild.*» [8, p. 110], «*Our old friend Ibn al-Haytham was one of the first to attempt to explain the physical basis of a rainbow in the tenth century. He described them as being produced by the 'light from the Sun as it is reflected by a cloud before reaching the eye'. This isn't too far from the truth*» [9, p. 56], «*In his classic science fiction novel 'Star Maker', Olaf Stapledon envisages, among many imaginative cosmoses, a 'musical universe, where creatures appeared to one another as complex patterns and rhythms of tonal character'*» [10, p. 147–148], «*Most of these authors don't seem to have realized the fact that if you can travel faster than light, the theory of relativity implies you can also travel back in time, as the following limerick says: There was a young lady of Wight Who travelled much faster than light. She departed one day, In a relative way, And arrived on the previous night*» [11, p. 108]. Ainsi, dans un texte vulgarisé tous les éléments se trouvent dans diverses corrélations sémantiques avec les autres textes et les systèmes sémiotiques.

A côté de ce dialogisme constitutif ou inter-discursif, il faut noter l'existence d'un dialogisme interlocutif. Dans ce type de dialogisme, selon Bakhtine, «un énoncé, cependant, est relié non seulement aux maillons qui le précèdent mais aussi à ceux qui lui succèdent dans la chaîne de l'échange verbal (...) l'énoncé, dès son tout début, s'élabore en fonction de la réaction-réponse éventuelle, en vue de laquelle il s'élabore» [5, p. 303–304]. Bakhtine précise que ce type de discours «anticipe la réponse active de l'autre et se modèle en conséquences» [5, p. 303]. Comme le montrent les quelques exemples suivants, les vulgarisateurs font un usage très important de la modalité interrogative pour créer un dialogue avec le lectorat: «*Which Friedmann model describes our universe? Will the universe eventually stop expanding and start contracting, or will it expand forever? It turns out the answer to this question is more complicated than scientists first thought*»

[11, p. 64], «*Friendly skeptics sometimes ask me: "Isn't it presumptuous for cosmologists to claim to explain anything about the vast cosmos?" My response is that what makes things hard to understand is how complicated they are, not how big they are. Under extreme conditions – inside the stars or in the hot early universe – everything breaks down to its simplest ingredients*» [10, p. 63], «*In Murray's eyes, looking at the rings of Saturn is like looking at the Solar System four and a half billion years ago, with the Sun at the centre surrounded by a disc of dust not unlike Saturn's rings. It is this similarity, the history of our solar system contained within the rings that convinces Murray of the importance of studying them. "If we can't understand a disc of material that's in our own backyard, what chance do we have of understanding a disc that's long since disappeared?"*» [9, p. 89], «*Notwithstanding their common features, an examination of the fundamental forces themselves serves only to compound the questions. Why, for instance, are there four fundamental forces? Why not five or three or perhaps only one? Why do the forces have such different properties? Why are the strong and weak forces confined to operate on microscopic scales while gravity and the electromagnetic force have an unlimited range of influence? And why is there such an enormous spread in the intrinsic strength of these forces? To appreciate this last question, imagine*» [8, p. 12]. Les exemples qui viennent d'être donnés montrent que la catégorie 'question-réponse' peut être absente (dans le cas des questions rhétoriques qui ne nécessitent pas de réponse) ou présente quand l'auteur introduit une question avec la réponse immédiate (ce qui donne l'illusion d'un dialogue interactif).

En d'autres termes, le dialogisme inter-discursif est basé sur le discours qui intègre d'autres discours sur le même sujet, et le dialogisme interlocutif nécessite la compréhension et la réponse. On peut déduire que les écrits de popularisation scientifique (en apparence formellement monologiques) privilégient les unités de l'interaction et les séquences communicationnelles.

En deuxième lieu, un texte de vulgarisation scientifique est non seulement une plateforme de l'échange communicationnel mais aussi celui de la formation et de l'élaboration des pensées. Comme précise Trognon dans '*Eléments d'analyse interlocutoire*', une communication ou une séquence conversationnelle «est à la fois un événement social et un événement cognitif» [12, p. 73]. Selon Mikhaïl Bakhtine, «La vérité naît entre les hommes qui la cherchent ensemble, dans le processus de leur communication dialogique» [13, p. 137]. Et il continue dans '*La Poétique de Dostoïevski*': «L'idée vit, non pas dans une conscience individuelle isolée, mais naît, se développe, trouve et renouvelle son expression verbale, engendre d'autres idées, seulement dans des rapports dialogiques, avec les idées d'autrui, c'est-à-dire dans sa conscience exprimée par le discours. C'est au point de contact de ces voix-consciences que naît et vit l'idée» [13, Ibidem]. Autrement dit, le dialogue vise à enrichir le bagage cognitif de chaque individu.

L'importance de développer les capacités cognitives chez un individu Bakhtine montre sur l'exemple du 'dialogue socratique'. 'Le dialogue socratique' s'appuie sur les notions de la syncrèse (c'est-à-dire «la confrontation de divers points de vue sur un sujet donné» [13, p. 166], et de l'anacrèse («es moyens de faire naître, de provoquer le discours de l'interlocuteur, de l'obliger à exprimer son opinion, et de pousser celle-ci jusqu'à ses limites» [13, Ibidem]. Selon Bakhtine, «la syncrèse et l'anacrèse dialogisent la pensée, la placent à l'extérieur, la

transforment en réplique, la rattachent à la communication dialogique entre les hommes» [13, Ibidem]. Dans le cadre de cette recherche, cet aspect cognitif de la popularisation scientifique, qui se centre sur le rapport entre une conscience et un signe, est de haute importance car il est omniprésent dans les écrits des vulgarisateurs.

En popularisation scientifique, le discours écrit joue un rôle constitutif dans le dialogue entre les savants et les non-spécialistes. La diffusion et massification des connaissances techniques et scientifiques s'impose comme une pratique communicationnelle à caractère socio-cognitif: le divulgateur instaure le dialogue avec un vaste public profane dans le but de briser la distance entre la science et le lectorat. L'auteur vise non seulement à transmettre des connaissances au simple lecteur, mais surtout à traduire des données savantes en un langage accessible aux non-spécialistes, expliquer des notions complexes et scientifiques par le biais d'un langage simple. Ainsi, on peut définir un écrit de vulgarisation scientifique comme une séquence dialogale à caractère sociocognitif: il n'est pas seulement le processus de l'échange communicationnel mais aussi un outil pour la formation et l'élaboration des pensées.

REFERENCES

1. *Jacobi, D.* Vulgariser la science – Le procès de l'ignorance / D. Jacobi, B. Schiele. – Seyssel : Milieux, 1988. – 284 p.
2. *Bakhtine, M.* Le marxisme et la philosophie du langage: essai d'application de la méthode sociologique en linguistique, trad. de russe par M. Yaguello / M. Bakhtine, V. N. Volochinov – Paris : Minuit, 1977. – 232 p.
3. *Cossutta, F.* Le dialogue: introduction à un genre philosophique / F. Cossutta. – Villeneuve d'Ascq : PU du Septentrion, 2005. – 274 p.
4. *Jacobi, D.* Diffusion et vulgarisation: itinéraires du texte scientifique / D. Jacobi. – Besançon : Presses Universitaires Franche-Comté, Besançon, 1986. – 186 p.
5. *Bakhtine, M.* Esthétique de la création verbale / M. Bakhtine. – Paris : Gallimard, 1952. – 408 p.
6. *Authier-Revuz, J.* Le fait autonymique: langage, langue, discours – quelques repères, (w:) red. Authier-Revuz A., Doury M., et Reboul-Touré S., Parler des mots. Le fait autonymique en discours / J. Authier-Revuz. – Paris : Presses universitaires de la Sorbonne Nouvelle, 2003. – P. 67–95.
7. *Kristeva, J.* Sémiotikè. Recherche pour une sémanalyse / J. Kristeva. – Paris : coll. Points, Seuil, 1969. – 384 p.
8. *Greene, B. R.* The Elegant Universe: superstrings, hidden dimensions, and the quest for the ultimate theory / B. R. Greene. – New York, London : W.W. Norton & Company, 2003. – 448 p.
9. *Cox, B.* Wonders of the Universe / B. Cox, A. Cohen. – London : Cohen HarperCollins Publishers, 2011. – 256 p.
10. *Rees, M.* Our Cosmic Habitat / M. Rees. – London : Princeton University Press, 2001. – 224 p.
11. *Hawking, S.* A Briefer History of Time / S. Hawking, L. Mlodinow. – London : Bantam, 2005. – 176 p.
12. *Trognon, A.* Eléments d'analyse interlocutoire, (w:) red. Gilly M., Roux J. P. et Trognon A., Apprendre dans l'interaction: Analyse des médiations sémiotiques / A. Trognon. – Aix en Provence et Nancy : PUP et PUN, 1999. – 400 p.
13. *Bakhtine, M.* La Poétique de Dostoïevski, trad. du russe par I. Kolitcheff / M. Bakhtine. – Paris : Seuil, 1988. – 366 p.

АВТОМАТИЧЕСКОЕ ВЫЯВЛЕНИЕ ОТЗЫВОВ И МНЕНИЙ В СЕТИ ИНТЕРНЕТ

М. А. Белюга

МНОГОУРОВНЕВЫЙ ПОДХОД К ИЗМЕРЕНИЮ СТЕПЕНИ СМЫСЛОВОГО ПОДОБИЯ ТЕКСТОВ

Современные исследования в области прикладной лингвистики характеризуются повышенным интересом к решению проблемы измерения степени смыслового подобия текстов. Решение данной задачи приобрело особую актуальность с точки зрения его применения в сферах цифрового и дистанционного обучения, информационного поиска, при разработке систем машинного перевода, систем автоматического аннотирования и реферирования текстов, вопросно-ответных систем, систем распознавания плагиата и многих других.

В свете обозначенной выше актуальности задачи многие исследователи предпринимали и предпринимаяют попытки разработки эффективных алгоритмов выявления схожих или идентичных по смыслу текстов, в том числе с использованием новейших достижений в области машинного обучения для формирования интегральной оценки смыслового подобия текстов на основании лексического сходства текстовых фрагментов без учета и с учетом порядка следования слов, их канонических форм и синонимов.

В данной работе представлен алгоритм оценки степени смыслового подобия текстов на английском языке, принципиальным отличием которого от уже существующих решений является сравнение предложений не только на лексико-грамматическом и синтаксическом уровнях, но и на уровне семантики. Алгоритм направлен на попарный анализ предложений сравниваемых текстов с тем, чтобы оценить степень их семантического сходства в виде интегральной оценки по непрерывной шкале значений от 0 до 5, где оценка 0 соответствует семантически несвязанным предложениям, а оценка 5 характеризует одинаковые по смыслу пары входных предложений [1]. Так, несмотря на лексическое и синтаксическое сходство приведенная ниже пара предложений (а) должна получать низкую оценку семантического подобия, тогда как оценка пары предложений (б) должна быть значительно более высокой вне зависимости от лексических различий:

а) “American air forces plane crashes in south Iraq.” – “American air forces aircraft crashes in the south of Lybia.”;

б) “Sarkozy makes official the re-election bid” – “France’s Nicolas Sarkozy announces his reelection bid.”.

Указанная интегральная оценка степени сходства предложений формируется посредством обучения регрессионной модели на основании метода опорных векторов, реализованной в программном пакете LIBSVM¹, с исполь-

¹ <https://www.csie.ntu.edu.tw/~cjlin/libsvm/>

зованием множества признаков подобия, вычисляемых для каждой пары входных текстов. Предполагается, что анализ семантики входных текстов способен привести качественные улучшения в работу алгоритмов, созданных в рамках описанного выше традиционного подхода.

Понятно, что решение поставленной задачи требует формирования набора базовых лингвистических, алгоритмических и программных ресурсов. В частности, был определен базовый лингвистический процессор с функциональностью лексического, лексико-грамматического, синтаксического и семантического анализов текста, а также обоснована и получена лингвистическая база знаний, включающая словари и эталонные корпуса текстов из разных предметных областей (заголовки газетных статей, описания изображений, сообщения пользователей различных форумов и проч.) для разработки и анализа алгоритма решения задачи. Так как качество ее решения напрямую зависит от точности и глубины лингвистического анализа текста, разработанный алгоритм был реализован с использованием лингвистического процессора IHS Goldfire, производящего обработку текстовых документов, начиная от предварительного их форматирования и заканчивая семантическим анализом [2; 3]. Данный лингвистический процессор ориентирован на промышленную обработку текстовых документов для целого ряда естественных языков (включая английский) и имеет лучшие на настоящее время показатели эффективности среди аналогичных разработок. Лексический, синтаксический и семантический анализ основывались на разработанной базе лингвистических правил, а также на собранных статистических данных, необходимых для получения признакового описания объектов (текстовых фрагментов), используемого в последующем для метода машинного обучения.

Принципиальная структурно-функциональная схема системы автоматического распознавания семантической эквивалентности текстовых фрагментов изображена на рис. 1.

На **предварительном этапе** при помощи ЛП IHS Goldfire текст входных предложений подвергался ряду операций нормализации, упрощающих дальнейшую обработку строковой информации:

- 1) поиск границ слов (токенизация);
- 2) нормализация написания слов и символов (*don't = do not, can't = can not; 'em = them; he's = he is* и т.д.);
- 3) приведение различных вариантов написания кавычек к единому виду;
- 4) анализ содержимого скобочных контекстов внутри предложения и их представление в виде отдельных предложений в случае необходимости;
- 5) удаление стоп-слов:
 - a. артикли и им подобные слова: *a, an, the, some, every, another, any, some, this, these, those* и т.д.;
 - b. неинформативные наречия: *very, a bit, also, however, likely* и т.д.;
 - c. дискурсивные маркеры: *I think, he said, ИМНО, as for me* и т.д.;
- 6) удаление знаков пунктуации и приведение всех слов входных предложений к нижнему регистру.

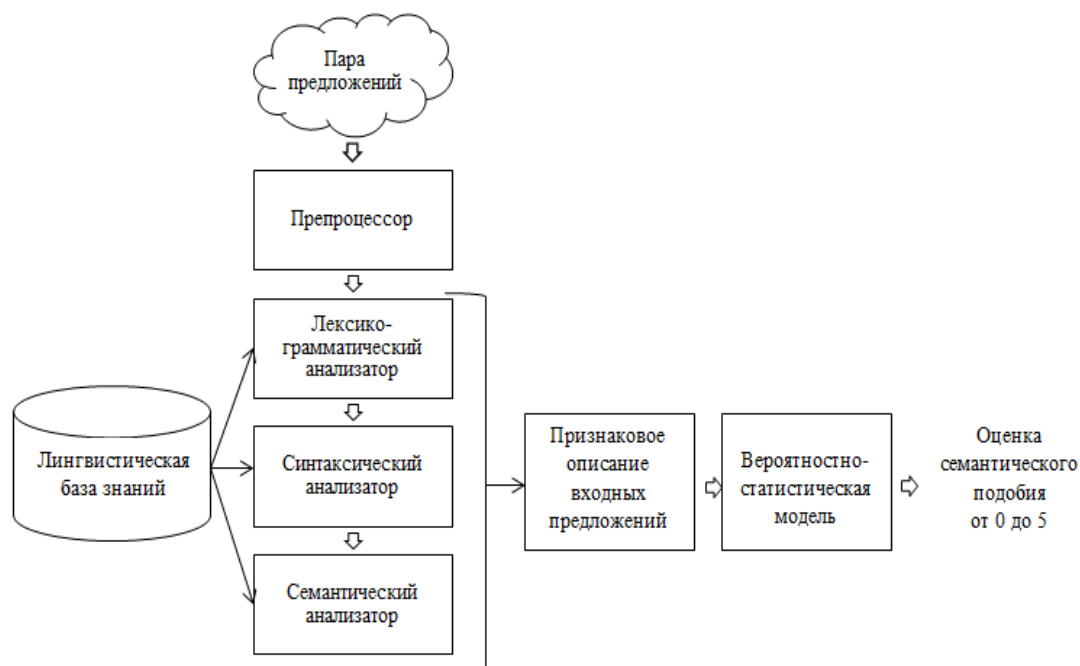


Рис. 1. Принципиальная схема решения задачи

Этап лексико-грамматического анализа заключался в идентификации семантически подобных слов и фраз при помощи:

- базы данных лексических, фразовых и синтаксических парафраз Paraphrase Database (PPDB) [4];
- разработанные нами для решения данной задачи словари, включающие:
 - 14870 синонимичных пар для прилагательных, глаголов и существительных (“*wrong*” – “*incorrect*”, “*link*” – “*connect*”, “*seller*” – “*vendor*”);
 - 1435 пар прилагательное + производное наречие (“*clear*” – “*clearly*”);
 - 2953 пар действие + агент (“*connect*” – “*connector*”);
 - 17556 пар глагол + отглагольное существительное (“*pulsate*” – “*pulsation*”);
 - 563 пары существительное + образованное от него прилагательное (“*sinusoid*” – “*sinusoidal*”);
 - 264 пары, состоящих из названия страны и соответствующей национальности (“*uk*” – “*british*”);
 - 262 пары, включающих название страны и её столицу (“*uk*” – “*London*”).

Деривационные списки были собраны автоматически с использованием слов из словаря ЛГК и деривативных аффиксов, а затем проверены на, случайным образом отобранных, корпусах объёмом 2 миллиона предложений из различных областей, включая тексты Wikipedia, научных статей и диссертаций, периодических изданий и патентов. При этом производное слово считалось валидным синонимом в том случае, если оно встречалось на корпусе не менее четырёх раз. Для составления синонимичных пар стран, столиц и национальностей были использованы ресурсы Wikipedia;

- программного продукта Google word2vec¹;

¹ <https://code.google.com/archive/p/word2vec/>

Данный этап завершался установлением семантической корреляции между подобными словами и фразами [5] с последующим расчетом предварительной оценки подобия предложений на основании меры Жаккара [6] и важности слова для каждого предложения (TF-IDF). Мера Жаккара рассчитывалась по формуле: $J(S_1, S_2) = \frac{|S_1 \cap S_2|}{|S_1 \cup S_2|}$, где S_1 и S_2 – вектора первого и второго предложений соответственно. Формирование дополнительных признаков на основании расчета меры Жаккара также производилось после этапов синтаксического и семантического анализа предложений. Мера подобия $sts(S_1, S_2)$ на основании оценки TF-IDF важности слов предложений рассчитывалась по формуле: $sts(S_1, S_2) = \frac{\sum_{\omega \in (S_1^a \cup S_2^a)} tfidf(\omega)}{\sum_{\omega \in (S_1 \cup S_2)} tfidf(\omega)}$, где числитель представляет собой сумму значений TF-IDF всех слов пары предложений, для которых были установлены отношения семантического подобия, а знаменатель – сумму значений TF-IDF всех слов обоих предложений. Таким образом, повышение значимости семантически подобных слов пары предложений непосредственно влияло на повышение данной меры подобия и наоборот.

На этапе синтаксического анализа при помощи лингвистического процессора IHS Goldfire производилось построение синтаксического дерева каждого предложения из входной пары. Полученные деревья сравнивались на предмет поиска корреляции в синтаксических ролях слов, отмеченных как подобные на этапе лексического анализа [7]. Например, в паре входных строк *“the notebook of my mother”* и *“my mother cooks amazing”* синтаксические роли слова *“mother”* различны: в первом случае это роль атрибута в именной группе, тогда как во втором – это роль подлежащего. Данный признак указывает на различие этих предложений на семантическом уровне.

На этапе семантического анализа вышеупомянутый лингвистический процессор использовался для выделения собственно концептов и семантических отношений между ними типа «субъект – акция – объект» (CAO), где каждый элемент отношения может иметь свои атрибуты. Данные структуры полностью соответствуют таким классическим типам знаний как объекты/классы объектов и факты об объектах. Из полученных отношений извлекались бинарные связи типа «субъект-акция», «акция-объект», «акция-непрямой объект» и др. с целью нахождения корреляции между предложениями на семантическом уровне. В данном случае наподобие предложений указывало полное или частичное совпадение «семантических ролей» слов, между которыми были установлены отношения подобия на этапе лексического анализа. При частичном совпадении, если подобные слова именных частей речи находятся в паре с глаголами, для которых отношений подобия не выявлено (например: *“predict”* и *“walk”*), то это свидетельствует о различиях на семантическом уровне. Оценка подобия предложений, формируемая на данном этапе, равнялась:

– сумме значений подобия для каждой из пар бинарных связей. Значение подобия для полностью совпадающей пары бинарных связей равнялось 1. Для частично совпадающих бинарных связей значение рассчитывалось следующим образом: 1 умножалась на долю совпадающих слов.

- числу минус 1 в случае отсутствия совпадений;
- числу 0 при отсутствии предиката и, соответственно, отсутствии отношения типа САО в одном или обоих входных предложениях.

Полученные промежуточные оценки подобиya использовались в качестве признакового описания пары предложений для **обучения регрессионной модели**, необходимой для формирования интегральной оценки семантического сходства предложений в непрерывном диапазоне значений от 0 до 5.

При разработке и тестировании алгоритма использовался разработанный эталонный корпус экспертных оценок подобиya для 11306 пар предложений из разных доменных областей (заголовки публицистических статей, фрагменты научных статей, художественных текстов и др.), который был поделен на обучающую и тестовую выборки. Проведенная оценка качества работы алгоритма показала, что точность предсказаний составила 82,9 %. На наш взгляд, данный показатель свидетельствует о состоятельности выбранного подхода и позволяет надеяться на дальнейшее улучшение полученных результатов за счёт более точной настройки алгоритма посредством добавления новых признаков подобиya и различия предложений, необходимых для обучения регрессионной модели.

ЛИТЕРАТУРА

1. Eneko Agirre; Carmen Banea; Claire Cardie; Daniel Cer; Mona Diab; Aitor Gonzalez-Agirre; Weiwei Guo; Inigo Lopez-Gazpio; Montse Maritxalar; Rada Mihalcea; German Rigau; Larraitz Uribe; Janyce Wiebe. SemEval-2015 task 2: Semantic Textual Similarity, English, Spanish and Pilot on Interpretability. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), Denver, CO, June.
2. *Todhunter, J.* System and method for automatic semantic labeling of natural language texts / J. Todhunter, I. Söyler, D. Pastanohau. – U.S. Patent 8 583 422, November 12, 2013.
3. *Чеусов, А. В.* Разработка лингвистических процессоров промышленной обработки текстовых документов / А. В. Чеусов // Искусственный интеллект. Интеллектуальные и многопроцессорные системы: матер. науч.-технич. конф., п. Кацивели, 25–30 сентября 2006 г.; в 3 т.; ред. В. О. Бронзов. – Таганрог : ТРТУ, 2006. – Т. 2. – С. 366–370.
4. Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The Paraphrase 152Database. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 758–764.
5. Md Arafat Sultan, Steven Bethard, and Tamara Sumner. 2015. DLS@CU: Sentence similarity from word alignment and semantic vector composition. In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 148–153, Denver, Colorado, June. Association for Computational Linguistics.
6. *Jaccard, P.* Étude comparative de la distribution florale dans une portion des Alpes et des Jura / P. Jaccard // Bulletin de la Société Vaudoise des Sciences Naturelles 37. – Lausanne, 1901. – P. 547–579.
7. Jesús Oliva, José Ignacio Serrano, María Dolores del Castillo, Ángel Iglesias. 2011. SyMSS: A syntax-based measure for short-text semantic similarity. Data & Knowledge Engineering.

СРЕДСТВА ФОРМИРОВАНИЯ ТОНАЛЬНОСТИ ПИСЬМЕННОГО ТЕКСТА

Современная лингвистика уделяет большое внимание исследованию социальных, психологических, этнических и других особенностей человека. Его невозможно представить беспристрастным сторонним наблюдателем событий, процессов и явлений объективной действительности, поэтому неотъемлемой частью сознания человека является оценочное отношение к реальному миру. Выражение отношения к познаваемому объекту – особый вид деятельности индивида, приводящий к его позитивной или негативной оценке. Одним из способов выражения оценочного аспекта взаимодействия человека и объективной действительности является язык, который располагает необходимыми средствами почти на всех уровнях.

Понимание смысла, как отдельного предложения, так и всего текста, подразумевает не только определение его информационного содержания, но и выделение прагматической направленности, выраженной категорией оценочности. На пересечении понятий фактуальной и оценочной составляющих смысла текста находится понятие тональности. Тонально-окрашенными принято считать такие элементы текста, которые несут в себе оценочную семантику: позитивную, нейтральную или негативную.

Способы выражения тональности в тексте весьма вариативны и проявляют себя на разных языковых уровнях путем использования графических, словообразовательных, лексических, синтаксических и стилистических средств языка. Центральное место в выражении тональности текста занимает лексический уровень, представленный позитивно-, негативно- и модальнооценочными словами и словосочетаниями [1, с. 17]. Значение слова или его варианта обладает эмоциональным компонентом, если выражает определённую эмоцию либо чувство. Эмоциями принято считать относительно кратковременные переживания, такие как радость, тревога, огорчение и т.д. К чувствам относятся более устойчивые отношения: ненависть, уважение, любовь и т.п. Эмоциональный компонент возникает на базе предметно-логического и со временем может его вытеснить [2].

Наиболее явным способом выражения эмоций являются междометия – совершенно особый пласт лексики, не обладающей предметно-логическим значением. По мнению И. В. Арнольд, «в междометиях сосредоточены все типические черты, отличающие эмоциональную лексику: синтаксическая факультативность, то есть возможность опущения без нарушения отмеченности фразы; отсутствие синтаксических связей с другими частями предложения; семантическая иррадиация, которая состоит в том, что присутствие хотя бы одного эмоционального слова придаёт эмоциональность всему высказыванию» [2, с. 106]. Тем не менее, многие эмоционально окрашенные слова (в том числе и междометия) выражают эмоцию в самом общем виде, без уточнения её положительного или отрицательного характера. Примером может служить междометие *Oh*, которое способно выражать радость, печаль

и многие другие эмоции. В таком случае при анализе тональности текста или его фрагмента приходится опираться на другие маркеры. Как отмечает И. В. Арнольд, «с эмоциональной лексикой не следует смешивать слова, называющие эмоции или чувства, например, *fear, delight, gloom, cheerfulness, annoу*, и слова, эмоциональность которых зависит от ассоциаций и реакций, связанных с денотатом, например, *death, tears, honor, rain*. С лингвистической точки зрения, это – разные группы слов» [2, с. 107].

Помимо эмоционального компонента слово может обладать и оценочной составляющей, которая непосредственно влияет на тональность текста. Оценочный компонент неразрывно связан с предметно-логическим компонентом, уточняет его и даже может входить в словарную дефиницию. Оценочная лексика характерна для описания общественной жизни и политических событий и, нередко, использует разные виды переносных значений, в то время как прямые значения нейтральны.

Способы выражения тональности текста также представлены на словообразовательном уровне. Например, суффикс *-ish* в сочетании с основой имени прилагательного может образовывать модальные «тактичные» слова в случае, если автор не хочет звучать очень категорично, например, *baldish, biggish, dullish*. Этот же суффикс, добавленный к основе имени существительного, образует имя прилагательное с отрицательной оценкой, например, *childish, doggish, sheepish*. Отрицательная оценка усиливается, если данный суффикс присоединяется к сложным основам, например, *stand-offish, honey-moonish*. Исключением являются имена прилагательные *boyish* и *girlish*, не обладающие отрицательной оценочностью. Наиболее значительными суффиксами отрицательной оценки являются *-ard, -ster, -aster, -eer* и полуаффикс *-monger*, например: *drunkard, gangster, poetaster, profiteer, panic-monger*. Уменьшительные суффиксы создают слова как с ласкательно-шутливым, так и с презрительным оттенком, например, *lambkin* (суффикс *-kin*), *starlet* (суффикс *-let*), *weakling* (суффикс *-ling*), *daddy* (суффикс *-y*), *oldie* (суффикс *-ie*). Синтаксически сложные слова, в частности, названия лиц, образованные не по моделям слов, а по синтагматическим моделям, содержат, как правило, отрицательную оценку, например, *Miss what's-her-name, a might-have-been*.

На грамматическом уровне тональность выражается с помощью морфологических и синтаксических средств. В морфологическом аспекте на первый план выходит частотная характеристика употребления языковой единицы. Наиболее распространённым морфологическим средством выражения тональности текста являются прилагательные и наречия, то есть признаковые части речи. Модальные глаголы и повторения также несут значительную оценочную нагрузку. К синтаксическим средствам, выражающим тональность, относят, в первую очередь, неполные и эллиптические предложения, усечение, инверсию и т.д. [1, с. 19].

Рассмотрим подробнее морфологические способы выражения тональности. Как известно, любая грамматическая форма может иметь несколько значений, одно из которых считается главным, а остальные переносными, то есть тонально-окрашенными. Таким образом, слова различных частей речи приобретают определённую тональную окраску в случае их употребления в

необычных лексико-грамматических и грамматических значениях. Как отмечает И. В. Арнольд, «такое расхождение между традиционно обозначающим и ситуативно обозначающим на уровне морфологии называется транспозицией (или иногда грамматической метафорой). Выражение эмоций, оценки и экспрессивность, а иногда и функционально-стилистические коннотации осуществляются при этом за счет нарушения привычных грамматических валентностных связей» [2, с. 138].

Каждая часть речи имеет при транспозиции свою специфику. Например, имена существительные, в соответствии с присущими им грамматическими категориями, проявляют свои экспрессивные возможности, в первую очередь, при необычном употреблении форм числа и падежа, а также при местоименном замещении. Если рассматривать имена прилагательные, то единственным способом выражения тональности будут манипуляции с категорией сравнения. Данная категория охватывает только качественные и количественные прилагательные. Употребление сравнительной или превосходной степени остальных прилагательных, которым данная категория несвойственна, придаёт им значительную экспрессивность. Наибольшим количеством грамматических категорий в современном английском языке обладает глагол. Исходя из этого, можно утверждать, что экспрессивный потенциал данной части речи значителен. Наиболее ярко выраженной тональностью обладают продолженные формы, также заслуживают внимания глаголы и глагольные формы с модальным значением, поскольку они передают отношение автора к сообщаемым фактам.

Эмоциональность и экспрессивность могут быть переданы в тексте не только особым подбором слов, но и их особым размещением (синтаксический уровень). В английском языке, как известно, порядок слов жёсткий: у каждого члена предложения есть свое постоянное место, определяемое способом его синтаксического выражения, связями с другими словами и типом предложения. Нарушение порядка следования членов предложения неизменно ведет к акцентированию того или иного слова и приобретению им специальных коннотаций экспрессивности или эмоциональности. Так возникает инверсия. Еще одним средством повышения эмоциональности является риторический вопрос.

Общеизвестно, что утвердительные и отрицательные предложения имеют не одинаковую вероятность употребления в тексте. Утвердительные предложения встречаются намного чаще. Таким образом, отрицательные предложения получают особую эмоциональную окраску. Явной тональностью обладает такое стилистическое средство как литота – употребление отрицательной частицы со словом, уже содержащим отрицательный префикс, например, *it is not unlikely, he was not unaware of*. Конструкция с литотой может оказывать разное воздействие в сочетании с разной стилистической окраской: от иронии до осторожности [2, с. 173–175].

Тональность письменного текста выражается также и на стилистическом уровне. Как известно, все стилистические средства подразделяются на изобразительные и выразительные. Изобразительными средствами языка считаются все виды образного употребления слов и словосочетаний, объединя-

ющие все виды переносных наименований общим термином «тропы». Изобразительные средства используются для описания объектов, предметов, явлений и в основном являются лексическими. К ним относятся такие виды использования слов и выражений в переносном значении, как метафора, метонимия, гипербола, ирония, перифраз и т.д. Следующий вид стилистических средств – выразительные средства или фигуры речи – не создает образов, а повышает выразительность речи и усиливает ее эмоциональность при помощи инверсий, риторических вопросов, параллельных конструкций и т.п. Тем не менее, деление стилистических средств на изобразительные и выразительные условно, так как тропы наравне с выразительными средствами могут выполнять экспрессивную функцию, а фигуры речи нередко участвуют в создании образности. На стилистическом уровне тональность текста находит выражение, прежде всего, в виде метафоры. Игра слов, персонификация и аллегория также достаточно употребительные и эмоционально насыщенные средства выражения тональности в тексте. Антономазия, метонимия и близкая к ней синекдоха иногда могут оспорить первенство метафоры в текстах определенной тематики [1, с. 20].

На тональность письменного текста могут влиять такие средства графического уровня языка, как шрифтовое оформление текста, его деление на абзацы, расположение строк, заглавные буквы, знаки препинания. Среди графических стилистических средств важное место принадлежит пунктуации, поскольку наряду с функцией членения предложения на составляющие его синтаксические части, членением текста на предложения и определением общей характеристики предложения (вопрос, восклицание, утверждение) пунктуация указывает и на многие элементы, важные в эмоционально-экспрессивном отношении. Пунктуации принадлежит важная роль в передаче отношения автора к высказываемым им мыслям, в намёке на подтекст, в подсказке эмоциональной реакции, которую ожидают от читателя. Тем не менее, стилистическая нагруженность различных знаков препинания неодинакова. Наибольшей выразительностью обладают восклицательный и вопросительный знаки. Важную роль играют также эмоциональные паузы, выраженные с помощью тире и многоточия. Написание слов заглавными буквами (THAT, NO) или курсивом (*will*, *grey-ish*) придает дополнительную эмфазу всему предложению. Таким образом, тональность всего текста складывается из тональностей входящих в него предложений, которые, в свою очередь, формируются с помощью средств разных языковых уровней.

ЛИТЕРАТУРА

1. Серова, С. А. Лингвистический анализ новостных газетных статей: оценочный аспект : автореф. дис. ... канд. фил. наук : 10.02.01 / С. А. Серова; Тамб. гос. ун-т. – Тамбов, 2007. – 19 с.
2. Арнольд, И. В. Стилистика современного английского языка (стилистика декодирования) / И. В. Арнольд. – СПб. : Просвещение, 1981. – 295 с.

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ФАКТОВ ИЗ ТЕКСТОВ ДОСЬЕ

В настоящее время для автоматической обработки больших неструктурированных текстовых массивов активно используются специализированные технологии *Information Extraction*, *Data Mining* и *Text Mining*. Разработанные на основе статистического и лингвистического анализа, а также методов искусственного интеллекта, указанные технологии предназначены для проведения смыслового анализа, поиска и извлечения различных понятий, фактов, описаний информационных объектов, мнений, отзывов об определенных объектах и т.п. По прогнозам аналитических компаний, спрос на подобные технологии и реализованные на их основе программные продукты существенно возрастет в течение ближайших 4–5 лет.

В данной статье описывается один из возможных подходов к моделированию автоматического извлечения фактов из текстов досье – автобиографий и резюме (заявок на работу). Необходимо отметить, что в основе любого процесса моделирования лежит процедура разработки формальной модели определенного типа. Формальная модель лингвистического явления представляет собой систему правил, описывающих и имитирующих структуру и/или поведение лингвистических объектов и позволяющую хотя бы частично воспроизвести эти объекты либо с помощью человека, либо с помощью компьютера. Для воспроизведения явления с помощью компьютера необходимо составить лингвистическую базу данных/знаний, описывающих определенное лингвистическое явление, и построить, опираясь на нее, алгоритм его функционирования.

Материалом исследования послужили двадцать текстов англоязычных автобиографий и двадцать текстов англоязычных резюме, взятых с разных сайтов сети Интернет. В ходе анализа отобранных текстов были выделены лексические и структурные маркеры, содержащие указание на определенную фактографическую информацию. Выделенные единицы были унифицированы и преобразованы в определенное число лексических и фразовых шаблонов. Поскольку тексты автобиографий и тексты резюме обладают некоторыми лингвистическими отличиями, проанализированные закономерности их организации нашли свое отражение, соответственно, в первой и второй частях лингвистической базы данных системы. С опорой на базу данных были построены два варианта формальной модели, отражающие особенности процесса извлечения фактов из текстов автобиографий и резюме. Несмотря на то, что тексты содержат практически идентичную информацию об их авторе, для извлечения фактов из документов необходимо использовать разные алгоритмы, поскольку, как отмечалось выше, способы представления информации в них не являются полностью аналогичными.

Процесс извлечения фактов из текстов досье заданного типа состоит из следующих этапов: ввода анализируемого текста досье, ввода запроса на поиск и извлечение фактов, просмотра результатов обработки запроса, ввода дополнительных поисковых запросов и просмотр дополнительных резуль-

татов поиска. После завершения процесса ввода текста автобиографии или текста резюме компьютер информирует пользователя о необходимости задания с помощью клавиатуры следующих параметров поиска: *Which information do you need?* (ключевые слова *name, age, birth, marital status, relatives, education, degree, career* для автобиографии и ключевые слова *name, age, birth, marital status, contacts, education, degree, career* для текста резюме). За каждым элементом поискового запроса закреплена определенная последовательность правил поиска фактов в тексте документа. При этом осуществляется поиск строк текста, содержащих соответствующие элементам поискового запроса лексические и фразовые шаблоны. Система поочередно анализирует каждую строку текста. Если в какой-либо строке текста обнаружен заданный шаблон, происходит извлечение конкретной фактографической информации. Такая информация, как имя автора, извлекается один раз, после чего осуществляется автоматический переход к следующему слову поискового запроса. Некоторая информация, например, конкретные факты об учебе и профессиональной деятельности автора текста часто находится в нескольких предложениях, которые могут располагаться в тексте в произвольном порядке. В таких случаях для качественного извлечения фактов необходимо проанализировать весь текст документа. Если все строки текста были проанализированы, но ни один шаблон не был обнаружен, происходит вывод результата поиска типа *Education: no information*. Если результаты поиска и извлечения фактов не соответствуют ожиданиям пользователя, существует возможность ввода дополнительных ключевых слов для организации нового поиска фактов по тексту документа. На данном этапе ключевые слова не закреплены за последовательностью правил, описывающих поиск и извлечение фактов определенного типа. Поэтому поиск строк, которые могут соответствовать введенному ключевому слову или группе слов, осуществляется по специальному алгоритму. По требованию пользователя данная функция может быть выполнена произвольное количество раз.

Одной из общих характеристик принципа поиска фактов в текстах автобиографий и заявок на работу является то, что при указании конкретных шаблонов в ходе составления алгоритма необходимо учитывать пробелы перед словами и после них, так как отдельные слова (особенно служебные части речи) могут совпадать с частями других слов. Допустим, если необходимо извлечь фрагмент строки от предлога *in* до конца строки, необходимо оформить условие, в котором перед предлогом и после него указаны пробелы. В противном случае система извлечет все слова, в которых есть последовательность символов *in*, например, *My name is Jin Flemings, and I was born in...* и т.п.

Для извлечения имени автора текста автобиографии производится поиск фразового шаблона *My name is* и выделение следующей за шаблоном части предложения до первого знака препинания. В большинстве случаев таким знаком препинания является точка. Однако автор текста может привести в одном предложении разнородные факты, вследствие чего искомая информация будет располагаться во фрагменте предложения от шаблона до первой запятой. Поэтому из шаблона, представленного в лингвистической базе

данных, была удалена первая буква (в зависимости от контекста она может быть строчной либо прописной). Для извлечения факта о возрасте автора текста используется лексический шаблон *years old*. С учетом того, что данный шаблон может быть обнаружен в предложениях, содержащих информацию не только об авторе текста (например, *My sister is 11 years old*), релевантной может быть лишь та строка, которая содержит местоимение *I* (*I am 20 years old*). Если возраст выражен числом, достаточно извлечь число, стоящее перед шаблоном. Однако возраст может быть указан словами, причем количество символов в словах заранее предугадать невозможно. В обоих случаях извлекаемая информация расположена между личным местоимением, следующим за ним вспомогательным глаголом и указанным выше шаблоном (*I am twenty years old*). Как правило, возраст автора в тексте автобиографии указан в самом начале документа, а последующие строки с аналогичным шаблоном употреблены для указания возраста ближайших родственников. Если пользователь хочет извлечь факт, касающийся возраста кого-либо из родственников автора текста, необходимо либо указать в поисковом запросе ключевое слово *relatives*, либо воспользоваться функцией дополнительного поиска. В алгоритме предусмотрено дополнительное правило поиска и извлечения факта о возрасте автора. Если перед лексическим шаблоном *years old* не было обнаружено личное местоимение, осуществляется поиск фразового шаблона *My age is* (*My age is 20 years old*). Для извлечения факта о дате и месте рождения автора текста используется лексический шаблон *born*. Искомой информацией является фрагмент строки текста от шаблона до конца предложения (в данном случае последовательность и характер используемых знаков препинания может варьироваться). Так как данный шаблон может быть употреблен в контексте, когда речь идет о других лицах, необходимо найти в этой же строке относящееся к нему личное местоимение *I*. Система не разделяет факты о дате и месте рождения и выдает один конечный результат, например, *Born: in 1991 in Berlin*. Как правило, дата рождения вводится предлогом *on*, а место рождения – предлогом *in*. Однако возможны варианты типа *I was born in 1991*, создающие дополнительные трудности при составлении правил поиска и извлечения фактов. Если необходимо извлечь факт о семейном положении автора, система производит поиск строк текста, содержащих определенные шаблоны. Чтобы убедиться, что они относятся непосредственно к автору текста, осуществляется поиск личного местоимения *I*. О факте *not married* свидетельствуют лексические единицы *not married, divorced, single, widowed*. Если в тексте документа встречается лексический шаблон *married*, необходимо проверить наличие перед ним отрицательной частицы *not*, влияющей на смысл извлекаемого факта. В ряде исследованных текстов их авторы указывают информацию о ближайших родственниках. Невозможно заранее предугадать степень детальности данной информации и значения приведенных в тексте фактов. Поэтому если необходимо найти информацию о родственниках автора, осуществляется поиск лексических шаблонов *mother, father, sister, daughter, wife* и т.д. Факт о получении образования представлен в тексте автобиографии соответствующими периодами. Система определяет

факты, относящиеся к разным этапам обучения, опираясь на лексические шаблоны *school, college, university*. Если возникает потребность узнать информацию о факте *kindergarten*, можно воспользоваться функцией дополнительного поиска фактов. Учебное заведение может быть не только местом учебы, но и местом работы. Для того чтобы отнести учебное заведение к месту учебы, необходимо найти рядом с наименованием учебного заведения такие глаголы (маркеры), как *study, go to, attend, graduate*. Заранее невозможно предугадать, на каком этапе обучения находится автор текста: планирует поступать в учебное заведение, является студентом или выпускником. В зависимости от временного интервала процесса обучения, автор текста может употреблять глаголы в различных временных формах, например, *I study at school/I studied at school*. В связи с этим осуществляется поиск не полной глагольной формы, а ее неизменяемой части. В итоге, данный факт имеет следующий вид – *тип учебного заведения*: строка текста, содержащая маркеры. Для извлечения из текста такого факта, как наличие у автора автобиографии научной степени, в поисковом запросе необходимо указать ключевое слово *degree*. Важная информация может находиться как перед аналогичным лексическим шаблоном, так и после него, причем нередко – одновременно в двух позициях. Для наиболее точного извлечения факта о научной степени строка, содержащая данный шаблон, выделяется из текста целиком. Поиск фактов, связанных с профессиональной деятельностью автора текста, осуществляется на основе лексических шаблонов *company, firm, office, job, position*. Место работы автора текста может также являться учебным заведением (*school, college, university*). Если в строке текста с одним из вышеуказанных слов обнаружена форма глагола *to work* и личное местоимение *I*, значение факта о профессиональной деятельности относится к автору текста. Трудовая деятельность может быть описана несколькими предложениями (в том числе, расположенными дистантно). Для накопления таких предложений используется лексический шаблон *career*. Названия таких шаблонов, как *Professional Experience, Education* и т.д. могут варьироваться.

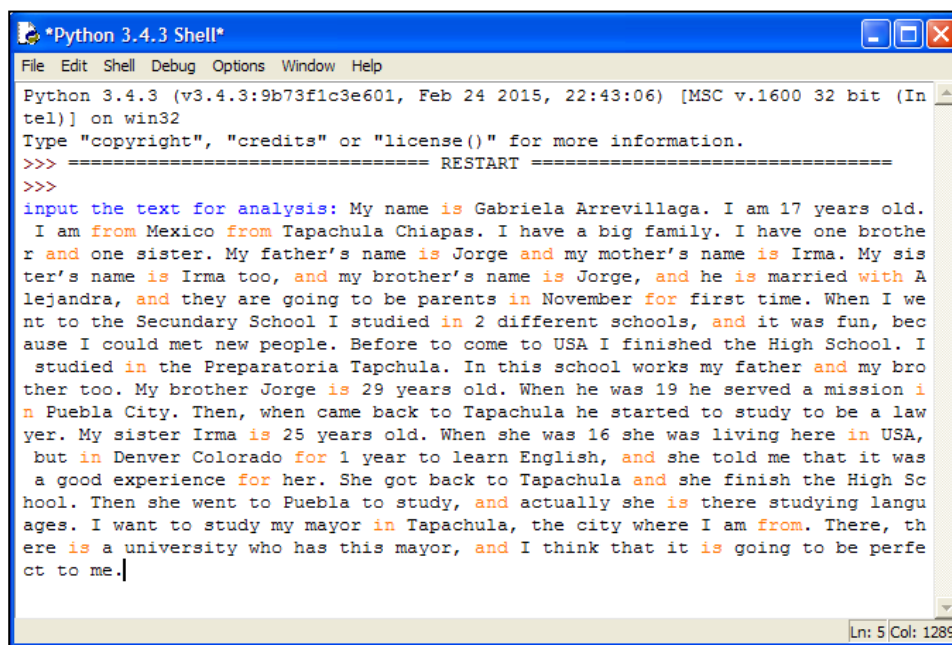
В тексте резюме поиск имени автора документа осуществляется на основе структурного шаблона *Name*, который должен располагаться в начале строки текста, содержащей данную информацию. Такая строка может быть извлечена целиком без изменений. В ряде исследованных текстов шаблон *Name* отсутствует, однако имя автора текста всегда занимает первую строку документа. Контактные данные находятся в строках текста со следующими структурными шаблонами: *Address, Phone, e-mail*. Если шаблон находится в начале строки, достаточно извлечь всю строку. Перед шаблоном *Phone* могут находиться дополнительные лексические шаблоны (*cell/phone, mobile/phone, office/phone*). Подобные шаблоны не являются релевантными при составлении правил извлечения фактов. Поэтому достаточно указать в правиле шаблон *Phone* и при положительном результате поиска извлечь всю строку целиком. Таким образом, дополнительные шаблоны будут сохранены, а правило извлечения факта будет более компактным и универсальным. При извлечении электронного адреса используется дополнительное правило поиска с учетом наличия символа *@* в каждой строке текста. Когда контактная

информация не сопровождается соответствующими структурными шаблонами, а представлена в верхней части документа сплошным текстом, при извлечении фактов возникают определенные трудности. При этом невозможно не только отделить факты друг от друга, но и выделить данный фрагмент текста без использования правил его семантического анализа. В такой ситуации из текста можно извлечь все строки, расположенные после структурного шаблона, следующего за личной и контактной информацией. Поиск фактов о профессиональной деятельности автора заявки на работу осуществляется на основе одного из следующих структурных шаблонов: *Professional Experience*, *Employment History*, *Work Experience*, *Construction Career Summary*, *Relevant Experience*. В шаблоне представлен период работы, наименование организации работодателя, должность автора текста и исполняемые им обязанности. Следовательно, при извлечении факта о карьере автора резюме из текста можно выделить весь фрагмент, соответствующий данному шаблону. Таким образом, чтобы извлечь все факты о профессиональной деятельности автора, необходимо выделить все строки текста, расположенные между шаблоном, указывающим на карьеру, и следующим шаблоном. В большинстве случаев, следующий фрагмент текста содержит факт, указывающий на уровень образования. Факт, отражающий информацию об образовании автора заявки на работу, представлен аналогичным образом. Информация о каждом учебном заведении включает в себя период обучения (полный период обучения либо год окончания учебного заведения), наименование учебного заведения и достигнутые результаты (полученная квалификация, ученая степень и т.д.). В данном случае используются следующие варианты структурного шаблона: *Education*, *Education and Further Training*, *Additional Courses* и т.д. Документ может содержать несколько фактов об образовании, например, факт о получении основного образования и факт о получении дополнительном образовании. Если важным является поиск и извлечение фактов об определенной должности, занимаемой автором, или о конкретном учебном заведении, в котором он обучался, можно воспользоваться функцией дополнительного поиска. Информация о научной степени автора (если такой факт содержится в тексте резюме) с большой вероятностью будет находиться в одной строке с наименованием учебного заведения. Тем не менее, существует возможность извлечь данный факт с опорой на структурный шаблон *Degree* (как и в тексте автобиографии).

Для проведения компьютерного эксперимента на основе двух вариантов формальной модели были написаны два программных кода (язык программирования *Python*). Программные продукты функционируют в командной строке среды *Idle 3.4.3*. Фрагменты результатов работы программ представлены ниже на рис. 1–2.

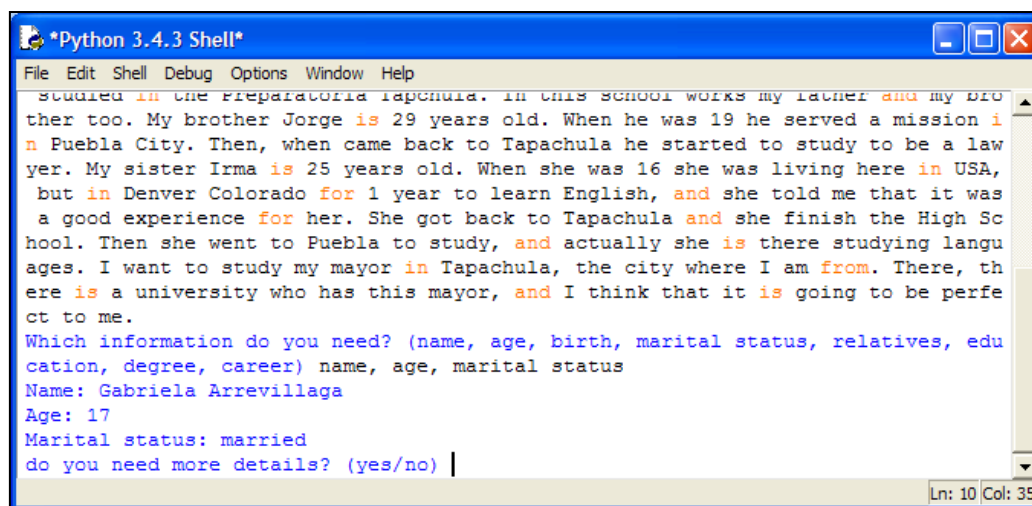
Анализ результатов работы первой программы (автоматическое извлечение фактов из текстов англоязычных автобиографий) показал, что компьютер может успешно извлекать факты из текста на основе лексических и фразовых шаблонов, представленных в первой части лингвистической базы данных. Результаты работы второй программы (автоматическое извлечение фактов из текстов резюме) также можно признать достаточно успешными.

Исключение составляют случаи, когда некоторые факты, в частности, контактная информация, не представлены в тексте лексическими единицами, совпадающими со структурными шаблонами. Компьютер не может разделить подобные факты без детального семантического анализа текста. По окончании компьютерного эксперимента были сделаны следующие выводы о возможностях совершенствования предложенной формальной модели.



```
Python 3.4.3 (v3.4.3:9b73f1c3e601, Feb 24 2015, 22:43:06) [MSC v.1600 32 bit (Intel)] on win32
Type "copyright", "credits" or "license()" for more information.
>>> ===== RESTART =====
>>>
input the text for analysis: My name is Gabriela Arrevillaga. I am 17 years old.
I am from Mexico from Tapachula Chiapas. I have a big family. I have one brother and one sister. My father's name is Jorge and my mother's name is Irma. My sister's name is Irma too, and my brother's name is Jorge, and he is married with Alejandra, and they are going to be parents in November for first time. When I went to the Secondary School I studied in 2 different schools, and it was fun, because I could met new people. Before to come to USA I finished the High School. I studied in the Preparatoria Tapchula. In this school works my father and my brother too. My brother Jorge is 29 years old. When he was 19 he served a mission in Puebla City. Then, when came back to Tapachula he started to study to be a lawyer. My sister Irma is 25 years old. When she was 16 she was living here in USA, but in Denver Colorado for 1 year to learn English, and she told me that it was a good experience for her. She got back to Tapachula and she finish the High School. Then she went to Puebla to study, and actually she is there studying languages. I want to study my mayor in Tapachula, the city where I am from. There, there is a university who has this mayor, and I think that it is going to be perfect to me.
```

Рис. 1. Рабочее окно программы с текстом автобиографии № 1



```
Studied in the Preparatoria Tapchula. In this school works my father and my brother too. My brother Jorge is 29 years old. When he was 19 he served a mission in Puebla City. Then, when came back to Tapachula he started to study to be a lawyer. My sister Irma is 25 years old. When she was 16 she was living here in USA, but in Denver Colorado for 1 year to learn English, and she told me that it was a good experience for her. She got back to Tapachula and she finish the High School. Then she went to Puebla to study, and actually she is there studying languages. I want to study my mayor in Tapachula, the city where I am from. There, there is a university who has this mayor, and I think that it is going to be perfect to me.
Which information do you need? (name, age, birth, marital status, relatives, education, degree, career) name, age, marital status
Name: Gabriela Arrevillaga
Age: 17
Marital status: married
do you need more details? (yes/no) |
```

Рис. 2. Поисковый запрос и результаты извлечения фактов из текста автобиографии № 1

1. Для улучшения качества процесса извлечения фактов необходимо разработать правила поиска дат, а также правила, позволяющие определять, к каким фактам относятся найденные даты.
2. Некоторые модули системы должны осуществлять более детальный поиск и извлечение фактов (например, поиск фактов о дате и месте рождения, образовании, профессиональной деятельности). Так, в рамках фор-

мальной модели факт об учебном заведении, которое окончил автор текста, будет извлечен по образцу: *University: From 1990 till 1994 I studied Physics in the University of Physics*. При использовании дополнительных правил извлечения фактов возможно получение искомого результата следующего вида:

University: University of Physics;

Years of studying: 1990 – 1994;

Major: Physics.

3. Если пользователь работает с большим количеством текстов автобиографий и резюме, для сохранения извлеченных фактов можно создать фактографическую базу данных. В дальнейшем по базе данных можно организовать дополнительный автоматический поиск фактов, ее можно постоянно обновлять. Допустим, авторы двух автобиографий работали или работают в одной организации. Если в одном из текстов указан адрес этой организации, данный факт может быть автоматически скопирован в массив фактов о втором лице. Кроме того, если в базе данных содержатся факты, извлеченные из текста автобиографии определенного лица, и затем происходит поиск фактов в тексте заявки на работу, написанного тем же лицом, обнаруженные в первом случае факты могут быть дополнены фактами, извлеченными из второго документа. Используя базу данных, можно проводить поиск и извлечение фактов не только об определенном лице или лицах, но и о конкретной организации, жителях определенного города и т.д.

Представленная в статье формальная модель может быть использована для поиска и извлечения фактов из текстов досье, связанных с жизнью и деятельностью конкретной личности либо группы лиц; автоматизации документооборота в отделах кадров; формирования и ведения тематических досье; проведения информационной разведки; отслеживания тенденций и различных показателей в обществе и т.д.

А. С. Семак

СТАТИЧЕСКИЙ АСПЕКТ СОДЕРЖАНИЯ ТЕКСТОВ ОЦЕНОЧНЫХ ОТЗЫВОВ НА ПРЕДПРИЯТИЯ ГОСТИНИЧНОГО ТИПА

Большинство лингвистов отмечает, что текст является многоаспектным, многоплановым явлением. К одному из аспектов текста относят «статику» и «динамику». Внешняя форма текста существует в материальном виде как последовательность лексических единиц. В процессе их декодирования они начинают приобретать содержание, которое формируется в сознании в виде мыслительного образования. Таким образом, текст, как результат речемыслительной деятельности, находится в статическом состоянии. Текст в процессе порождения, восприятия и понимания находится в динамическом состоянии [1, с. 31].

Как отмечают исследователи, в процессе восприятия текста сначала воспринимается семантика предложений, а потом их синтаксическая структура [2]. На семантическом уровне текст состоит из высказываний и сверхфра-

зовых единств. На композиционно-семантическом уровне выделяют главы, параграфы, разделы, абзацы и т.д. [3]. Данные элементы являются взаимозависимыми и образуют смысловое единство всего текста. Как показывают лингвистические исследования смысловое единство текста заключается в определенном наборе «ключевых слов» («смысловых вех», «топиков», «ядер» и т.д.) [4]. Очевидно, что исходя из наличия в тексте таких свойств как «целостность» и «связность» можно утверждать, что в каждом тексте присутствуют конкретные объекты и субъекты в определенном, фиксированном отрезке времени и ограниченном количестве мест. Эти компоненты, объективируемые в «ключевых словах», распределены по тексту неравномерно из-за индивидуальной цели автора текста. Это неравномерное распределение формирует иерархию «ключевых слов», которая соотносится с иерархией денотатов [5, с. 79].

Таким образом, для определения смыслового содержания текста требуется определить его «ключевые слова». Существует три группы методов извлечения ключевых слов, словосочетаний или предложений из текста: статистические, позиционные и логико-семантические. Основным критерий статистических методов – частота употребления знаменательных слов. Позиционные методы исходят из предпосылки, что наиболее значимая информация содержится в заглавии, начале и конце текста. Логико-семантические методы используют целый ряд критериев для извлечения наиболее важной информации. Например, смысловой вес слов, наличие семантической связи между предложениями, употребление в тексте слов из названия, наличие глагола в предложении и т.д. Несмотря на количество методов и критериев отбора ключевых компонентов текста, можно отметить, что они ориентированы непосредственно на результат этого процесса, а не на моделирование самого процесса понимания текста человеком.

Несмотря на недостатки вышеуказанных методов, стоит заметить, что они обладают рядом достоинств. Статистический метод отличается высокой точностью результатов, поэтому именно он послужил методом определения «ключевых слов» в нашем исследовании. В качестве материала исследования послужили оценочные отзывы на предприятия гостиничного типа. Под «оценочным отзывом» понимается оформленный тип текста, содержащий авторскую оценку о некоем явлении действительности, размещенный в сети Интернет (или на других ресурсах) с целью обмена информацией между пользователями. Причиной выбора оценочных отзывов для анализа стала активизация лингвистических исследований в области определения тональности. Системы определения тональности широко применяются в сфере мониторинга качества услуг и товаров, бизнес-аналитики, кинопроката и политического прогнозирования. Под определением тональности понимается определение характера субъективной информации, содержащейся в тексте. Тональность может быть положительной, нейтральной и отрицательной. На настоящий момент количество методов определения тональности очень велико, но их принято распределять на четыре группы: методы, основанные на правилах, основанные на словарях, основанные на машинном обучении с учителем и на машинном обучении без учителя. Каждая группа методов

обладает рядом достоинств и недостатков, характерных для различных задач анализа тональности. Стоит заметить, что ни один из вышеуказанных методов не рассматривает оценочный отзыв как связную структуру. Например, в методе, основанном на словарях тональной лексики, происходит сравнение определенных лексических единиц с единицами тонального словаря: если прилагательное, встречающееся в тексте, имеет в таком словаре положительную тональность, то тональность существительного, к которому относится это прилагательное, так же будет положительной. Методы, основанные на машинном обучении, исходят из принципа использования заранее размеченных текстов-образцов, на основе которых происходит обучение компьютерного классификатора. Очевидно, что при таких подходах наличие смысловых связей внутри текста игнорируется во время анализа. Таким образом, для определения тональности посредством формализации семантики текста требуется определить его статику и динамику.

Для последующего анализа нами были использованы отзывы с оценочных Интернет сайтов booking.com, tripadvisor.com. Причиной выбора этих сайтов послужила политика администрации этих ресурсов направленная на предотвращение публикации информации коммерческого содержания. Например, сайт www.booking.com позволяет оставлять отзыв об отеле только после фактической оплаты пребывания в нем.

Целью статистического метода являлось определение основных объектов текста: действующие лица, действия, состояния, места, время и т.д. Статистический анализ показал, что на протяжении всего текста употребляются опорные слова, которые раскрывают тему. Распределение данных слов не является равнозначным. Слова с наибольшим семантическим весом являются *главными опорными словами* (ГОС). Слова, раскрывающие отдельные микроситуации являются *второстепенными опорными словами* (ВОС). Статическое содержание всего текста отражено в последовательности ГОС и ВОС [5].

В качестве основной логико-семантической единицы текста в нашем случае выступает абзац, т.к. именно в предметно-логическом содержании абзаца текста раскрывается его микротема. Под предметно-логическим содержанием абзаца понимается перечисление фактов, объектов действительности так или иначе связанных с сообщением текста. В независимости от жанра текста, абзац строится по определенным правилам, которые отражают структурно-смысловые модели изложения мыслей [6, с. 335]. Данные модели усваиваются человеком еще в раннем детстве и хранятся в памяти в виде языковых шаблонов [7]. Они представляют собой элементы психической реальности, которые напрямую участвуют в порождении высказывания. Причем, синтаксические шаблоны могут охватывать больше одного предложения и формировать шаблон-высказывание. Количество и специфичность таких шаблонов определяется текстами одной предметной области или одного автора. Совокупность таких шаблонов-высказываний в лингвистике называется «сложным синтаксическим целым» или «сверхфразовым единством». Но из-за отсутствия формальных признаков таких составляющих подвергнуть компьютерной обработке сверхфразовое единство и сложное синтаксическое целое чрезвычайно трудно.

Исходя из вышесказанного, можно сделать вывод, что абзац является отражением фрагмента психической ситуации и формирует смысловое и логическое единство текста, что подтверждается рядом психолингвистических экспериментов [8, 9]. В ходе данных экспериментов было установлено, что членение текстов на абзацы происходит не автоматически, а носит целенаправленный характер. Так же абзац исключительно удобен для формализации из-за наличия формального признака (красной строки). Таким образом, текст можно представить в виде последовательного изложения его абзацев.

На начальном этапе анализа отзывов был составлен алфавитно-частотный словарь каждого текста с указанием общего количества словоупотреблений и количество абзацев с данным словом. Далее алфавитно-частотного словаря были удалены все служебные и общеупотребительные слова. Например, были удалены предлоги, союзы, артикли, вспомогательные и модальные глаголы и числительные. Показатели были скорректированы с учетом объединения словоформ, лексических повторов и различных видов синонимий. На последнем этапе обработки алфавитно-частотного словаря были удалены все словоформы, употребленные только в одном абзаце ($m = 1$), т.к. они относятся к содержанию только одного абзаца, а не всего текста. В итоге был получен словарь потенциальных опорных слов. Дальнейшее распределение на ГОС и ВОС осуществлялось по формуле:

$$K_{\text{важ}} = \frac{(F * m)}{(N * n)},$$

где:

F – абсолютная частота слова в тексте;

m – число абзацев, в которых встретилось слово;

N – общее число слов в тексте;

n – общее число абзацев в тексте.

Извлечение ГОС и ВОС происходит на основе сравнения коэффициента важности с граничными значениями $K_{1\text{важ}}$ и $K_{2\text{важ}}$, которые вычисляются для каждого текста исходя из количества абзацев и общего числа словоформ.

Например, в одном¹ из анализируемых оценочных отзывах имеется 6 абзацев и 359 слов. В ходе вышеописанного анализа было выделено 4 ГОС (табл. 1) и 10 ВОС (табл. 2).

Т а б л и ц а 1

Главные опорные слова оценочного отзыва

№ слова	ГОС	Общее количество словоупотреблений	Общее количество абзацев, где встретилось слово
1	We	13	4
2	Breakfast	7	4
3	Room	6	4
4	Hotel	5	4

¹ Отзыв размещен на сайте: <http://www.tripadvisor.com>

Т а б л и ц а 2

Второстепенные опорные слова оценочного отзыва

№ слова	ВОС	Общее количество словоупотреблений	Общее количество абзацев, где встретилось слово
1	Night	3	2
2	Pounds	3	2
3	Check-in	2	2
4	Could	2	2
5	Got	2	2
6	Morning	2	2
7	Took	2	2
8	Usual	2	2
9	Very	2	2
10	Went	2	2

Далее в каждом абзаце текста анализировалась позиция ГОС и ВОС и то, как они раскрывают тему текста. Таким образом, предметно-логическое содержание данного отзыва можно представить в виде цепочки микротем (табл. 3). Каждой микротеме был присвоен код. Всего, исходя из результатов подобного анализа сорока оценочных отзывов, было выявлено 26 типов микротем.

Т а б л и ц а 3

Предметно-логическое содержание абзацев оценочного отзыва

№ Пункта	Код микротемы	Предметно-логическое содержание абзацев текста
1	M001	Описание процесса бронирования ГОС 3 в ГОС 4
2	M002	Оценочная характеристика ГОС 4
3	M003	Оценочная характеристика ГОС 3 в ГОС 5
4	M004	Общая оценочная характеристика ГОС 3
5	M005	Описание процесса выписки ГОС 1 из ГОС 5
6	M006	Итоговая характеристика ГОС 5 и ГОС 4

Полученные данные отражают статический аспект текстов оценочных отзывов и необходимы для дальнейшей формализации логико-семантических отношений.

ЛИТЕРАТУРА

1. Новиков, А. И. Семантика текста и ее формализация / А. И. Новиков. – М. : Наука, 1983. – 215 с.
2. Леонтьев, А. А. Психолингвистические единицы и порождение речевого высказывания / А. А. Леонтьев. – М. : Наука, 1969. – 307 с.
3. Валгина, Н. С. Теория текста / Н. С. Валгина. – М. : Логос, 2003 – 278 с.
4. Вейнрих, У. Опыт семантической теории / У. Вейнрих // Новое в зарубежной лингвистике. – М., 1981. – Вып. 10. – С. 50–176.
5. Зубов, А. В. Статистический аспект содержания текста и его формальное выражение / А. В. Зубов // Квантитативная лингвистика и автоматический анализ текста : учен. зап. Тарт. ун-та. – Тарту, 1986. – Вып. 745. – С. 75–91.

6. *Зубов, А. В.* Вероятностно-алгоритмическое моделирование статической и динамической составляющих содержания текста / А. В. Зубов // OSTIS 2011: матер. межд. науч.-технич. конф., Минск, 10–12 февр. 2011 г. / редкол. : В. В. Голенков (отв. ред.) [и др.]. – Минск : БГУИР, 2011. – С. 331–342.
7. *Адмони, В. Г.* Синтаксическая семантика – это семантика синтаксических структур / В. Г. Адмони // Проблемы синтаксической семантики. – М., 1976. – С. 3–8.
8. *Страхова, В. С.* Внешние средства организации текста / В. С. Страхова // Лингвистика текста : сб. науч. тр. МГПИИЯ. – М., 1971. – Вып. 141. – С. 156.
9. *Бондаренко, Г. В.* Текст – смысл – ситуация (к постановке проблемы) / Г. В. Бондаренко, Ю. А. Шрейдер // Вопросы информационной теории и практики – М., 1978. – Вып. 36. – С. 80–91.

М. А. Шумская

АВТОМАТИЧЕСКОЕ ФОРМИРОВАНИЕ ОТВЕТА НА ВОПРОС ПОЛЬЗОВАТЕЛЯ КАК БАЗОВАЯ ФУНКЦИЯ ВОПРОСНО-ОТВЕТНОЙ ПОИСКОВОЙ СИСТЕМЫ

В последние годы в рамках технологий *Text Mining* развивается новое направление, связанное с разработкой интеллектуальных вопросно-ответных поисковых систем. Основные отличия такого рода системы от классической поисковой системы заключаются в ее умении обрабатывать запросы пользователя, сформулированные в виде вопросительных предложений на естественном языке, а также представлять результаты поиска не в виде ранжированных адресов сайтов, релевантных запросу, а в виде конкретного текстового фрагмента, содержащего нужную пользователю информацию.

Автором данной статьи была создана формальная модель англоязычной вопросно-ответной поисковой системы. Материалом исследования послужили тексты 19-ти научно-популярных статей на английском языке, а также 45 специальных вопросов по их содержанию, сформулированные жителями разных стран, для которых английский язык является родным, вторым или иностранным. Первая часть формальной модели (анализатор запросов) отвечает за анализ запроса пользователя, представленного в виде специального вопроса, начинающегося с вопросительного слова *When* или *Where*. Вторая часть формальной модели (синтезатор ответов), опираясь на результаты поиска, синтезирует конкретный ответ на поставленный вопрос. Рассмотрим подробнее процедуру автоматического формирования ответа на вопрос пользователя.

Работа синтезатора ответов состоит из двух основных этапов. На первом этапе во всем массиве найденных документов система должна корректно определить нужный текст (или несколько текстов), который будет соответствовать тематике и смыслу вопроса пользователя. Затем, из набора предложений текста (текстов) ей нужно выбрать предложения-кандидаты, содержащие необходимую пользователю информацию. Второй этап процесса синтеза ответа включает в себя проведение синтаксического анализа предложений-кандидатов, извлечение из них необходимой информации (исключительно на

основе синтаксического анализа), и непосредственно синтез ответа на основе определенной синтаксической структуры и данных, полученных в первой части формальной модели на этапе синтаксического анализа вопроса.

В целом лингвистическая база данных синтезатора ответов, как и база данных анализатора запросов, включает несколько компонентов.

1. Список синтаксических структур вопросов и соответствующих им структур ответов, представленный в табл. 1.

Т а б л и ц а 1

Список структур вопросов и соответствующих им структур ответов

Структура предложения-запроса	Структура предложения-ответа
Aux.v Pred = be	
WRB/PL aux.v Subj MV	Subj aux.v MV #MPL
WRB/PL Pred Subj	Subj Pred #MPL
Aux.v = do	
WRB/PL aux.v Subj Pred	Subj Pred #MPL
WRB/PL aux.v Subj Adv. Pred	Subj Adv. Pred #MPL
Aux.v = will can	
WRB/TM aux.v Subj Pred	Subj aux.v Pred #MTM
WRB/TM aux.v Subj Pred D.Obj	Subj aux.v Pred D.Obj #MTM

2. Алфавитный тегированный словарь (состоящий из слов предложений-кандидатов в ответы), составленный с опорой на демо-версию POS-тегера Иллинойского Университета, США. Фрагмент словаря представлен в табл. 2.

Т а б л и ц а 2

Фрагмент алфавитного тегированного словаря синтезатора ответа

's/POS	church/NN	is/VBZ	Serbian/JJ
1966/CD	citizen/NN	island/NN	sides/NNS
3600/CD	clinic/NN	its/PRP\$	somewhere/RB
7/CD	construction/NN	Jimmy/NNP	stage/NN
a/DT	Croatia/NNP	man/NN	temples/NNS
Alabama/NNP	dozen/NN	Maryland/NNP	Tesla/NNP
among/IN	emigrated/VBD	Nash/NNP	the/DT
.....			

3. Список слов, формирующих модуль времени и модуль места. Фрагмент списка представлен в табл. 3.

Т а б л и ц а 3

Фрагмент списка слов, входящих в состав модулей времени и места

Модуль времени	Модуль места
<i>about</i> 5,000 light-years	<i>at</i> St. Peter's Basilica
<i>around</i> 3600 B.C.	<i>in</i> Barcelona
<i>by</i> spring 2016	<i>in</i> Bethesda, Maryland
.....	

4. Список контекстуальных синонимов, который формируется для каждого текста. Фрагмент списка представлен в табл. 4.

Т а б л и ц а 4

Фрагмент списка контекстуальных синонимов

inhabit	live
reign	rule
Kepler-78	Planet, it
Tesla	Nikola Tesla, he
Wales	Jimmy Wales, he
.....	

5. Список последовательностей тегов, которые соответствуют словам, входящим в именные группы (*noun phrase, NP*). Фрагмент списка представлен в таблице 5

Т а б л и ц а 5

Фрагмент списка последовательностей тегов слов именных групп

NN	DT NNP NNP
NNS	DT NNP NNS
NNP	JJ JJ NNS
NNPS	NNP DT NNP
DT NNS	DT NN IN NNPS
JJ NNP	DT NN NN NN
NN NNP	DT NN POS NNS
NN NNS	DT JJ NN IN NN
.....	

6. Список форм глаголов, фрагментарно представленный в табл. 6.

Т а б л и ц а 6

Фрагмент списка форм глаголов

VB	VBZ (для ед. ч.)	VBP (для мн. ч.)	VBD	VCN	VBG
<i>arrive, conclude, consecrate, die, grace, live, locate, reserve, rule</i>	VB + s	VB	VB + d	VB + d	VB-e + ing
<i>bomb, crash, finish, vanish</i>	VB + es		VB + ed	VB + ed	VB + ing
<i>crown, discover, exist, found, found, happen, inhabit, reign</i>	VB+''s''				
<i>be</i>	is	are	was – ед. ч. were – мн. ч.	been	being
.....					

Правила приписывания синтаксических функций словам в зависимости от их принадлежности к различным частям речи и позиции в предложении, а также правила согласования членов предложений прописаны, соответственно в алгоритме анализатора запросов и синтезатора ответов.

Рассмотрим подробнее компоненты лингвистической базы данных синтезатора ответов. Тегированные словари составлялись на основе слов текстов статей и вопросительных запросов с применением демо-версии тегера, т.е. фактически каждый текст проходил POS-обработку, а затем на основе полученного словаря проводился анализ вопросительных предложений. Однако промышленной системе, способной обслуживать множество пользователей с опорой на огромный текстовый массив, был бы необходим единый тегированный словарь, содержащий многие тысячи слов и специальных символов, которые могут идентифицироваться как токены (токен – один символ или последовательность символов от пробела до пробела). В основном токенами являются слова естественного языка, однако в него может входить, например, и апостроф. Так, некоторые системы автоматического POS-тегирования текстов распознают слово *Mike's* как единый токен, другие (как и использованный нами тегер) делят слово на два токена: *Mike* и *'s*. Именно поэтому представленный в таблице 2 фрагмент словаря содержит не только слова и словоформы, но и специальные символы и цифры. Из данных таблицы 3 хорошо видна определенная закономерность в образовании словосочетаний, выполняющих синтаксическую функцию модуля времени и места, что объясняется аналитической природой английского языка. Следовательно, представляется возможным написание правил, по которым эти модули будет определяться автоматически. Сложнее обстоит дело с формированием списка контекстуальных синонимов, поскольку часто слова имеют схожее значение лишь в рамках определенного контекста. Тем не менее, при анализе больших текстовых массивов можно выделить некоторый набор самых часто встречающихся вариантов контекстуальных заменителей и составить из них базовый список, который будет дополнять личный список каждого отдельного текста. Подобные уникальные списки синонимов формируются при первом анализе текста системой и хранятся вместе с поисковым образом каждого документа.

Важным компонентом синтезатора ответа являются правила распознавания и приписывания словам синтаксических функций. В отличие от количества правил, содержащихся в анализаторе запросов (вопросительных предложений), правила синтезатора ответов не так многообразны, поскольку в данном случае нет необходимости проводить полный и глубокий разбор предложений-кандидатов на ответ. Например, из предложения-кандидата ответа на вопрос *When was Jimmy Wales born?* компьютеру необходимо извлечь только конкретную фактографическую информацию, в то время как субъект и предикат ситуации известны из самого вопроса. Таким образом, протегировав предложение-кандидат и включив в него дополнительные теги модулей времени и места, компьютер формирует следующую запись:

**Jimmy/NNP Wales/NNP was/VBD born/VBN #MPLin/IN
Huntsville/NNP , Alabama/NNP#MPL , #MTMon/IN August/NNP 7/CD ,
1966/CD#MTM .**

Из примера хорошо видно, что при большом, или практически полном соответствии слов запроса пользователя и слов предложения текста глубокий синтаксический анализ с полным определением синтаксических функций

является излишним. Еще один заслуживающий внимания вопрос связан с длиной каждого модуля. Как показано на примере и в таблице 3, иногда слова с одинаковыми тегами могут быть объединены в один более крупный блок.

Не менее важным компонентом синтезатора ответов является набор правил согласования членов предложения. Поскольку система уже владеет достаточной информацией для синтеза ответа (сюда можно отнести данные, полученные в ходе анализа вопросов, а также структуры ответов), то набор правила согласования будет небольшим. В самом общем виде они выглядят следующим образом.

- Если подлежащее, выраженное нарицательным существительным или именем собственным, имеет тег NN или NNP (только для существительных не входящих в список исключений),

- то стоящий за ним глагол с тегом VB должен быть изменен на VBZ, в соответствии с таблицей 6.

- Если подлежащее, выраженное нарицательным существительным или именем собственным, имеет тег NNS или NNPS,

- то стоящий за ним глагол с тегом VB должен быть изменен на VBP, в соответствии с таблицей 6.

- Если подлежащее, выражено нарицательным существительным или именем собственным, а глагол с тегом VBD не является формой глагола *be*,

- то глагол с тегом VBD остается неизменным, в соответствии с таблицей 6.

- Если подлежащее, выраженное нарицательным существительным или именем собственным, имеет тег NN или NNP, а следующий за ним глагол – форма глагола *be* с тегом VBD,

- то глагол *be* принимает форму *was*.

- Если подлежащее, выраженное нарицательным существительным или именем собственным, имеет тег NNS или NNPS, а следующий за ним глагол – форма глагола *be* с тегом VBD,

- то глагол *be* принимает форму *were*.

- Если в вопросительном предложении предикат состоит из вспомогательного глагола *do* и смыслового глагола с тегом VB,

- то в ответе глагол с тегом VB принимает время глагола *do* и число, согласованное с предшествующим существительным.

- Если в вопросительном предложении предикат состоит из вспомогательного глагола *be* и смыслового глагола с тегом VBN,

- то в ответе предикат будет состоять из формы глагола *be*, согласованной в числе с предшествующим существительным, но в сохраненном из вопроса времени, а глагол с тегом VBN останется неизменным.

- Если в вопросительном предложении предикат состоял из смыслового глагола *be*,

- то в ответе предикат будет состоять из формы глагола *be*, согласованной в числе с предшествующим существительным, но в сохраненном из вопроса времени.

Приведенные выше правила составляют основу алгоритма работы синтезатора ответов. Рассмотрим его подробнее. Работа синтезатора ответа начинается с поиска текстов, а затем входящих в них предложений, которые содержат слова из поискового образа запроса (слова из *Блока Б* анализатора запроса пользователя). На выходе получается одно либо несколько наиболее подходящих для синтеза ответа предложений-кандидатов. В том случае, если предложений-кандидатов найдено не было, с помощью контекстуального словаря и правил нормализации производится изменение поискового образа запроса, который отправляется в систему для повторного поиска.

На следующем этапе с опорой на алфавитный тегированный словарь словам предложений-кандидатов приписываются теги частей речи. После этого найденные предложения проверяются на наличие в них модулей времени и места, а затем на содержание искомого тега модуля, который был получен на этапе анализа вопросительного предложения (запроса пользователя) и хранится в *Блоке Б*. Таким образом в системе остается лишь одно, самое вероятное предложение-кандидат. С этого момента начинается непосредственно синтез ответа пользователю. После того как была получена синтаксическая информация, содержащаяся в предложении-кандидате, компьютер находит в базе данных структуру из *Блока А* (анализатор запроса пользователя) и структуру ответа, которая ей соответствует. После этого структура ответа заполняется подходящими словами. Чтобы исключить вероятность ошибки при синтезе ответа, компьютер проводит проверку на наличие двух одинаковых рядом стоящих слов.

Заключительным этапом синтеза ответа на вопрос пользователя является согласование слов в предложении с опорой на хранящиеся в системе правила. Это позволяет сделать ответ грамматически более точным, непохожим на простой набор слов, что отражает определенную степень интеллектуальности вопросно-ответной поисковой системы.

A. Bobkov, S. Gafurov, V. Krasnoproshin, H. Vissia

SEMANTIC-BASED INFORMATION EXTRACTION IN BUZZTALK

Information extraction is gaining much popularity [1; 2]. The field of information extraction is well suited to various types of business and government intelligence applications. Diverse information is of great importance for decision making on products, services, events, persons, organizations.

Creation of systems that can effectively extract meaningful information requires overcoming a number of challenges: identification of documents, knowledge domains, specific opinions, opinion holders, events, activities, as well as representation of the obtained results.

The purpose of this paper is to introduce a system for solving the problem of effective extraction of meaningful information. Semantic patterns approach and an ontology-based approach are proposed as a solution to the problem.

Numerous models and algorithms are proposed for information extraction [3]. The major part of the most useful information is represented as a great variety of texts and documents in a natural language. But the problem of effective information extraction from texts in a natural language still remains unsolved. Processing of texts in a natural language necessitates the solution of the problem of extracting meaningful information. Semantic relations play a major role [4].

In information extraction and text mining, word collocations show a great potential to be useful in many applications (machine translation, natural language processing, lexicography, etc.).

"Collocations" are usually described as "sequences of lexical items which habitually co-occur, but which are nonetheless fully transparent in the sense that each lexical constituent is also a semantic constituent" [5]. Such expressions as "fine weather", "high winds", "cosmetic surgery" are examples of collocations.

The traditional method of performing automatic collocation extraction is to find a formula based on the statistical quantities of words to calculate a score associated to each word pair. Proposed formulas are mainly: "mutual information", "t-test", "z test", "chi-squared test" and "likelihood ratio" [6].

Word collocations from the point of semantic constituents have not yet been widely studied and used for extracting meaningful information, especially when processing texts in a natural language.

The proposed semantic patterns approach is based on word collocations on the semantic level and contextual relations. In general, a semantic pattern includes: 1) *participants* (a person, company, natural/manufactured object, as well as a more abstract entity, such as a plan, policy, etc.) involved in the action or being evaluated; 2) *actions* - a set of verb semantic groups and verbal nouns; 3) *rules for semantic patterns actualization*.

An ontology-based approach is used for semantic patterns actualization [7].

Ontologies have become common on the World-Wide Web ([8]). Ontologies on the Web range from large taxonomies categorizing Web sites (such as on Yahoo!) to categorizations of products for sale and their features (such as on Amazon.com). For any given knowledge domain, the ontology represents the concepts which are held in common by the participants in a particular domain.

Since ontologies explicitly represent knowledge domain semantics (terms in the domain and relations among them), they can be effectively used in solving information extraction problems, word sense disambiguation in particular.

The proposed approach has been successfully realized in BuzzTalk portal ([9]) for subject domains recognition, opinion mining, mood state detection, event extraction and economic activities detection.

BuzzTalk is offered to companies as a SaaS model (Software as a Service) and it answers questions like:

- What is the latest information about my brand, competitors or industry?
- What are people writing about my brand, product, organization or CEO?
- What are important trends in my industry?
- What are the big events inside my industry sector?
- Where are my customers/clients located?
- When and where are people discussing my brand?

What are the burning world and local issues?

Who is involved in burning issues?

What are the consequences?

BuzzTalk presents a new way of finding content.

The difference between a traditional search engine and a discovery engine such as BuzzTalk, is that search engines list all results for a specific search whereas our engine allows you to monitor topic-specific developments within your search.

BuzzTalk collects all text documents from over 58.000 of the most active websites around the globe, two thirds are news sites and one third are blog sites. The authors of these documents are mainly scientists, journalists and opinion leaders.

BuzzTalk presents a list of articles in chronological order based on publication date. This list grows each day. You can sort and filter this list based on a variety of criteria such as sentiment, mood state, happenings, etc., thus to experience the wealth of real time information without the pain of information overload. For example, you can easily find all publications within your theme that relate to product releases, employment changes, merger & acquisitions and many more.

Below are examples of information extraction in BuzzTalk.

Opinion mining is gaining much popularity within natural language processing [10]. Web reviews, blogs and public articles provide the most essential information for opinion mining. This information is of great importance for decision making on products, services, persons, events, organizations.

The proposed ontology-based approach for semantic patterns actualization was realized in the developed knowledge base, which contains opinion words expressing:

- 1) personal emotional state (e.g. happy, delighted, proud, sad, angry, horrified);
- 2) appreciation (e.g. flexible, efficient, stable, reduced, ideal, backward, poor, highest);
- 3) judgement (e.g. active, decisive, caring, dedicated, intelligent, negligent, evil).

Opinion words can be expressed by: an adjective (*brilliant, reliable*); a verb (*like, love, hate, blame*); a noun (*garbage, triumph, catastrophe*); a phrase (*easy to use, simple to use*). Adjectives derive almost all disambiguating information from the nouns they modify, and nouns are best disambiguated by directly adjacent adjectives or nouns.

Information about the force of evaluation (low, high, the highest) and orientation (positive/negative) is also included in the knowledge base. For example, *safe* (low force, positive orientation), *safer* (high force, positive orientation), *the safest* (the highest force, positive orientation), *unsafe* (low force, negative orientation).

In the knowledge base opinion words go together with their accompanying words, thus forming “opinion collocations” (e.g. *deep depression, deep devotion, warm greetings, discuss calmly, beautifully furnished*). By an “opinion collocation” we understand a combination of an opinion word and accompanying words, which commonly occur together in an opinion-oriented text. The use of opinion collocations is a way to solve the problem of opinion word sense disambiguation (e.g. *well-balanced political leader* and *well-balanced wheel*) and to exclude words that do not relate to opinions (cf. *attractive idea* and *attractive energy*).

We assume that the number of opinion collocations, which can be listed in a knowledge base, is fixed.

The use of opinion collocations within the ontology-based approach opens a possibility to assign names of knowledge domains to them, because opinion collocations are generally domain specific. For example, *helpful medical staff* (“health care”), *helpful hotel reception staff* (“travel-hotel”), *stable economy* (“economics”), *well-balanced politician* (“politics”). More than one knowledge domain may be assigned to an opinion collocation, e.g. *fast service* (“economics-company”, “travel-hotel”).

Processing of the extracted opinion collocations is carried out in their contextual environment. The developed algorithm checks for the presence of modifiers that can change the force of evaluation and orientation indicated in the knowledge base.

The developed knowledge base also provides additional information about quality characteristics and relationships for different objects on which an opinion is expressed (e.g. *software product* evaluation includes: usability, reliability, efficiency, reusability, maintainability, portability, testability; *travel-hotel* evaluation includes: value, rooms, location, cleanliness, check in/front desk, service).

The results of opinion collocations processing are grouped and evaluated to recognize the quality of the opinion-related text. The results are also visualized.

A valuable addition to opinion mining is detection of individual/public mood states. The relationship between mood states and different human activities has proven a popular area of research [11].

BuzzTalk mood detection uses the classification of the widely-accepted “Profile of Mood States” (POMS), originally developed by McNair, Lorr and Droppleman [12].

In BuzzTalk, mood state detection is based on: 1) mood indicators (e.g. “I feel”, “makes me feel”, etc.); 2) mood words (e.g. anger, fury, horrified, tired, taken aback, depressed, optimistic); 3) special contextual rules to avoid ambiguation. BuzzTalk automatically recognizes the following mood states: “Anger”, “Tension”, “Fatigue”, “Confusion”, “Depression”, “Vigor”.

Mood state detection alongside with opinion mining can give answers to where we are now and where will be in future.

BuzzTalk detects 233 economic activities from texts in a natural language. The economic activities cover all major activities represented in NACE classification (Statistical Classification of Economic Activities in the European Community), which is similar to the Standard Industrial Classification and North American Industry Classification System.

An example of economic activities detection

General manager Tim Deakin said Orkney Cheddar was produced with locally-sourced milk, following a traditional recipe and process.

Extracted instances:

Economic activity = **Manufacture of Dairy Products**.

BuzzTalk performs real-time extraction of 35 events for decision making in different spheres of business, legal and social activities. The events include: "Environmental Issues", "Natural Disaster", "Health Issues", "Energy Issues",

"Merger & Acquisition", "Company Reorganization", "Competitive Product/Company", "Money Market", "Product Release", "Bankruptcy", "Bribery & Corruption", "Fraud & Forgery", "Treason", "Hijacking", "Illegal Business", "Sex Abuse", "Conflict", "Conflict Resolution", "Social Life", etc.

For example

A New York woman faced charges for faking cancer to solicit money from unsuspecting donors and a relative.

Extracted instances:

Event = **Fraud & Forgery**

A subject domain is recognized on the basis of a particular set of noun and verb phrases unambiguously describing the domain. For solving the problem of disambiguation special filters, based on the contextual environment (on the level of phrases and the whole text), are introduced.

Subject domains and their concepts are organized hierarchically to state "part-of", "is a kind of" relations.

The proposed semantic patterns approach has been successfully realized in BuzzTalk portal for opinion mining, mood state detection, event extraction and economic activities detection. The approach ensures high accuracy, flexibility for customization and future diverse applications for information extraction.

Semantic word collocations are a major factor in the development of a wide variety of applications including information extraction and information management (retrieval, clustering, categorization, etc.).

Implementation results show that the proposed knowledge-based approach is correct and justified and the technique is highly effective.

REFERENCES

1. Moens, M. Information Extraction: Algorithms and Prospects in a Retrieval Context / M. Moens. – Springer, 2006. – 246 p.
2. Baeza-Yates, R. Modern Information Retrieval: The Concepts and Technology behind Search / R. Baeza-Yates, B. Ribeiro-Neto. – Addison-Wesley Professional, 2011. – 944 p.
3. Buettcher, S. Information Retrieval: Implementing and Evaluating Search Engines / S. Buettcher, C. Clarke, G. Cormack. – MIT Press, 2010. – 632 p.
4. Khoo, Ch. Identifying Semantic Relations in Text for Information Retrieval and Information Extraction / Ch. Khoo, H. Myaeng S. – Springer, 2002. – P. 161–180.
5. Cruse, D. A. Lexical Semantics / D. A. Cruse. – Cambridge University Press, 1986. – 310 p.
6. Manning, C. D. Foundations of statistical natural language processing / C. D. Manning, H. Schütze. – Cambridge, MA : MIT Press, 1999. – 620 p.
7. Bilan, V. An Ontology-Based Approach to Opinion Mining / V. Bilan, A. Bobkov, S. Gafurov, V. V. Krasnoproshin, J. van de Laar, H. Vissia // Proceedings of 10-th International Conference PRIP'2009. – Minsk, 2009. – P. 257–259.
8. Fensel, D. Foundations for the Web of Information and Services: A Review of 20 Years of Semantic Web Research / D. Fensel. – Springer, 2011. – 416 p.
9. <http://www.buzztalkmonitor.com>.
10. Pang, B. Opinion Mining and Sentiment Analysis / B. Pang, L. Lee. – Now Publishers Inc, 2008. – 148 p.
11. Clark, A. V. Mood State and Health / A. V. Clark. – Nova Publishers, 2005. – 213 p.
12. McNair, D. M. Profile of Mood States / D. M. McNair, M. Lorr, L. F. Droppleman. – San Diego, Calif. : Educational and Industrial Testing Service, 1971.

АВТОМАТИЧЕСКИЙ АНАЛИЗ И СИНТЕЗ ТЕКСТОВ И РЕЧИ

Т. В. Аникеева, И. И. Панова

КОНСТАНТНЫЕ И ВАРИАТИВНЫЕ ПРИЗНАКИ В ПРОСОДИИ АНГЛИЙСКОЙ ПУБЛИЧНОЙ РЕЧИ ПРЕДСТАВИТЕЛЕЙ РАЗЛИЧНЫХ ПРОФЕССИЙ

Сопоставительное рассмотрение содержательных и классифицирующих подходов к толкованию дискурсивной сущности публичной речи убеждает в правомерности принятия ее главных интегральных признаков, влияющих на языковую специфику ее формообразования, таких как адресатная направленность, монологичность, социальная ориентированность, коммуникативная и прагматическая заданность. Являясь продуктом совместного творчества говорящего и аудитории, т.е. такой формой, в которой реализуется особая интерактивная и прагматически ориентированная речевая деятельность индивидуума в официально-деловой сфере общения [1], публичная речь социально и культурно обусловлена, и ее структурирование, следовательно, регламентировано правилами и предписаниями ораторской традиции [2].

Несмотря на значительное количество лингвистических исследований публичной речи, предметом изысканий которых, чаще всего, являются отдельные языковые аспекты ее описания [3; 4; 5], комплексное толкование характера взаимодействия языковых средств различных уровней в процессе реализации ее наиболее важных коммуникативно-функциональных и эмоционально-модальных особенностей до сих пор представляет научный интерес. В то же время, публичная речь как социальное и культурное явление речевой действительности с необходимостью предполагает определение специфики ее композиции, содержательности и базовых функций в системе речевых актов. Одной из наименее изученных областей публичной речи также являются объем и границы ее значимой вариативности в зависимости от социолингвистических факторов, и, в том числе, обусловленных социогрупповой и профессиональной практикой речетворчества носителей языка.

В данной статье излагаются результаты экспериментально-фонетического изучения характера контрастивного расслоения просодических структур фраз, составляющих английские публичные сообщения представителей различных профессий, которые являются достоверными индикаторами их воспринимаемой профессиональной маркированности.

Результаты исследования основаны на сопоставительном экспериментально-фонетическом анализе аутентичных текстов в естественных условиях их публичной реализации английскими политиками, юристами, учителями и актерами.

При отборе образцов устных текстов важным условием являлась необходимость их соответствия требованиям однотипных экстралингвистических факторов их представления, а именно:

1) выступление перед большой аудиторией вне профессионального сообщества каждого испытуемого;

2) отсутствие профессиональной терминологии как потенциального сигнала их профессиональной принадлежности;

3) социально-популярная тема выступления, максимально дистанцированная от профессиональной занятости говорящего.

Исследуемые отрывки текстов идентичной временной ограниченности в количестве 20 единиц были вычленены из аутентичных записей выступлений носителей британской орфоэпической нормы (Received Pronunciation-RP). Длительность каждого отрывка составляла одну минуту звучания; тематика сообщений касалась социально значимой на определенном этапе проблематики в жизни британского общества.

Экспериментально-фонетический анализ проводился в три этапа. На первом этапе анализа была выполнена идентификация профессий говорящих по предъявляемым устным текстам без опоры на письменные носители тремя группами аудиторов: носителями английского языка, неносителями языка, владеющими английским языком на высоком уровне и немецкими гражданами, не владеющими английским языком. Усредненный процент идентификации профессий говорящих в каждой указанной группе составлял 85 %, 78 % и 65 % соответственно каждой указанной группе.

На втором этапе проводился аудитивный анализ аудиторами-фонетистами по признакам членения текстов на просодические единицы (интонационные группы и фразы), паузальной сегментации текстов, дистрибуции фразовых ударений различной степени, высотно-уровневых характеристик фраз и тонального завершения финальных и нефинальных интонационных групп, которые были зарегистрированы ими в просодической транскрипции.

Третий этап включал акустический анализ по параметру ЧОТ (частоты основного тона), измеряемый в герцах (Гц): максимальный уровень первого ударного слога во фразах, абсолютный максимум ЧОТ в тексте, средний частотный максимум всех ударных слогов в тексте, средний максимальный уровень ЧОТ ядерных слогов во фразах, средний минимальный уровень ЧОТ текста, абсолютный минимум ЧОТ в тексте); средняя длительность слога в тексте (СДС) измерялась в миллисекундах (мсек).

Усредненные количественные величины параметрических значений представлены на диаграммах.

Как свидетельствуют показатели количественных величин, в каждой исследованной группе просодическое структурирование публичных сообщений основано на наличии нескольких константных совокупностей устойчивых признаков интонирования фраз на перцептивном и акустическом уровнях их анализа, сигнализирующих о надежности идентификации различными аудиторами профессиональной принадлежности говорящих, наряду с признаками тождества вариантов просодической выразительности публичной речи как самостоятельного дискурсивного феномена.

Т а б л и ц а 1

Локализация величин ЧОТ в текстах (Гц)

	max _{нач}	абс. max	средн. max	max _{яд. сл.}	средн. min	абс. min
Teacher 1	191	250	160	250	69	65
Teacher 2	345	355	307	355	83	62
Actress 1	178	230	203	230	126	42,9
Actor 2	175	269	235	269	130	42
Lawyer 1	237	292	154	252	110	42,9
Lawyer 2	289	342	210	342	44	42
Politician 1	351	359	248	318	166,4	77,9
Politician 2	355	368	300	327	80	47

В целом анализ показал, что орфоэпическая нормативность английской публичной речи представляет собой определенную зону просодических вариантов ее реализации, обусловленных, в том числе, и действием так называемых «вторичных» социолингвистических факторов, среди которых одной из наиболее значимых предпосылок внутристилистической вариативности интонирования английской официально-деловой речи является профессиональная принадлежность говорящих. Проведенный комплексный анализ на перцептивном и акустическом уровнях показал различный характер проявления признаков, участвующих в дифференцировании всех сравниваемых 20 текстов. К первой группе признаков, демонстрирующих наиболее эксплицитные межпрофессиональные различия, относятся количественные показатели особенностей сегментации текстов на просодические единицы и их темпоральная организация. Средние значения по данным признакам представлены на диаграмме 1.

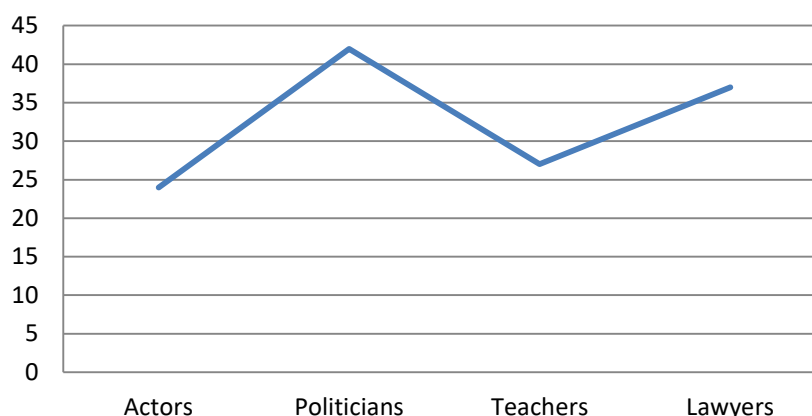


Диаграмма 1. Количество просодических единиц в текстах

Как показывают количественные данные, наибольшее число интонационных групп в речи учителей и юристов составило 27 и 36 просодических единиц соответственно. Наибольшее количество просодических единиц за минуту звучания текста продемонстрировали политики (42 интонационные группы в минуту), а наименьшее их количество зарегистрировано в речи актеров (24).

Показатели среднеслоговой длительности (СДС) политиков, составил 220 мсек, учителей – 200 мсек, юристов – 190 мсек, актеров – 170 мсек.

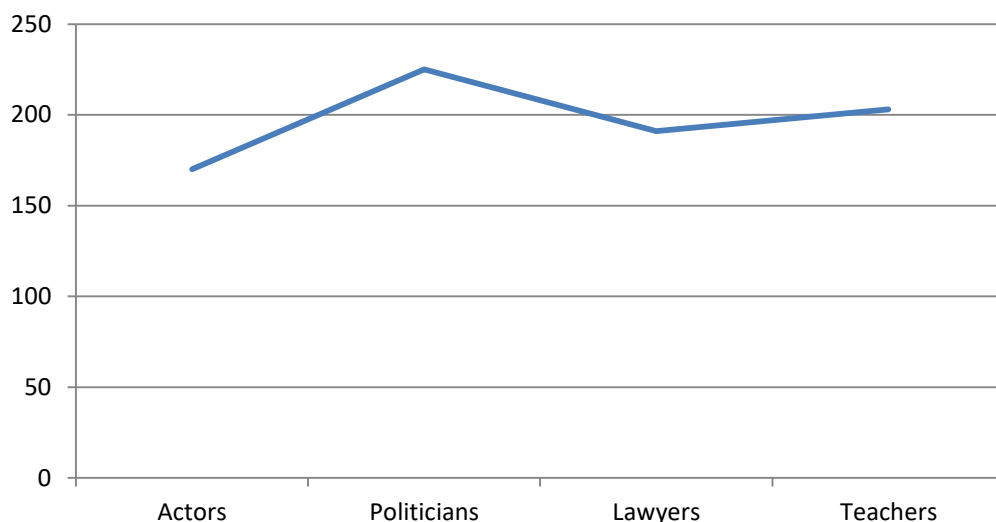


Диаграмма 2. Среднее значение СДС в текстах (мсек)

Наиболее контрастные различия в среднеслоговой длительности текстов характерны для политиков и актеров, и менее значительные в сообщениях учителей и юристов. Процент различия сообщений юристов и учителей по исследованным признакам значительно снижается, и степень их дистантности менее выражена по сравнению с оппозицией «актеры – политики». Наиболее контрастные «полярные» различия величин параметровых признаков, которые образуют константные различия отдельных групп испытуемых, установлены между политиками и актерами в паузальном членении текстов длинными паузами (17 % и 60 % соответственно), и тональном завершении нефинальных интонационных групп (в частотности употребления типов тонального завершения: низкого и высокого восходящих тонов у актеров, и высокого нисходящего – у политиков).

Наибольшая дистантность сообщений, по показаниям величин их параметрических значений, формирующих стабильные зоны межпрофессиональных контрастов, установлены между актерами и политиками. Наиболее высокий процент идентификации их профессиональной принадлежности всеми группами аудиторов, очевидно, опирался на комплексное восприятие делимитативных, темпоральных, тональных характеристик их публичных выступлений.

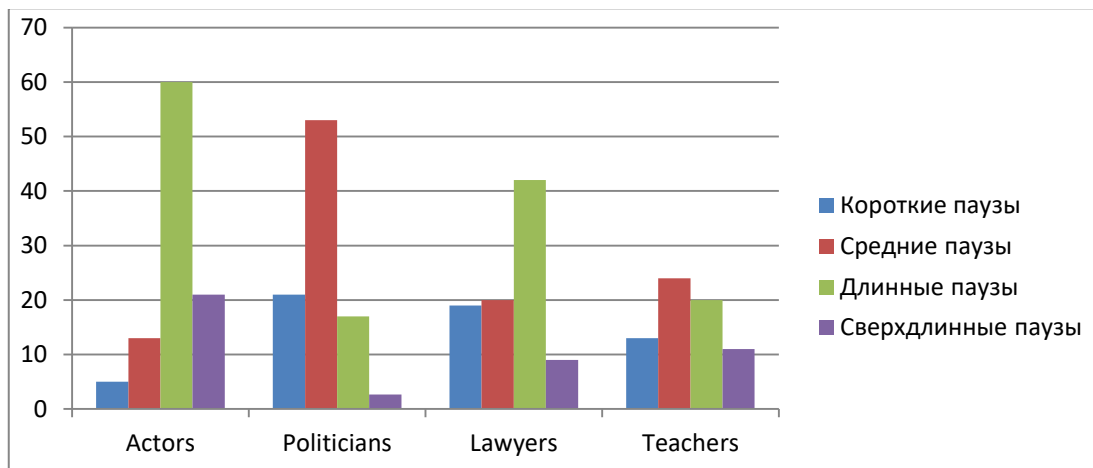


Диаграмма 3. Паузальное членение текстов

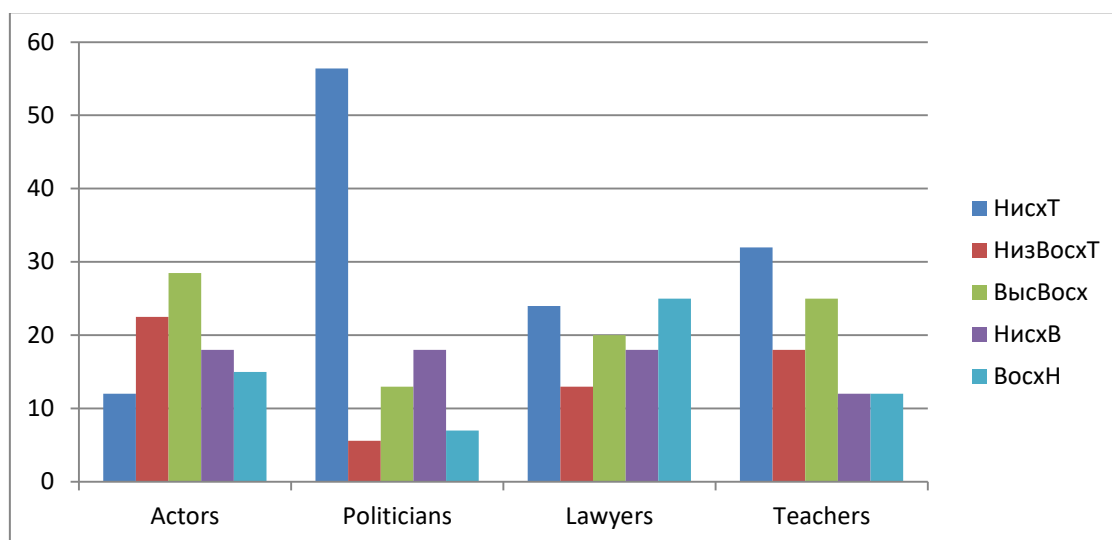


Диаграмма 4. Тональное завершение нефинальных интонационных групп

Степень дистантности просодических структур фраз в речи юристов и учителей носит градуальный характер относительно других групп испытуемых, и она менее выражена по сравнению с контрастностью речи актеров и политиков.

Количественные данные проведенного эксперимента свидетельствуют о том, что в целом наиболее значимые признаки различия в публичной речи представителей сравниваемых в исследовании профессий при их парном рассмотрении заключаются в различном выборе типов тонального завершения финальных и нефинальных интонационных групп, степени вариативности направления тона на отдельных предъядерных участках интонационного контура фраз, темпе речи и среднеслоговой длительности ядерных слогов, а также уровню и диапазону частоты основного тона. Особую группу признаков, интегрирующих английскую публичную речь как стилистическую макроединицу, независимо от профессиональной принадлежности испытуемых формируют однотипное членение текстов на одноакцентные и двухакцентные интонационные группы, равномерное распределение фразовой слоговыделенности, частотность совпадения синтаксического и просодического членения, однотипность внутрифразового и межфразового членения текстов паузами малой продолжительности в трех сравниваемых группах испытуемых (исключение составляют актеры). Указанные признаки в их взаимодействии создают просодический «портрет» подготовленного устного публичного дискурса в условиях односторонней коммуникации.

ЛИТЕРАТУРА

1. Блох, М. Я. Публичная речь и ее просодический строй / М. Я. Блох, Е. Л. Фрейдина. – М. : Прометей, 2011. – 236 с.
2. Просодия публичной речи / Е. Л. Фрейдина [и др.]. – М. : Прометей, 2013. – 224 с.
3. Брантов, С. А. Просодическая составляющая риторической аргументации в публичной речи: на материале британских лекций : автореф. дис. ... канд. филол. наук : 10.02.04 / С. А. Брантов. – М., 2004. – 20 с.

4. Данилина, В. В. Политическая ораторская речь в ритмико-текстологическом аспекте : автореф. дис. ... канд. филол. наук : 10.02.04 / В. В. Данилина. – М., 2002. – 24 с.
5. Кочеткова, Т. В. Языковая личность носителя элитарной речевой культуры : автореф. дис. ... докт. филол. наук : 10.02.01 / Т. В. Кочеткова. – Саратов, 1999. – 53 с.

И. А. Болдак

СОЦИАЛЬНАЯ ДЕТЕРМИНИРОВАННОСТЬ НАУЧНОЙ БИОГРАФИИ

Биография как отдельная литературная форма в последние десятилетия вызывает все больший интерес филологов. В центре внимания ученых-лингвистов оказались такие аспекты изучения биографии, как фактор адресованности [1], коммуникативные стратегии автора [2], автобиография как жанр [3; 4], категория авторской «вовлеченности/отстраненности» в англоязычной биографической прозе [5], проявление концепта «замечательный человек» в биографическом тексте [6]. Тем не менее, проблема описания типологических особенностей разных видов биографических произведений по-прежнему остается дискуссионной. Изучение текста научной биографии в плане представления логического содержания позволяет определить смысловые особенности текстов научной биографии, которые не всегда обладают ярко выраженными формально-языковыми характеристиками.

Научная биография отличается по ряду параметров от общелитературной биографии. В общелитературной биографии личность писателя или художника представлена как часть отдельно взятой истории и культуры одной страны. В научной биографии личность ученого выступает как часть общей мировой культуры и науки. В содержательном плане научная биография представляет собой не столько изложение событий жизни человека, сколько описание этапов его научной деятельности, оценку его вклада в науку, и уже после этого анализ факторов, оказавших влияние на становление и развитие творческой личности ученого [7, с. 149–151]. В текстах научных биографий раскрываются главные черты индивидуальной научной биографии человека на фоне «биографии» самой науки.

Научная биография описывает результат социально-культурной научной деятельности человека, направленной на создание более точной картины мира с целью ее использования в практической жизни человека в отличие от общелитературной биографии, которая влияет на мировоззрение адресата, на его жизненный опыт.

Еще одно отличие научной биографии от общелитературной состоит в том, что описание научной деятельности ученого не связано непосредственно с его личным жизненным опытом. Она образует отдельную плоскость в рамках жизненного пути ученого и в центре ее внимания оказывается, прежде всего, социально-научный аспект деятельности человека. Только по отношению к нему описываемая жизнь приобретает значение и особый смысл, к которому применимы понятия уникальности, важности для дальнейшего развития науки и общества в целом.

Социальная детерминированность научной биографии проявляется в ее структуре по-разному. Деятельность отдельного ученого описывается как часть коллективной биографии науки. Во всех текстах научных биографий отражается вклад конкретного ученого в коллективное научное творчество. Научная биография помогает понять, как взгляды и достижения конкретного человека повлияли на направление, характер и процесс научной деятельности социума в целом. Она также помогает проникнуть в суть теории, созданной выдающимся ученым и уяснить, что при всех неточностях, допущенных ученым, его главный вклад в развитие науки не вызывает сомнения и является исключительно важным. Через жизнеописания ученых многие вопросы науки и истории науки могут восприниматься более глубоко, что способствует более эффективному процессу познания [8, с.43] .

Вклад ученого в развитие науки как социального феномена обязательно предполагает связь с другими научными исследованиями. Так, в научной биографии Ньютона, приведенной в энциклопедии “Men and Women of Science” [9], упоминаются теории таких ученых, как Коперник, Кеплер, Эдмунд Галлей, Готфрид Лейбниц, Аристотель, Роберт Хук [9, с. 23–26]. В биографии Галилея подчеркивается связь его теории с теорией Коперника, в биографии Дэйви – его связь с исследованиями Майкла Фарадея и т.д. Например: “*Galileo’s spectacular findings thoroughly convinced him of the truth of the Copernican theory about the solar system*” [9, с. 12]; “*Davy examined the relationship between electricity and magnetism. These researches probably influenced his assistant, Michael Faraday, who was destined to become one of the leading scientists in the fields of electricity and magnetism*” [9, с. 38]; “*Ampere made his fundamental discoveries in the 1820’s, spurred by the Danish physicist Hans Oersted’s 1820 observation that a magnetic needle is deflected when placed near a wire-carrying current*” [9, с. 39]; “*Maxwell took Faraday’s idea and developed it into a complete electromagnetic theory*” [9, с. 52].

Исследования отдельного ученого являются одной из вех развития коллективного общенаучного социума, и это проявляется через четкое указание на названия и годы основных публикаций. Данная семантическая составляющая структуры научной биографии практически всегда представлена в исследуемых текстах, например: “*Newton first published some of his ideas on light and colors in the papers for the Royal Society of London. His essay, “On Colors”, appeared in 1672. In 1675 he published the essay, “An Hypothesis Explaining the Properties of Light”. Newton published his entire body of theories on light in a book called Opticks in 1704*” [9, с. 25]; “*... Descartes developed a thorough method for gaining and testing knowledge. He wrote about his method in several books. In Discourse on Method, published in 1637, Descartes presented his basic philosophy and examined certain scientific questions*” [9, с. 19].

Социальная детерминированность научной биографии проявляется и через описание отношений учитель-ученик, что свидетельствует о преемственности научных исследований. Так, в биографии Эрнеста Резерфорда акцентируется тот факт, что его учителем является Джозеф Джон Томсон, а его учеником – Ханс Гейгер [9, с. 70–71]. Под руководством Джозефа Джона Томсона Резерфорд подготовил свою дипломную работу, а вместе с Х. Гейге-

ром он разработал прибор, известный как счетчик Гейгера. В биографии Аристотеля подчеркивается, что в возрасте 18 лет он поступил в Академию знаменитого ученого Платона в Афинах [9, с. 6]. В биографии Тихо Браге, датского астронома, упоминается тот факт, что его научная работа позволила его ученику и ассистенту И. Кеплеру подтвердить теорию Коперника [9, с. 12].

Практически в каждой научной биографии подчеркивается общественное признание вклада ученого и дается перечисление наград, например: *"Many honors and awards came to Thomson during his long, productive life. In addition to his Nobel Prize, Thomson won the Order of Merit in 1912. He was knighted by King Edward VII in 1908"* [9, с. 62]; *"In addition to his science awards and honors, Rutherford received civic honors as well. He was made a knight in 1914. Then, in 1931, he was made a member of nobility – Baron Rutherford of Nelson. ... He is buried in Westminster Abbey, a great honor for an English citizen."* [9, с. 71]; *"In addition to the Nobel Prize, Fermi received the Congressional Medal of Merit in 1946. ... The Enrico Fermi Award was established in his honor ... An artificially produced chemical element, fermium, was named for him in 1952."* [9, с. 82].

Такая семантическая составляющая социальной детерминированности, как общественная жизнь в рамках научного социума, в структуре научной биографии может выделяться в отдельный абзац, что свидетельствует о ее важности. Так, в частности, отдельным абзацем она представлена в биографии сэра Чарльза Лилля: *"Lyell participated in many civic activities. He served on a team with the great physicist, Michael Faraday, to find ways to prevent mining disasters. He was also a commissioner for the Crystal Palace Exposition in 1851-52. ... Lyell also helped reform the curriculum (course of study) at Oxford University"* [9, с. 44].

Указание на национальную принадлежность, на наш взгляд, также подчеркивает социальную детерминированность научной биографии. Национальная принадлежность приводится чаще всего в первом же предложении: *"Sir William Herschel (1738-1822) was a German-born British astronomer"* [9, с. 31]; *"Lord Kelvin (1824-1907) was a great British physicist and inventor"* [9, с. 45]; *"Louis Pasteur (1822-1895) was a great French chemist and microbiologist"* [9, с. 49]. Иногда данные о национальной принадлежности идут в конце, где приводятся факты личной биографии: *"Johann Mendel was born in Austria Silesia ... in 1822"* [9, с. 55]; *"Sigmund Freud was born in Moravia ... of Jewish parents in 1856"* [9, с. 67]. Эти данные могут дублироваться и в начале и в конце текста, как в биографии Нильса Бора [9, с. 7475], Энрико Ферми [9, с. 81–82]. Национальная принадлежность может содержаться и в косвенном указании: *"Pierre Curie's life ended early and tragically when, in April, 1906, he was killed by a speeding wagon in the streets of Paris. ... By the time he was 18, Pierre was working as a laboratory assistant at the Sorbonne in Paris"* [9, с. 63].

Таким образом, представленность в структуре научной биографии таких семантических составляющих, как связь с другими научными исследованиями, взаимоотношения учитель-ученик, общественная жизнь в рамках научного социума и т.п., свидетельствует о ее социальной детерминированности.

ЛИТЕРАТУРА

1. *Евсеева, Ю. Н.* Адресованность как определяющий фактор лексических и структурных особенностей текста биографии : автореф. дис. ... канд. филол. наук : 10.02.04 / Ю. Н. Евсеева. – М., 2010. – 19 с.
2. *Козлова, О. В.* Взаимосвязь интер- и метатекста как актуализация коммуникативных стратегий автора в немецком биографическом дискурсе : дис. ... канд. филол. наук : 10.02.04 / О. В. Козлова. – Иркутск, 2010. – 159 с.
3. *Кованова, Е. А.* Риторика автобиографического дискурса (на материале автобиографий американских деятелей политики и искусства) : автореф. дис. ... канд. филол. наук : 10.02.04 / Е. А. Кованова. – СПб., 2005. – 19 с.
4. *Волошина, С. В.* Автобиографический рассказ как объект лингвистического исследования / С. В. Волошина // Вестник ТГУ. – Томск, 2008. – № 3. – С. 11–15.
5. *Сафонова, М. А.* Стилистическое своеобразие англоязычной биографической прозы в составе публицистического функционального стиля : автореф. дис. ... канд. филол. наук : 10.02.04 / М. А. Сафонова. – М., 2011. – 23 с.
6. *Терпугова, А. В.* Биографический текст как объект лингвистического исследования : дис. ... канд. филол. наук : 10.02.19 / А. В. Терпугова. – М., 2011. – 251 с.
7. *Головаха, Е. И.* Жизненный путь личности / Е. И. Головаха, А. А. Крепик // Вопросы языкознания. – 1982. – № 8. – С.149–151.
8. *Соколовская, З.* Возможности биографического жанра / З. Соколовская // Библиотекарь. – 1986. – № 4. – С. 43–44.
9. Men and Women of Science and Index. The World Book Encyclopedia of Science. – Chicago : World Book, Inc., A Scott Fetzer company, 1991. – Vol. 8. – 128 p.

Т. В. Бусел

СОВРЕМЕННЫЕ ПОДХОДЫ К ПОРОЖДЕНИЮ ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

Проблема порождения связного текста на естественном языке приобрела особую актуальность в связи с развитием перспективных информационных и коммуникационных технологий, в частности глобальной компьютерной сети Интернет, а также необходимостью быстрого ввода, обновления, обмена текстовой информацией в режиме реального времени. Решение данной проблемы возможно через моделирование и последующую автоматизацию процесса порождения текстов на естественном языке.

Несмотря на достаточно длительную историю развития подходов к порождению текстов, разнообразие предложенных методов, эта проблема еще не до конца решена. Большинство действующих компьютерных моделей используют шаблонные технологии, главная особенность которых заключается в том, что содержание будущего текста в них представлено в виде готовых фрагментов, что касается лингвистически мотивированных технологий, то они находятся в стадии исследования.

Моделирование процесса порождения текстов на естественном языке предполагает решение вопросов, связанных с изучением жанровой и коммуникативных характеристик текстов, риторических приемов организации

содержания текста, языковых средств выражения связанности текста, формализацией грамматики и лексических описаний. Таким образом, рассматриваемая проблема находится в неразрывной связи с рядом актуальных вопросов современного теоретического языкознания и компьютерной лингвистики: способами представления смысловой структуры текста и его единиц; закономерностями построения текста; вариативностью языкового выражения его содержания; определением качественного своеобразия функционирования языковых единиц различных уровней в тексте; построением типологии текстов.

В настоящее время известны многочисленные подходы к процедуре порождения текстов с помощью компьютера, которые подробно освещаются в работе американского ученого Эдварда Хови [1]. Зависят они в основном от того, для какой цели создается текст. По степени сложности и выразительности, существующие методы порождения сообщений принято подразделять на четыре класса:

1. *Canned-based method*. Для порождения сообщений создаются таблицы неизменяющихся шаблонов, которые используются системой в зависимости от ситуации. Этот метод предназначен для порождения простых цепочек слов (например, сообщение об ошибке в работе программного продукта т.д.).

2. *Template-based method*. Данный метод связан с созданием различного рода диалоговых систем, применяемых в справочных и обучающих системах, как правило, это шаблонные системы (*template systems*), которые используют готовые реплики или комбинируют готовые фрагменты текста таким образом, что они занимают заданные позиции в дискурсе или стереотипном тексте.

3. *Phrase-based method*. Более сложный метод, при котором используются универсальные фразовые шаблоны на синтаксическом уровне и на уровне дискурса, поэтому их также называют «планами текста» (*text plans*). В таких системах, фразы строятся в соответствии с определенной моделью (например, подлежащие + сказуемое + дополнение), а затем каждая составляющая данной модели находит свое воплощение в соответствии с грамматическими правилами, заложенными в систему. Процесс построения модели предложения завершается тогда, когда каждая его составляющая выражается конкретным словом или сочетанием слов. Такие системы являются достаточно эффективными, но имеют определенные ограничения, вызванные необходимостью детально описывать межфразовые связи и способы их реализации, для построения грамматически правильных предложений.

4. *Feature-based method*. Это наиболее сложный метод, он требует привлечения обширных лингвистических знаний, но, в то же время, он и наиболее привлекателен. Синтез сообщения осуществляется на основе набора свойств (грамматических признаков). При таком подходе предложение определяется набором характеристик составляющих его слов (например, наличие/отсутствие отрицания, настоящее/прошедшее время) и правилами их сочетаемости.

Архитектура современных систем генерации текстов на естественном языке (ГЕЯ), как правило, представлена тремя основными составляющими: *оболочка, планировщик и лингвистический компонент*. Рассмотрим их основные функции, опираясь на работы [2, с. 12–14; 3, р. 53–74].

Оболочка (underlying application program) определяет назначение СТГ и характер баз знаний, из которых черпается информация для построения текста. Оболочка выполняет две основные функции: иницирует процесс генерации и определяет цели, которые должны быть достигнуты высказываниями.

Планировщик определяет пути достижения высказываниями поставленных оболочкой целей в данном предметном контексте. Он обеспечивает:

- 1) выбор информации, которая должна быть выражена или опущена;
- 2) определение того, как она должна быть представлена (как событие, например, “the economy developed” или как объект, например, “the development of the economy” и т.д.);

- 3) выбор способа взаимодействия с лингвистическими данными (лексика открытых классов, синтаксические конструкции). В частности, планировщик выполняет следующие задачи:

- структурирование текста – определение порядка пропозиции и границ предложений в выходном тексте;
- выбор лексики (без ограничений на синтаксический класс);
- построение синтаксической структуры предложений выходного языка;
- языковое оформление отношений кореференции (анафора, дейксис, эллипсис).

Каждая из перечисленных выше задач планировщика только теоретически может рассматриваться полностью изолированно от других. При экспериментальном моделировании, особенно в составе СТГ, их функции перемежаются.

Лингвистический компонент порождает тексты в соответствии со спецификациями планировщика. Он обеспечивает грамматическую правильность текста и принимает большую часть, если не все синтаксические и морфологические решения. Полностью обеспечивает процессы синтаксического и морфологического синтеза текста на основе синтаксической структуры.

Процесс ГЕЯ принято представлять с использованием известной в теории информационных систем идеи конвейера обработки данных. Путем обобщения опыта создания действующих систем ГЕЯ, построена схема, представленная на рис. 1, которая отражает общую картину преобразований в системе ГЕЯ.

Из схемы видно, что система генерации состоит из трех относительно независимых блоков. Макропланирование – построение плана текста. Микропланирование – построение планов предложений и языковое оформление – реализация построенных планов предложений средствами конкретных ЕЯ. Рассмотрим более подробно каждый из этапов генерации.

Основная цель этапа **макропланирования** – создание плана текста. Для этого из входных данных система выбирает данные, релевантные поставленной коммуникативной цели, и организует их в последовательность изложения в будущем тексте.

Определение вида входных данных является кардинальным вопросом для ЛМ систем. Существуют три вида возможных входов для систем ГЕЯ – числовые данные, структурированные объекты и логические формулы. Более конструктивным является изучение входов, с которыми работают экспериментальные ЛМ системы. Можно выделить три вида таких входов.

1. Базы данных. Особенность этого типа источника состоит в том, что информация не организована для передачи адресату. Тип текста, который можно построить на основе этой информации, и его структура, должны быть определены извне.

2. Семантическое представление – представление содержания текста, созданное человеком с помощью системы интерфейсного типа «человек – компьютер», т.е. такой системы, которая позволяет построить семантическое представление из предлагаемых интерфейсом понятий на основе внутренней речи человека. Этот процесс называется “symbolic authoring” [4, с. 570].

3. Представление знаний на формальном языке.

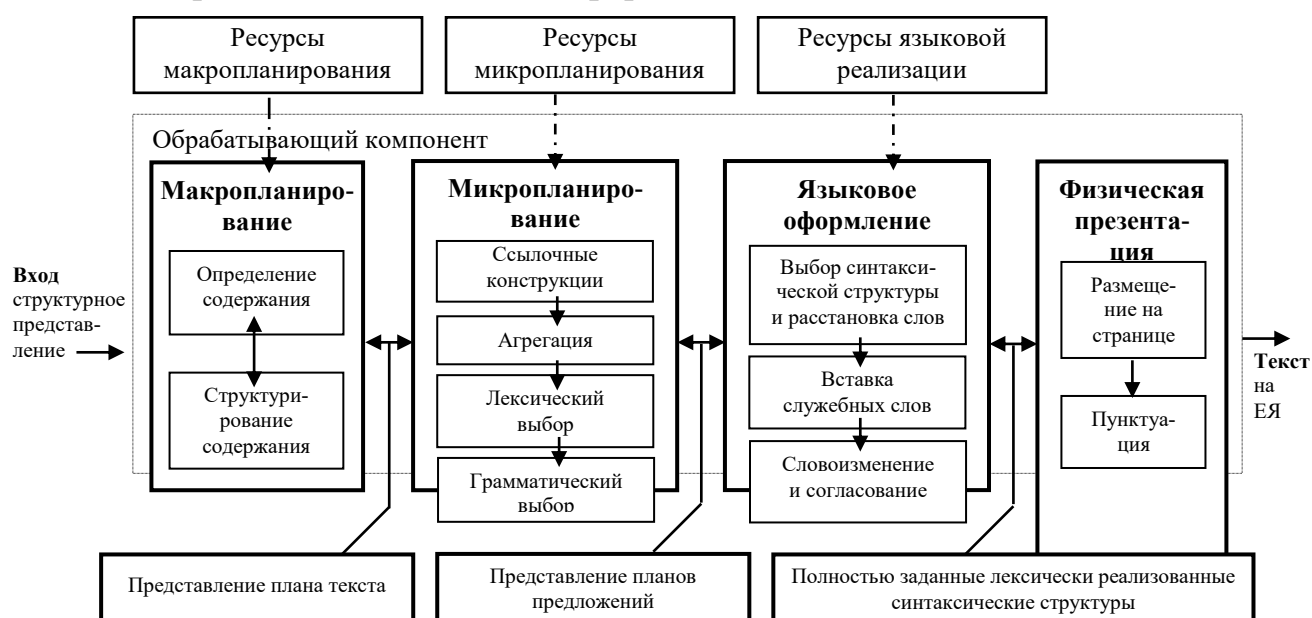


Рис. 1. Общая схема генерации

П л а н т е к с т а – это представление информации, составляющей содержание будущего текста, организованное в виде единой структуры, которая может быть отображена в структуру текста на ЕЯ. Для представления этой информации может использоваться концептуальное представление, состоящее из объектов и отношений между ними. Объекты концептуального представления – это экземпляры сущностей ПО, порожденные из них согласно информации, представленной системе генерации, отношения между объектами – это отношения между соответствующими сущностями модели ПО, которые планируется рассмотреть в создаваемом тексте.

Способ записи концептуального представления определяется решаемой задачей каждой конкретной системы генерации. Отношения концептуального представления имеют предикативную природу, т.е. каждому отношению можно сопоставить предикат, и представить в виде фрейма, в котором каждый из объектов отношения имеет фиксированную роль. Поэтому отношения легко интерпретировать как планы отдельных высказываний – сообщения.

Сообщения являются концептуальными планами отдельных высказывание текста. Они могут представляться различными средствами с разной степенью общности, начиная с представления сообщений как идентификаторов обсуждаемых в тексте тем, и кончая их представлением в виде логических формул и шаблонов высказывание.

Концептуальное представление не дает исчерпывающего ответа о порядке изложения содержащейся в нем информации, о том какие фреймы должны быть выделены из этого представления, и в каком порядке они должны быть включены в результирующий текст. Поэтому оно не может считаться планом текста.

Вопрос о том, по каким правилам строится текст, как выстроить отдельные высказывания текста, чтобы достичь желаемого воздействия на адресата, изменения его знаний, убеждений и намерений, изучает риторика. Поэтому вопрос построения плана текста – это вопрос построения его риторической структуры.

Риторическую структуру имеет текст, призванный воздействовать на адресата. Не случайно в античности риторика обслуживала речи ораторов. Но воздействие на адресата – не единственный мотив организации текста. Стратегия построения текста может также определяться самими данными, однако все они могут быть представлены в терминах риторической структуры, которая в данном случае рассматривается скорее как форма представления структуры текста.

Основные принципы риторического моделирования структуры текста сформулированы в теории риторических структур – *Rhetorical Structure Theory (RST)*. Как отмечают В. Манн и С. Томпсон «риторическая структура – это древесное представление определяющее, как будет организован создаваемый текст» [5, р. 243–281]. Его терминальные вершины – сообщения соответствуют отдельным высказываниям текста. Внутренние вершины риторического представления описывают, как сообщения структурированы вместе и связаны друг с другом. Внутренние вершины задают риторические отношения между связанными с ними фрагментами текста, например, цель, причина, последовательность, уточнение, побуждение, разрешение и т.д. Каждая внутренняя вершина разделяет свое содержание как минимум на две части: главную (*nucleus*) и второстепенную (*satellite*).

Группировка текстовых фрагментов риторическими отношениями накладывает ограничения на будущее разделение порождаемого текста на абзацы и предложения. Существование единого покрытия текста риторической структурой соответствует ощущению человеком его цельности.

Существуют два основных подхода к генерации риторической структуры текста: «основанный на планирующих операторах на предикативных схемах» [6, с. 5–6].

Первый подход более теоретически мотивирован – задача выбора структуры текста полностью перекладывается на ресурсы генератора – планирующие операторы.

Второй – более подходит для практических задач. Он задает предварительные шаблоны структуры текста – предикативные схемы, уточняемые при построении структуры конкретного текста. Данный подход, впервые предложенный К. Макьюин [7, с. 2–42], основан на замечании, что для конкретной коммуникативной цели люди регулярно используют одни и те же виды информации в одном и том же порядке. Предикативная схема – это шаблон, который определяет, как должен быть организован текст, используя для этого более мелкие схемы, сообщения и дискурсные отношения между ними.

После построения плана текста и сообщений выполняются задачи микропланирования. Целью данного этапа является составление плана отдельных предложений генерируемого текста на основе сообщений с учетом общей структуры текста.

Для представления планов предложений на практике используются различные техники:

1. План предложения может состоять из набора уже готовых фрагментов высказывания, которые на этапе языкового оформления нужно лишь немного доработать для лучшей согласованности частей предложения.

2. План предложения может представлять собой полуграмматическую структуру, содержащую отдельные фрагменты высказывания, грамматические единицы и семантические признаки.

3. План предложения может быть представлен семантической структурой. Своеобразие семантических представлений состоит в сочетании достаточно высокой абстракции и одновременно близости к ЕЯ, что позволяет рассматривать их как универсальное предъязыковое представление данных или стандарт представления планов предложений.

Семантическое представление предложения строится из одного или нескольких соседних сообщений с учетом окружающей их риторической структуры. Для того чтобы провести такое преобразование на этапе микропланирования, выполняются четыре основные задачи:

1. **А г р е г а ц и я**, в ходе которой происходит объединение простых фраз в более сложные структуры предложений (простое сочинение, синтаксическое подчинение и т.д.).

2. **Л е к с и к а л и з а ц и я** концептов сообщения, т.е. выбор подходящих слов для выражения их содержания.

3. **В с т а в к а с с ы л о ч н ы х к о н с т р у к ц и й**. Для обеспечения лучшей слитности текста при повторном упоминании объекта в высказываниях для его идентификации выбираются различные слова или словосочетания (местоимения, дефинитные описания и т.д.).

Таким образом, на этапе микропланирования построенные сообщения преобразуются, учитывая их расположение в плане текста, в планы отдельных предложений.

На этапе языкового оформления эти планы реализуются средствами лексики и грамматики конкретного ЕЯ в грамматические структуры, которые затем преобразуются в предложения ЕЯ текста. Готовые предложения собираются в результирующий текст.

Этот этап называется также поверхностной реализацией и базируется на положениях трансформационной грамматики, разработанной Н. Хомским, который разделил лингвистические представления на глубинные и поверхностные. Глубинное грамматическое (семантическое) представление фактически содержит план поверхностной грамматической структуры высказывания. На этапе поверхностной реализации генератор выбирает грамматические конструкции – функциональные роли (подлежащее, прямое дополнение и т.д.), определяет линейный порядок (определяемый грамматикой), части речи (существительное, глагол и т.п.), сложность предложения (простое, сложное) и окончательную форму слов (морфология), вставляет служебные слова (союзы, предлоги, артикли). Ресурсы этого уровня описывают лексический, морфологический и синтаксический уровни лингвистической модели языка.

Таким образом, в процессе генерации входное представление последовательно преобразуется между следующими лингвистическими уровнями: концептуальный уровень, семантический уровень, риторический уровень, синтаксический уровень и текстовый уровень. Считается, что первые три уровня описывают надязыковые явления. Последние два уровня описывают явления, специфичные для конкретного языка. Генерация в такой уровневой модели может быть определена как лингвистически мотивированный процесс построения текста на ЕЯ последовательным преобразованием его порождаемой структуры от концептуального уровня к текстовому.

Анализ приведенных выше подходов к созданию систем генерации показывает, что данные модели, как правило, не рассчитаны на порождение текстов по заданному содержанию.

Нами была разработана вероятностно-алгоритмическая модель порождения текста, позволяющая порождать тексты деловых электронных писем различных типов на английском языке по заданному содержанию. Созданная модель является алгоритмической, относящейся к классу воспроизводящих лингвистических моделей, которые с одной стороны, имитируют структуру и поведение реальных лингвистических объектов, а с другой – позволяют воспроизводить эти объекты. Учитывая многоаспектность, многоуровневость порождаемых объектов, при создании данной модели мы стремились наиболее полно и точно отразить в ней реальные свойства языка.

Процесс генерации текста компьютерной системой происходит в две стадии. Первая определяет содержание и структуру будущего текста, это *стратегический компонент*, его еще называют “планировщиком” текста. Вторая стадия – *лингвистический* (или *тактический*) *компонент*, – определяет, как строить текст делового письма, какие лексические, синтаксические и коммуникативные средства естественного языка нужны для порождения текста.

На основе анализа различных текстов деловых писем была создана база данных для работы программы порождения. Для каждого текста делового письма в такой базе данных указаны:

1. Логико-семантическая формула текста, которая представляет собой линейную последовательность абзацев с определенным предметно-логическим содержанием.

2. Таблица основного статического содержания (ТОС), в которой указаны главные и второстепенные опорные слова, которые отражают главные субъекты и объекты ситуации, описываемой в тексте.

3. Алфавитно-частотный словарь текста.

4. Семантико-синтаксические формулы абзацев (СЕСФА) текста, представленные на специальном семантико-синтаксическом языке, в основе которого лежат семантические функции.

Подбор СЕСФА для заданного основного содержания и порядок их следования в пределах семантико-синтаксической формулы текста определяется двумя типами факторов:

а) вероятностными, выявленными в процессе статистического анализа следования абзацев с разным предметно-логическим содержанием в текстах деловых писем;

б) детерминированными, полученными в результате качественного изучения ТОС и СЕСФА.

Определяющими при этом являются факторы детерминированные.

При внедрении в промышленные системы обработки связных текстов созданная модель вероятностно-алгоритмического порождения текстов на естественном языке позволит грамотно и быстро порождать различные типы деловых документов по заданному содержанию, соответствующие международным стандартам, и может быть использована даже специалистами, не владеющими в совершенстве английским языком. Поскольку английский язык является основным средством международного делового общения во всем мире, данная вероятностно-алгоритмическая модель порождения может иметь широкое практическое применение.

ЛИТЕРАТУРА

1. Hovy, E. Language Generation / E. Hovy// Survey of the state of art in human language technology [Электронный ресурс]. – Режим доступа : <http://www.isi.edu/natural-language/people/hovy/publications.html>. – Дата доступа : 10.06.2008.
2. Соколова, Е. Г. Лингвистические компоненты в экспериментальных системах генерации текстов (по работам ученых США и Канады) / Е. Г. Соколова // НТИ. Сер. 2. Информационные процессы и системы. – 1993. – № 4. – С. 10–14.
3. Bateman, J. A. An overview of computational text generation / C. Butler, editor // Computers and Texts : An Applied Prospective. – Basil Blackwell, Oxford, England, 1992. – P. 53–74.
4. Соколова, Е. Г. Генерация текстов на естественном языке – состояние вопроса и прикладные системы / Е. Г. Соколова, М. В. Болдасов // НТИ. Сер. 2. Информационные процессы и системы. – 2005. – № 10. – С. 12–22.
5. Mann, W. Rhetorical structure theory toward a functional theory of text organization / W. Mann, S. Thompson // Text. – 1988. – Vol.8. – № 3. – P. 243–281.
6. Болдасов, М. В. Генерация текстов на естественном языке – теории, методы, технологии / М. В. Болдасов, Е. Г. Соколова // НТИ. Сер. 2. Информационные процессы и системы. – 2006. – № 7. – С. 1–14.
7. McKeown, K. R. Text Generation: Using discourse strategies and focus constraints to generate natural language text / K. R. McKeown. – Cambridge University Press., 1986. – 246 p.

ТЕОРЕТИЧЕСКИЕ АСПЕКТЫ ПОНИМАНИЯ ПИСЬМЕННОГО ТЕКСТА КОМПЬЮТЕРОМ

В широком смысле понимание – сложное гносеологическое понятие, рассматриваемое такими науками, как философия и психология. Более конкретно к сути процесса понимания текста как частного вида понимания подходят психолингвисты и лингвисты. Необходимо заметить, что особые черты данная проблема приобретает в ходе изучения специфики понимания текста компьютером. Так, в работе В. С. Файна [1, с. 9–17] предложены два подхода к определению понятия «компьютер понимает текст»: психолингвистический и утилитарный, или практический. Согласно психолингвистическому подходу, все этапы процесса понимания текста компьютером проходят точно так же, как у человека. Однако при таком подходе не всегда ясно, что является конечным продуктом компьютерного понимания текста. Обычно говорят, что результатом является выделенный компьютером смысл текста. Но до сих пор нет единого определения этого понятия, поскольку смысл текста для каждого пользователя различен. Еще один недостаток данного подхода заключается в том, что заложенная в компьютер модель понимания текста должна отражать соответствующую психическую деятельность человека, осуществляемую им в ходе понимания текста. Однако многие составляющие процесса понимания текста и речи человеком до сих пор изучены не достаточно полно. При утилитарном подходе к объяснению процесса понимания текста компьютером предполагается, что способ решения данной задачи значения не имеет. Главное, чтобы результат работы компьютера был эквивалентен результату, полученному при решении той же задачи человеком, и чтобы получение такого результата было экономически выгодным. В данном случае не ставится задача моделирования всей психической деятельности человека, не имеет значения тот факт, понимает компьютер смысл текста или нет. Главное, чтобы компьютер правильно выполнил конкретное задание пользователя. В современных компьютерных системах понимания текста используются элементы обоих подходов.

В последние годы специалисты предлагают модель «мягкого понимания» текста компьютером, отражающую способность компьютера порождать различные осмысленные интерпретации исходного текста в зависимости от разных условий и составляющих процесса его восприятия. Такая модель «должна сочетать в себе структуры последовательного, буквального, поэлементного понимания, с одной стороны (узко лингвистический подход), и чтение «крупным взглядом», глазами специалиста, с другой стороны (информационный подход, экстралингвистическое понимание)» [2, с. 30]. Поэтому в рамках данного подхода создается информационно-лингвистическая модель понимания текста. Она базируется на новом плюралистическом подходе к пониманию текста на естественном языке и методах собственно лингвистического анализа на всех языковых уровнях, а также более грубом информационном анализе (выделении из текста релевантной для произвольного пользователя информации).

Согласно современным теориям автоматического понимания текста, компьютер понял текст, если он может кратко изложить основное содержание этого текста, т.е. создать его реферат; ответить на вопросы по тексту; нарисовать к тексту картинку, схему и т.д.; приведенные в тексте сведения представить в иной форме (например, в виде таблицы или графика); сравнить содержание двух текстов и определить, что в них общего и чем они отличаются; на основе анализа текста, написанного на одном языке, выдать адекватную информацию на другом языке, т.е. выполнить перевод текста; в результате анализа одного или нескольких текстов извлечь такие знания, которые можно поместить в некоторую базу знаний [2, с. 17–19].

Сегодня компьютерное понимание текста реализуется в рамках решения конкретной, относительно узкой задачи обработки текста. Соотнося уровни понимания текста компьютером с основными уровнями языка, можно выделить следующие уровни автоматического понимания: морфологический; синтаксический; семантический; гиперсинтаксический, или прагматический [3, с. 36; 4, с. 115–121]. Морфологическое понимание текста сводится к автоматическому приписыванию каждому слову текста его морфологических характеристик, например, части речи, рода, числа, падежа для имени существительного в русском языке, части речи, времени, лица, числа, вида и др. для русского глагола и т.д. Это – наиболее освоенный вид компьютерного понимания текста. Для его реализации используются различные методы построения формальных морфологий, строгих систем окончаний и приписывания каждому словупотреблению текста определенного морфологического типа. В последние годы такие операции легко реализуются в рамках автоматического лексико-грамматического анализа слов в процессе создания корпусов текстов. Суть синтаксического понимания текста заключается в автоматическом выделении в каждом предложении его главных и второстепенных членов и установлении между ними различных типов связей в виде некоторого дерева зависимостей. С этой целью используются различные формальные грамматики: грамматика непосредственно составляющих, формальные грамматики Н. Хомского, трансформационные грамматики, расширенные сети переходов, цепочечный анализ и т.д. [2, с. 78–100; 5]. В процессе семантического понимания текста автоматически устанавливаются значения выделенных ранее членов предложения. Например, компьютер определяет, что подлежащее выражено одушевленным или неодушевленным именем существительным; сказуемое выражено глаголом движения или чувствования; обстоятельство является обстоятельством места, или времени, или причины и т.п. При этом для каждого слова соответствующей синтаксической составляющей создается ролевой фрейм, который описывает то необходимое окружение, с которым данное слово всегда связано на этапе его понимания. Как результат семантического понимания строится семантический граф предложения. Он связывает предложение с определенной базой знаний, хранящейся в памяти компьютерной системы. Гиперсинтаксическое, или прагматическое понимание текста связано с выяснением семантических отношений между его предложениями и выявлением соответствующей тексту ситуации реальной действительности.

Компьютерная система, понимающая письменный текст, должна иметь в своей базе знаний следующую информацию:

- 1) знания о языке, на котором написан текст;
- 2) правила использования знаний о языке;
- 3) знания об определенных фрагментах объективной действительности (например, знания по физике, математике, метеорологии, автомобильному транспорту и т.д.);
- 4) правила использования соответствующих знаний о фрагменте реального мира;
- 5) знания о пользователе системы, желающем получить от компьютера некоторое понимание находящегося в его памяти текста (уровень понимания, цель понимания, намерения использования результатов понимания и т.д.) [4, с. 117].

В зависимости от сложности используемых компьютером знаний и их количества различают следующие типы автоматического понимания письменных текстов:

- 1) понимание-узнавание;
- 2) понимание-уподобление;
- 3) понимание-прогнозирование;
- 4) понимание-объединение;
- 5) понимание-объяснение [6, с. 128–139].

Понимание-узнавание использовалось в первых автоматических системах обработки текста. Его суть сводилась к тому, что компьютер узнавал морфологическую структуру каждого слова предложения и на этой основе строил синтаксическое дерево предложения, отражающее синтаксические отношения между входящими в предложение словами [4, с. 117]. Сегодня автоматический морфологический анализ слов большинства языков не вызывает затруднений. Более сложной задачей является автоматический синтаксический анализ предложения, т.е. построение синтаксического дерева зависимостей между его словами. Здесь используются различные теоретические подходы: теория предсказуемого анализа, грамматика непосредственно составляющих (НС–грамматика), трансформационные грамматики Н. Хомского, расширенная сеть переходов, цепочечный анализ и др. [7, с. 81–98]. Последние из этих теорий учитывают информацию не только о морфологических признаках слов, но и их семантических признаках. Следующий вид компьютерных систем, работающих по принципу «понимание-узнавание», связан с поиском в тексте ключевых слов, словосочетаний и предложений. К таким системам можно отнести системы автоматического индексирования, аннотирования и реферирования текста, основанные на статистическом и позиционном подходах.

Понимание-уподобление сводится к следующему. Считается, что компьютерная система понимает текст, если она может ответить на вопросы по данному тексту. При таком понимании компьютер находит в своей памяти ответы на заранее подготовленные разработчиками системы вопросы, касающиеся какой-то достаточно ограниченной темы. На эти вопросы в памяти компьютера имеются заранее подготовленные ответы. Последние

представляют собой наиболее вероятные ответы человека. Здесь используются не только морфологические и синтаксические знания о предложениях, но и определенные правила логического вывода.

Автоматические системы, работающие по принципу «понимание-прогнозирование», ориентированы на понимание текстов определенной достаточно узкой предметной области (компьютерные сети, обучение, термодинамика и т.п.). В данном случае процесс понимания опирается на лингвистические знания, представленные в виде словарей соответствующей предметной области и грамматики языка, а также на знания о предметной области. Последние чаще всего формализуются в виде семантической сети, в которой зафиксированы типичные для данной предметной области ситуации. Считается, что компьютер понимает текст, если он может определить, какой объект или событие конкретной предметной области соответствует отдельным фрагментам текста. Как показывает опыт создания таких систем, одним и тем же фрагментам текста могут соответствовать различные объекты, процессы и события соответствующей предметной области, поэтому определить их можно лишь с некоторой степенью вероятности. Величина такой вероятности зависит от цели создания компьютерной системы, сложности (достаточности) заложенных в нее знаний и, в конечном итоге, от решаемой задачи. Отличительной особенностью подобных систем является их умение прогнозировать будущий результат.

Основная особенность автоматических систем, работающих по принципу «понимание-объединение», заключается в том, что они должны уметь объяснить пользователю, почему приняли то или иное решение, выполнили то или иное действие. В базе знаний таких систем кроме знаний о мире (некоторой предметной области), представленных в виде определенных утверждений (аксиом, теорем), есть знания и о процессах (алгоритмах) оперирования этими утверждениями.

Компьютерная система, относящаяся к типу «понимание-объяснение», должна иметь в своей базе знаний следующие данные: знания о языке (его морфологии, синтаксисе, семантике); знания о правилах использования данных о языке; знания о мире (предметной области); знания о правилах использования данных о мире; знания о целях, намерениях, предпочтениях пользователя системы. Необходимость такого полного набора знаний объясняется двумя причинами.

1. По своей сути понимание является субъективным процессом. Человек извлекает из заложенного автором содержания текста свой смысл, который зависит от его уровня образования, знаний, жизненного опыта и т.п.

2. Практическая ценность понимания текста человеком заключается в умении объяснить то, что он понял текст, и только после этого использовать результаты понимания в своей деятельности.

То же самое требуется и от компьютерной системы понимания текста. Такая система должна уметь объяснить пользователю, каким образом она поняла текст, т.е. верно ли она выполнила его задание. Но с системами автоматического понимания текста работает не один конкретный пользователь, а большое количество разных людей. Следовательно, в такую систему

необходимо вложить интерсубъективные принципы понимания, чтобы объяснение общего принципа понимания система могла давать любому пользователю. Для реализации такого подхода в современных интеллектуальных понимающих системах используется диалог «компьютер–пользователь». В ходе диалога уточняются поставленные пользователем задачи, выясняются его запросы и мнения, объясняются принимаемые системой решения.

В дополнение к выделенным типам систем компьютерного понимания текста можно ввести еще один тип – понимание через преобразование кодов. Такие преобразования осуществляются, например, при переводе текста с одного естественного языка на другой или при преобразовании текста в некоторый специальный код (при использовании специального промежуточного языка – языка-посредника в процессе автоматического перевода текста). В данном случае словам исходного текста, представляющим субъектов, объекты, процессы, явления и отношения между ними, в другом тексте ставятся определенные коды (слова или специальные знаки).

Рассмотренные выше типы автоматического понимания письменного текста промышленно реализованы в виде целого ряда лингвистических процессоров, например, автоматического индексирования текста; аннотирования и реферирования текста; диалогового понимания текста, или экспертных систем; перевода текста; глубинного семантического анализа текста с целью извлечения фактов, понятий, оценок, мнений, знаний и т.д. Но, как отмечают специалисты по автоматической обработке текста, «между уровнем понимания текста человеком-специалистом и информационной системой имеется существенное различие, проявляющееся в том, что практически любой текст содержит, в частности, такую информацию, которая в принципе не может быть выявлена никакой информационной системой» [3, с. 30]. И далее: «Никакое глубинно-синтаксическое и глубинно-семантическое описание текста не может компенсировать машине отсутствие у нее «доступа» к реальному явлению, лежащему в основе этого текста» [3, с. 32]. Эту мысль подтверждают и известные психологи, заявляя, что «человеческое мышление не сводится к процессам работы с информацией» [8, с. 145]; «информационный подход... не выражает действительного строения мыслительной деятельности человека» [9, с. 31]. Тем не менее, хочется надеяться, что человечество найдет выход из всех проблем, касающихся понимания письменной речи, и придет к созданию совершенных компьютерных систем.

ЛИТЕРАТУРА

1. *Файн, В. С.* Распознавание образов и машинное понимание естественного языка / В. С. Файн. – М. : Наука, 1987. – 199 с.
2. *Леонтьева, Н. Н.* Автоматическое понимание текстов. Системы. Модели. Ресурсы / Н. Н. Леонтьева. – М. : ACADEMIA, 2006. – 303 с.
3. *Блехман, М. С.* Инженерная лингвистика и система «понимания» текста / М. С. Блехман // Квантитативная лингвистика и автоматический анализ текстов : учен. зап. Тарт. гос. ун-та. – Тарту : ТГУ, 1987. – Вып. 774. – С. 29–48.
4. *Совпель, И. В.* Понимание текстов на естественном языке / И. В. Совпель // Искусственный интеллект : в 3 кн. – М.: Радио и связь, 1990. – Книга 2 : Модели и методы. – С. 115–121.

5. Синтаксический компонент в системах машинного перевода: обзорная информация : сб. науч. ст. / под общ. ред. В. В. Жаркова. – М., 1981. – Вып. 5. – 198 с.
6. Психологические проблемы автоматизации научно-исследовательских работ : сб. науч. ст. / под общ. ред. П. В. Новикова. – М. : Наука, 1987. – 233 с.
7. Линдсей, П. Переработка информации у человека: введение в психологию / П. Линдсей, Д. Норман. – М., 1974. – 398 с.
8. Войскунский, А. Е. От психологии компьютеризации к психологии Интернета / А. Е. Войскунский // Новые педагогические и информационные технологии в системе образования. – 3-е изд. – М. : ACADEMIA, 2008. – С. 140–151.
9. Уотермен, Д. Руководство по экспертным системам / Д. Уотермен. – М. : Мир, 1989. – 378 с.

Т. П. Карпилович

РАСПОЗНАВАНИЕ СМЫСЛА АНАЛИТИЧЕСКОГО МЕДИАТЕКСТА ФОРМАЛЬНЫМИ МЕТОДАМИ

Современный этап автоматической обработки текстов характеризуется бурным развитием корпусных технологий, дальнейшим совершенствованием традиционных систем машинного перевода, аннотирования и реферирования, а также постоянным вниманием к таким новым направлениям, как распознавание мнений, эмоциональных оценок и состояний коммуникантов, извлечение из текстов знаний различных типов. Эти новые направления фактически направлены на выявление прагматики текста – такого семиотического измерения, которое вплоть до настоящего времени считалось невозможным формализовать для алгоритмической обработки.

Разработанные нами ранее алгоритмы поиска мнений участников медийного общения, включая автора текста [1], базируются, в первую очередь, на анализе таких поверхностных сигналов, как лексико-синтаксические структуры прямой и косвенной речи и далее – на лексических и грамматических средствах, маркирующих высказывания мнения. В качестве следующего шага в формализации прагматики медиатекста представляется целесообразным рассмотреть возможности распознавания основной идеи, или смысла целого текста (в качестве соответствующего понятия в современном научном и разговорном обиходе часто используется английское слово *message*). Рассмотрим вначале основные подходы к пониманию термина *смысл текста* в современной лингвистике.

Прежде всего, следует заметить, что необходимо разграничивать термины *смысл текста* и *семантика текста*. С точки зрения решения алгоритмических задач, представляется возможным рассматривать понятие семантики текста как синоним его содержания. Сравните в этом отношении следующее определение: «семантику текста составляет структура его содержания как целостное и системное образование, возникающее в интеллекте человека, в его отношении к линейной структуре текста» [2, с. 11]. Сходную точку зрения высказывает А. И. Новиков: с позиций разработанного им денотатного подхода, семантическую структуру текста (его внутреннюю форму)

составляет «совокупность денотатов, связанных предметными отношениями в целостный семантический комплекс, представляющий собой модель ситуации, задаваемую языковыми средствами данного языка» [3, с. 117]. В его понимании содержание – это ментальное образование, базирующееся на денотативных (референтных) структурах, в то время как смысл представляет собой мысль о действительности, интерпретацию того, что сообщается в тексте [4]. Тем не менее в ряде работ понятия семантики текста и его содержания разграничиваются. Так, семантическая структура связывается со знаковым характером текста, содержательная – с кодированием заключенной в нем информации, а смысловая – с декодированием, т.е. с восприятием произведения как целого [5, с. 69].

Важным для настоящего исследования является дифференциация таких прагматических видов информации, как содержательно-фактуальная (СФИ) и содержательно-концептуальная информации (СКИ), данная И. Р. Гальпериным. Второе из вышеуказанных понятий, имеющее непосредственное отношение к смыслу всего текста, определяется как «индивидуально-авторское понимание отношений между явлениями, описанными средствами СФИ, понимание их причинно-следственных отношений, их значимости в социальной, экономической, политической, культурной жизни народа. <...> Такая информация извлекается из всего произведения и представляет собой творческое переосмысление указанных отношений, фактов, событий, процессов» [6, с. 28].

Соотношение содержания и смысла зависит от типа текста. Можно согласиться с А. И. Новиковым в том, что «минимальное расхождение содержания и смысла характерно для текстов научной, технической тематики, для деловой прозы, а максимальное – для различных жанров художественных текстов» [4, с. 112]. Безусловно, наиболее сложным представляется процесс определения смысла художественного текста, который выявляется не только на основе лингвистических параметров, но и комплекса экстралингвистических факторов. Этой проблеме посвящен целый ряд работ в функциональной стилистике и лингвистике текста (А. Н. Васильева, Г. И. Богин, Н. А. Купина, Н. С. Болотнова и др.). Что касается таких типов текстов, как публицистические и разговорные, то их смысловые структуры и глобальные смыслы остаются наименее изученными.

Как следует из вышеприведенных определений смысла целого текста, он представляет собой авторскую интерпретацию текста, в большинстве случаев его словесная формулировка не дается в готовом виде. Вместе с тем можно предположить, что в каждом тексте есть некоторые поверхностные (т.е. формальные) признаки, сигнализирующие о его глубинных структурах, в том числе и об общем (глубинном) смысле. К таким признакам относятся конкретные слова, лексико-семантические группы слов, словосочетания или лексико-синтаксические структуры. Именно на таких признаках и могут базироваться алгоритмы распознавания главного смысла того или иного типа текста, на основе которого создаются компьютерные программы.

В нашей концепции коммуникативно-когнитивного моделирования текстовых жанров понятия семантики и прагматики текста рассматриваются

в неразрывном единстве [1], поскольку семантика текста без прагматики является неполной, а прагматика без семантики не существует. Семантика текста в нашей модели анализируется на основе понятия суперструктуры, предложенной Т. ван Дейком [7]. Хотя суперструктура постулируется этим автором как формальная схема дискурсивного жанра, в конкретном анализе текстов газетных новостей, наряду с формальными категориями, т.е. категориями композиции (*заголовок, резюме, лид* и др.), он использует семантические категории (*предшествующие события, фон, вербальные реакции* и др.). Используя понятие суперструктуры, мы наделяем его семантическими свойствами и определяем как общую структуру содержания текста, т.е. как семантическую структуру.

Конкретный анализ текстов показал, что каждый жанр обладает типовой суперструктурой, включающей семантические компоненты, характерные для подавляющего большинства текстов данного жанра [8; 9; 10; 11; 12]. При этом установлено, что семантические составляющие репрезентируются в текстах с помощью определенных языковых маркеров, что дает возможность их алгоритмического распознавания [8; 12] и построения сжатых аналогов исходных текстов – аннотаций и рефератов.

В прагматически-ориентированных алгоритмах анализа текста, например, поиска мнений или определения тональности текста, нами используется понятие *семантико-прагматическая структура*. Его основу составляет рассмотренное выше понятие семантической структуры, модифицированное с точки зрения прагматической значимости того или иного семантического компонента. Прагматическая значимость, или прагматическая «заряженность» семантического компонента суперструктуры определяется на основе совокупности критериев: количества высказываний, выражающих мнения; количества оценочных высказываний; степени эмоциональной насыщенности; наличия непосредственной связи с основной идеей текста. Так, анализ такого вида аналитической статьи, как проблемная статья, показал, что в ее суперструктуре наиболее прагматически нагруженными являются компоненты «описание проблемной ситуации» и «пути решения проблемы». При этом второй компонент содержит мнения не только автора текста (журналиста), но, главным образом, мнения других участников медийной коммуникации – представителей государственных, частных, общественных организаций. В связи с этим, смысл такой разновидности аналитической статьи можно рассматривать как коллективное мнение о путях разрешения проблемы.

В современной прессе аналитические статьи представлены и многими другими субжанрами. Так, распространенным видом являются статьи, которые публикуются под рубриками «Opinion», «Мнение», «Меркаванне», «Погляд» и др., часто подразделяемые на ряд подрубрик. Интересно отметить, что в текстах подрубрики «В стране / Страна» в белорусской прессе мнение автора (журналиста), как правило, выражено эксплицитно, хотя привлекаются точки зрения и других участников медийной коммуникации. В отличие от проблемных статей, в таких статьях рубрики «Мнение» авторская точка зрения превалирует, и она сигнализируется эксплицитными средствами: личные и притяжательные местоимения (*я, мой, мы, наши*), эллиптические струк-

туры с пропущенным подлежащим 1-го лица (*уверен, не верю, понимаю, думаю, рискую, смею, начал, обнаружила* и др.). Например: ... *И, кстати, я не права – интрига остается* (СБ Беларусь сегодня, 6.05.2016); *И здесь именно в музыке я вижу огромный потенциал* (СБ Беларусь сегодня, 18.05.2016).

Суперструктура таких текстов может быть как сходной, так и отличной от проблемной статьи. Например, в статье «Капитальные расчеты», опубликованной в подразделе «В стране» рубрики «Мнение» (СБ Беларусь сегодня, 5.05.2016) рассматривается проблемная ситуация, связанная с отчислениями граждан на капитальный ремонт жилья. Суперструктура этой статьи, как и большинство статей данной рубрики, посвященных внутренним проблемам страны, включает такие компоненты, как *описание проблемной ситуации, формулирование проблемы, выявление ее причин, пути решения проблемы*. Как было отмечено выше, в соответствии с установленной нами ранее семантико-прагматической структурой проблемной статьи одним из наиболее важных и прагматически нагруженных ее составляющих является компонент «пути решения проблемы». Именно в этой структурной части, как показывает анализ, содержится главный авторский смысл рассматриваемого текста, который в концентрированной форме обобщен в последнем предложении следующего текстового фрагмента, передающего мнение журналиста:

Но если мы собираемся выйти на 100-процентное возмещение затрат на ЖКХ, то уже не отдельные наши читатели, а все начнут задавать вопросы: сколько моих денег собрали и куда они ушли? К полному отчету надо бы готовиться. Кроме того, оценив накопленную сумму, жильцы, возможно, захотят сами выбрать подрядчика. <...> А если к жилью бережно относиться, можно и 50 лет не думать о комплексном капремонте. Должен быть выбор – вот главный посыл (СБ Беларусь сегодня, 5.05.2016).

Проведенный анализ статей раздела «В стране» рубрики «Мнение» показал, что алгоритм распознавания их авторского смысла должен быть инкорпорирован в детальный алгоритм выявления семантико-прагматического компонента «пути решения проблемы», который базируется на комплексе языковых маркеров этого компонента [1]. После установления текстового фрагмента, содержащего точку зрения автора на способы решения обсуждаемой проблемы, в качестве смысла всего текста выбирается его финальный отрезок, который, как правило, представляет обобщенное мнение автора и может дополнительно маркироваться такими языковыми средствами, как *главное; главный вопрос, главный посыл* и др.:

Главное, показать себя на мировом рынке. А лучшие – по-хорошему удивить (СБ Беларусь сегодня, 14.04.2015);

<...> самый **главный** вопрос – как убедить человека быть в ответе за собственную единственную и бесценную жизнь? (СБ Беларусь сегодня, 4.05.2015);

В отличие от раздела «В стране» рубрики «Мнение», статьи раздела «В мире» часто имеют иную прототипическую структуру: это последовательность аргументативных блоков, каждый из которых состоит из тезиса, аргументов различных типов и следствия. Например:

<Тезис> С приходом к власти президента Роухани Иран сделал ряд существенных шагов в этом направлении. <Аргумент> В первую очередь это касается снятия санкций в связи с иранской ядерной программой, <Следствие> что позволило Ирану существенно нарастить торгово-экономическое и технологическое сотрудничество с другими государствами (Рэспубліка, 3.05.2016).

Как показал анализ текстов, посвященных ситуации в мире, в большинстве случаев их структура характеризуется четкостью и логичностью развертывания, что выражается в частой оформленности аргументативных блоков в рамках отдельных абзацев. Подобное структурирование придает сходство данного субжанра аналитической статьи с научной статьей и облегчает его алгоритмический анализ. Как любой научный текст, рассматриваемые статьи часто завершаются четким выводом, который можно интерпретировать, на наш взгляд, как выражение главного смысла. При этом такой заключительный вывод может маркироваться не только финальной позицией в тексте, но и маркерами обобщения: *в целом, таким образом, ясно, очень важно; вынікі, у выніку и др.* Например:

Но ясно одно: движение иракских исламистов переживает бурный расцвет, и ожидать его политической кончины пока не приходится (СБ Беларусь сегодня, 8.07.2014);

Тем не менее может возникнуть эффект домино: кризис распространится на все компании, оказывающие услуги в нефтяной промышленности (СБ Беларусь сегодня, 13.01.2015);

*Можна чакаць, што ў **выніку** мінулых выбараў Еўрапарламент стане менш эфектыўным інструментам еўраінтэграцыі. А ўрады краін, дзе еўраскептыкі дасягнулі важных вынікаў, будуць вымушаны ўлічваць гэты факт у сваёй рабоце ў рамках Еўрасаюза (Звязда, 27.05. 2014);*

Таким образом, противостояние «либерального» и консервативного крыла иранского общества входит в новую стадию, и новый состав иранского меджлиса, по всей видимости, станет отражением этой непростой внутренней ситуации (Рэспубліка, 3.05.2016).

Как показывают приведенные примеры, вывод, или общий смысл статей данного субжанра, наряду с подведением итогов рассмотрения конкретной ситуации, часто включает и прогноз развития описываемого положения дел в будущем.

Несмотря на сходства с научными текстами, рассматриваемые газетные аналитические статьи сохраняют все характеристики публицистического стиля, в частности, его стилистические свойства. Этим объясняется возможность вербального оформления вывода в виде вопроса к читателям, вопросно-ответного блока или риторического вопроса, незавершенного высказывания, перифразы, парцелляции и др.:

Как мне кажется, для нас очень важно вовремя вписаться в концепцию нового Шелкового пути, которую так активно сейчас продвигает Китай. <...> Для нас это может означать инвестиции в новые рабочие места. Настало время собирать камни. Великие в том числе (СБ Беларусь сегодня, 5.05.2015);

Можжа, не трэба абагульненняў, калі іх атрымліваецца зрабіць толькі шляхам замоўчвання і скажэння фактаў? Можжа, трэба імкнуцца крытычна і аб'ектыўна ставіцца не толькі да асобных старонак гісторыі, а выкарыстоўваць такі падыход увогуле? (Звязда, 26.02.2016).

Таким образом, проведенный анализ показал, что формальное распознавание общего смысла медийных статей рубрик «Мнение», «Погляд» и др. в белорусской прессе целесообразно проводить на основе семантико-прагматических структур таких текстов и их языковых маркеров. Установленная в результате первого этапа работы алгоритма конкретная структура будет определять дальнейшее ветвление принципиального алгоритма и выявление наиболее вероятного вербального выражения глобального авторского смысла в структурном компоненте, имеющем наивысший прагматический потенциал.

ЛИТЕРАТУРА

1. Карпилович, Т. П. Семантико-прагматический интерфейс в алгоритмах анализа медиадискурса / Т. П. Карпилович // Вестник МГЛУ. Сер. 1, Филология. – № 4 (71). – 2014. – С. 28–35.
2. Слюсарева, Н. А. Проблемы функционального синтаксиса современного английского языка / Н. А. Слюсарева. – М. : Наука, 1981. – 207 с.
3. Новиков, А. И. Семантика текста и ее формализация / А. И. Новиков. – М. : Наука, 1983. – 215 с.
4. Новиков, А. И. Текст, его содержание и смысл / А. И. Новиков // Текст: Структура и функционирование. – Барнаул, 1997. – С. 108–113.
5. Каменская, О. Л. Текст и коммуникация / О. Л. Каменская. – М. : Высш. шк., 1990. – 152 с.
6. Гальперин, И. Р. Текст как объект лингвистического исследования / отв. ред. Г. В. Степанов. – изд. 8-е. – М. : Книжный дом «Либроком», 2014. – 144 с.
7. Дейк, Т. А. ван. Язык, познание, коммуникация / Т. А. ван Дейк. – М. : Прогресс, 1989. – 312 с.
8. Карпилович, Т. П. Моделирование процесса смысловой компрессии текста: когнитивно-дискурсивный подход / Т. П. Карпилович. – Минск : Изд-во МГЛУ, 2003. – 226 с.
9. Сысоева, Т. А. Прагматическая структура малоформатных медийных текстов / Т. А. Сысоева // Вестник МГЛУ. Сер. 1, Филология. – № 6 (49). – 2010. – С. 130–137.
10. Лешкевич, И. В. Семантическая структура газетных рецензий в британской и американской прессе / И. В. Лешкевич // Медiateкст и культура / Минский гос. лингвист. ун-т; под ред. Т. П. Карпилович. – Минск, 2015. – С. 26–34.
11. Шевцова, А. К. Дискурсивные маркеры компонентов суперструктуры английских и белорусских радиодискуссий / А. К. Шевцова // Актуальные проблемы лингвистики и гуманитарных наук : сб. науч. ст. по материалам междунар. науч.-практ. конф., Москва, 23 марта 2012 г. / Рос. ун-т дружбы народов ; редкол. : Е. В. Сафонова (отв. ред.) [и др.]. – М. : РУДН, 2012. – С. 288–299.
12. Глазко, П. П. Суперструктура публицистического текста как основа его автоматического реферирования / П. П. Глазко // Сб. тр. VI Междунар. науч. конф. «Прикладная лингвистика в науке и образовании», Санкт-Петербург, 5–7 апр. 2012 г. – СПб, 2012. – С. 60–63.

СОЧЕТАНИЯ ИДЕНТИЧНЫХ СОГЛАСНЫХ КАК ПОТЕНЦИАЛЬНАЯ ТРУДНОСТЬ РАСПОЗНОВАНИЯ СЛОВА В АНГЛИЙСКОЙ СПОНТАННОЙ РЕЧИ

Артикуляторные, акустические и перцептивные характеристики удвоенных согласных, или геминат, явились объектом многочисленных специальных исследований [1; 2; 3].

Общеизвестно, что геминаты представляют собой фонетически удлинённые согласные, при артикуляции которых происходит задержка размыкания по сравнению с одиночными согласными [4; 5]. В языках различают геминаты трех типов: лексические, ассимилятивные и составные (конкатенативные) [5]. Лексические геминаты являются частью лексикона и фонемного инвентаря (например, [pap:a] ‘mush’ в итальянском языке) [6]. Ассимилятивные геминаты являются результатом уподобления одного согласного предшествующему или последующему звуку на морфемном стыке внутри слова. Составные геминаты возникают при сочетании двух идентичных согласных на внутрисловных межморфемных или межсловных границах, например, *un+named* ‘безымянный’, *funname* ‘смешное имя’, *can name* ‘можно назвать’ [2]. Очевидно, что составные геминаты, особенно на межсловных границах, представляют собой случайные сочетания идентичных согласных, ограниченные дистрибутивными свойствами отдельных фонем в данном языке.

В лингвистической литературе лексические и ассимилятивные (исторические) геминаты относятся к «истинным», а составные геминаты описываются как «ложные». Длительность «истинных» геминат представляет собой часть репрезентации лексикона, в то время как длительность «ложных», постлексических, геминат определяется условиями результатом их реализации в потоке речи. «Истинные» геминаты отличаются от «ложных» тем, что даже на морфемном стыке они контрастивны [3]. Так, например, в корейском языке слова *mask* ‘маска’ и *exemption* ‘освобождение (от налога и т.п.)’ отличаются длительностью согласного, и о значении слова нельзя догадаться по морфемному составу. В английском языке, с другой стороны, длительность внутрисловной составной геминаты, например, /nn/ в слове *unnamed* ‘безымянный’, не влияет на смысл. Значение слова можно вывести на основании знаний о производной морфеме ‘un’ и глагольной морфеме ‘name’. Нужно заметить, что этот же аргумент относится и к ассимилятивным геминатам. Представляется в этой связи, что различие ассимилятивных и конкатенативных межморфемных геминат не имеет строгих критериев, и оба этих типа могут быть объединены в один – внутрисловные геминаты.

Традиционно дискуссионным вопросом в изучении геминации является их бифонемность или монофонемность, т.е. признание (в последнем случае) существования долгих и кратких согласных в языках. В отношении «ложных» геминат такой проблемы нет: бифонемность сочетания не ставится под сомнение. Задачи изучения этого типа удвоенных согласных заключается в установлении нормы удвоения, как с точки зрения его орфоэпической обязательности, так и с точки зрения его фонетически адекватного выражения.

Среди исследователей «истинных» и «ложных» геминат существуют разногласия относительно наличия/отсутствия различий между ними. К различительным фонетическим признакам геминат относят длительность предшествующего гласного, длительности смычки и время задержки голоса относительно момента взрыва согласного (VOT). А. Лахири и Дж. Ханкамер считают, что оба типа геминат (например, в бенгальском языке) реализуются одинаково и не различаются по названным признакам [7]. К аналогичным выводам приходит А. Лехисте, указывая на то, что в английском языке губно-губные смычно-взрывные «истинные», лексические, геминаты на внутрисловном межморфемном стыке /tappa/ и «ложные», постлексические, геминаты /lar real/ на межсловном стыке существенно не различаются по длительности смычки и средней амплитуде [8]. В то же время результаты новейших исследований свидетельствуют о том, что длительность геминат на межсловном стыке превышает длительность геминат на межморфемном стыке.

Противоречивые данные по длительности приводят ученых к необходимости поиска других, более надежных, по их мнению, признаков различения двух типов геминат. Так, Р. Ридуани отмечает, что при различении «истинных» и «ложных» геминат необходимо опираться на синтагматическую модель соотношения последовательности длительностей ‘гласный + согласный’, а не на парадигматические, сегментные, различия, такие как длительность смычки, VOT, амплитуду, или на соотношение длительности одиночного согласного к геминате [9].

Наряду с вышеназванной дифференциацией интерес исследователей вызывает и вопрос соотношения длительности геминат одиночных согласных. По экспериментальным данным П. Ладефогед и И. Мэддисона, длительность геминат в чтении в 1,5–3 раза превышает длительность одиночных согласных [5]. Отношение длительности одиночных согласных к длительности геминат варьирует в широких пределах в различных языках. Так, например, по данным П. Делятра, соотношение между длительностью одиночного согласного и геминаты в английском языке составляет 1:1.4, в то время как во французском языке оно равняется 1:1.19 [10]. Отмечается также, что длительность предшествующего гласного короче перед геминатой, чем перед одиночным согласным. Следует указать, однако, что сделанные авторами выводы относятся к материалу, состоящему из специально подобранных слов и словосочетаний, прочитанных носителями языка в изолированной позиции и в составе специально подобранных фраз.

Вместе с тем, имеющиеся в настоящее время доказательства специфики реализации фонологических единиц и их соединений в спонтанной речи дают основания утверждать, что удвоенные согласные в спонтанной речи существенно отличаются по вариативности и степени удлинения по сравнению с чтением. В более ранних работах не учитывалась и роль фразово-позиционных факторов в реализации удвоенных согласных, в том числе и характер межсловной связи (внутриакцентной или межакцентной). Представляется очевидным также, что необходимо дифференцированное изучение удвоенных согласных отдельно для всех типов фонем в подсистеме консонантизма, при этом внутри каждого типа согласных отдельно для глухих и звонких и далее для конкретных представителей типа. Названные аспекты

общей проблемы обуславливают актуальность дальнейших специальных экспериментальных исследований фонетической структуры удвоенных согласных. Одной из причин важности ее изучения, в том числе в английском языке, является необходимость получения надежных и полных фонетических данных о реализации удвоенных согласных для прикладных целей. Общеизвестно, в частности, что неадекватная длительность идентичных согласных на межсловных стыках является одним из типичных недостатков (отклонений) синтезированной речи, также как и речи носителей языка.

Как показывают наблюдения над английской спонтанной речью, квантитативная маркированность удвоенных согласных способствует разграничению слов и разборчивости речи. Если при этом признать, что разграничение слов происходит благодаря объективному различению между словоконечными и словоначальными аллофонами согласных, то требует доказательства сохранения ими своей специфики и ее «достаточности» для распознавания границ слов при сочетании идентичных согласных.

В этой связи возникает вопрос о минимальном пороге восприятия удвоенности по сравнению с восприятием удлиненности согласного, как, например, под влиянием экспрессивности. Так, в работе Л. В. Рускевич было показано, что удлинение щелевого согласного [s] всего на 10–20 % приводит к восприятию удлинения звука и, как следствие, восприятию экспрессивного произнесения слова [11]. Достаточно ли аналогичного увеличения длительности для восприятия удвоенности согласных? Предпринятое нами исследование позволит получить экспериментально-обоснованный ответ на заданный вопрос благодаря сопоставлению фонетической реализации сочетания идентичных согласных на межсловных стыках по сравнению с соединением позиционно различных аллофонов одиночных согласных, суммарная длительность которых может быть определена аргіогі для всех типов.

Предпринятое нами исследование проводилось на материале спонтанной речи 3 носителей языка (мужчин в возрасте от 40 до 55 лет), постоянно проживающих в южной части Англии и владеющих нормами английского произношения.

Из звучащих спонтанных высказываний были выделены слова (n=150), содержащие идентичные носовые сонанты [m], [n], боковой сонант [l] и щелевые согласные [s], [z] на межсловном стыке, а также слова, содержащие словоначальные и словоконечные аллофоны изучаемых одиночных согласных (n=150).

Акустический анализ проводился при помощи компьютерной программы Praat 4.0.35 в соответствии с принятой методикой. В задачу акустического анализа входило определение и сравнение квантитативных и амплитудных характеристик исследуемых удвоенных английских согласных по сравнению с одиночным согласным в сочетании согласный + гласный (Cg) или гласный + согласный (гC).

Согласно полученным данным, в спонтанной речи наблюдается значительная вариативность как между согласными, так и внутри каждого индивидуального согласного: диапазон вариативности достаточно широк и находится в пределах от 92 до 163 м/с для [m+m], от 80 до 211 м/с для [n+n], от 72 до 200 м/с для [l+l] и от 90 до 228 м/с для [s+s] (табл. 1).

Как видно из таблицы, носовые сонорные [m] и [n] не совпадают по длительности. Так, суммарная длительность [n] превышает длительность [m] на 10–15 %. Данная вариативность указывает на то, что каждую группу согласных, вплоть до каждого конкретного звука, следует изучать отдельно с учетом факторов фразовой и словесной просодии и находить коэффициенты удлинения для отдельных классов в соответствии со спецификой звуков.

Т а б л и ц а 1

Длительность геминат в спонтанной речи носителей языка

Согласный	Дикторы					
	Д1		Д2		Д3	
	t, m/c	V, %	t, m/c	V, %	t, m/c	V, %
[m] + [m]	123	23	100	8	112	20
[n] + [n]	146	22	120	10	131	12
[l] + [l]	136	18	119	36	121	13
[s] + [s]	141	24	157	10	138	23

В ходе эксперимента стимулы были далее разделены с учетом ударности/неударности слогов на межсловном стыке на три группы:

- безударный + безударный слог, например, <...> *from my* <...> ;
- безударный + ударный слог, например, <...> *from Minsk* <...>;
- ударный + ударный слог, например, <...> *Sam Morris* <...>.

Полученные результаты анализа представлены в табл. 2. Как видно из таблицы, варьирование длительности всех изучаемых удвоенных согласных на межсловных стыках обнаруживает регламентированный характер, определяемый акцентно-временной организацией английской (британской) произносительной нормы. Так, длительность геминат минимальна в сочетании безударного слога с безударным, несколько увеличивается в сочетании безударного слога с ударным и максимальна при соединении двух ударных слогов.

Выявленная устойчивая тенденция к увеличению длительности удвоенных согласных на межсловном стыке типа ‘ударный + ударный’ слог, безусловно, является предсказуемой и является доказательством значимости позиционно-просодических аллофонических модификаций согласных.

На следующем этапе исследования предполагалось проверить гипотезу о том, что суммарная длительность двух одиночных согласных (словоконечного и словоначального) в английской спонтанной речи не равна длительности геминаты. С этой целью из экспериментального материала нами были отобраны слова, содержащие аллофоны словоначальных и словоконечных изучаемых согласных, реализуемые в аналогичных фразовых условиях с геминатами: <...> *She graduated from my University*. <...> vs <...> *She graduated from it* <...> vs <...> *The car was waiting for me outside*. <...>

В табл. 3 представлены результаты сопоставительного анализа длительность геминат и одиночных согласных в спонтанной речи в безударной позиции.

Т а б л и ц а 2

Длительность геминат английской спонтанной речи
в зависимости от ударности/неударности слога в стыке

Согласный	Диктор	Тип стыка		
		безударный + безударный	безударный + ударный	ударный + ударный
[m] + [m]	Д1	99	133	150
	Д2	92	130	–
	Д3	93	131	158
[n] + [n]	Д1	101	153	–
	Д2	107	145	154
	Д3	–	131	–
[l] + [l]	Д1	120	129	–
	Д2	75	136	174
	Д3	97	127	–
[s] + [s]	Д1	103	136	146
	Д2	134	152	196
	Д3	103	142	–

Т а б л и ц а 3

Длительность геминат и одиночных согласных
в спонтанной речи в безударном слоге, м/с

Согласный	Длительность			Сумма длительности 'начальный+конечный'
	гемината	начальный	конечный	
[m]	92	50	60	110
[n]	107	54	68	122
[l]	75	49	61	110
[s]	134	109	114	223

Как видно из приведенных данных, длительность геминат превышает длительность одиночных согласных в 1,5–1,8 раза. У всех изучаемых согласных наблюдается увеличение суммарной длительности словоначального и словоконечного согласного при довольно большой степени вариативности данных показателей.

Основной результат эксперимента состоит в том, что два идентичных согласных на межсловном стыке всегда сохраняют свою фонемную самостоятельность, но сокращаются по длительности. Важным результатом является также выявление специфики каждого из элементов соединения в создании эффекта удвоения для разных типов согласных.

Как уже отмечалось, межсловный стык не планируется, а является свободным сочетанием сегментных единиц. Тем не менее, экспериментальные данные позволяют говорить о наличии у носителей языка ментального эталона фонетической реализации таких сочетаний для каждого

типа и каждого конкретного класса согласных. Специфика этой реализации заключается в дополнительной функции – усилить распознаваемость межсловной границы в условиях ее объективного усложнения при соединении идентичных согласных. Наряду с делимитативной функцией позиционного аллофонического варьирования согласных, специальное усиление и подчеркивание разъединения удвоенных согласных при их внешней фонетически полной интегрированности является важным средством оптимизации слухового восприятия речи.

ЛИТЕРАТУРА

1. *Mccarthy, J.* OCP Effects: Gemination and antigemination / J. Mccarthy // *Linguistic Inquiry*. – 1986. – Vol. 17. – P. 207–263.
2. *Kaye, A.* Gemination in English / A. Kaye // *English Today*. – 2005. – Vol. 21. – P. 43–55.
3. *Aoyama, K.* Cross-linguistic tendencies and durational contrasts in geminate consonants: examination of Guinaang Bontok geminates / K. Aoyama, A. R. Lawrence // *Journal of the International Phonetic Association*. – 2006. – Vol. 36. – № 2. – P. 145–157.
4. *Кодзасов, С. В.* Общая фонетика : учебник / С. В. Кодзасов, О. Ф. Кривнова. – М. : Рос. гос. гуманитар. ун-т, 2001. – 590 с.
5. *Ladefoged, P.* The sounds of the world's languages / P. Ladefoged, I. Maddieson. – Oxford; Cambridge : Blackwell Publ., 1996. – XXI. – 425 p.
6. *Payne, E.* Phonetic variation in Italian consonant gemination / E. Payne // *Journal of the International Phonetic Association*. – 2005. – Vol. 35. – № 2. – P. 153–189.
7. *Lahiri, A.* The timing of geminate consonants / A. Lahiri, J. Hankamer // *Journal of Phonetics*. – 1988. – № 16. – P. 327–338.
8. *Lehiste, I.* An instrumental study of consonant germination / I. Lehiste, K. Morton, M. Tatham // *Journal of Phonetics*. – 1973. – № 1. – P. 131–148.
9. *Ridouane, R.* Gemination in Tashlhiyt Berber: an acoustic and articulatory study / R. Ridouane // *Journal of the International Phonetic Association*. – 2007. – Vol. 37. – № 2. – P. 119–142.
10. *Delattre, P.* Pharyngeal features in consonants of Arabic, German, Spanish, French and American English / P. Delattre // *Phonetica*. – 1971. – Vol. 23. – P. 129–155.
11. *Рускевич, Л. В.* Взаимодействие просодических и лексико-семантических средств в выражении экспрессивности в современном английском языке (экспериментально-фонетическое исследование) : дис. ... канд. филол. наук : 10.02.04 / Л. В. Рускевич. – Минск, 2014. – 170 л.

А. В. Крывальцэвіч, Я. С. Качан, Ю. С. Гецэвіч

АПРАЦОЎКА КОЛЬКАСНЫХ ВЫРАЗАЎ У МАБІЛЬНЫМ ПРЫКЛАДАННІ ТУРЫСТЫЧНАЙ ТЭМАТЫКІ ЯК СРОДАК ВЫВУЧЭННЯ КАЛЬТУРНА-МАЎЛЕНЧЫХ АСАБЛІВАСЦЯЎ КРАІНЫ

Інфармацыйныя тэхналогіі сучаснасці прадстаўляюць сабой дынамічную вобласць, развіццё якой паўплывала на ўсе сферы жыцця чалавека. Але нельга не заўважыць, што інфарматызацыя адбываецца не толькі праз распаўсюджванне персанальных камп'ютараў, але і праз павышэнне якасці працы мабільных тэлефонаў, а больш дакладна – смартфонаў.

Згодна заяве кіраўніка каманды Пошука Google, Аміта Сінгхала (Amit Singhal), летам 2015 года колькасць мабільных пошукаў у Google упершыню перавысіла колькасць дэсктопных [1]. З гэтага можна зрабіць выснову, што павялічылся колькасць карыстальнікаў мабільнымі прыстасаваннямі, а гэта, у выніку дае павелічэнне выкарыстання мабільных прыкладанняў.

Мабільнае прыкладанне – гэта праграмае забеспячэнне прыкладнога характару, якое выкарыстоўваецца для аптымізацыі працы карыстальніка на мабільных прыстасаваннях.

Таму важна далучыцца да распрацовак мабільных прыстасаванняў у сферы турызму, якія могуць быць карыснымі як для жыхароў Рэспублікі Беларусь, так і для замежных грамадзян.

На падставе зместу дзяржаўнай праграмы развіцця турызму ў Рэспубліцы Беларусь, вядома, што турызм з’яўляецца адной з важнейшых сфер эканомікі [2].

У якасці прыкладу ў дадзеным даследаванні разглядаецца мабільнае прыкладанне *KrokApp*, якое распрацоўваецца ў лабараторыі распазнавання і сінтэзу маўлення АПП НАН Беларусі.

Канцэпцыя мабільнага прыстасавання накіравана на тры асноўныя мэты: экскурсійная прэзентацыя гарадоў без удзелу чалавека-экскурсавода, фарміраванне цікавасці да гісторыі Беларусі і стварэнне станоўчага іміджа краіны. Мэтавай аўдыторыяй з’яўляюцца жыхары Рэспублікі Беларусь і турысты, якія валодаюць англійскай мовай.

Структура мабільнага прыкладання *KrokApp* уключае ў сябе сукупнасць наступных старонак: старонка-лагатып; галоўнае меню, у якім ёсць магчымасць выбару мовы (беларуская, англійская, руская); старонка выбару тыпу маршруту (экскурсійны, кафэ і рэстараны або шопінг); старонка-пералік гарадоў (Мінск, Гродна, Брэст, Віцебск, Магілёў, Гомель); старонка выбару тыпу экскурсіі (тэматычная экскурсія, аглядная экскурсія); старонка-пералік кропак агляду для кожнага горада адпаведна; старонкі-агляды кропак агляду для кожнага горада адпаведна.

На дадзены момант створаны прататып, які можна спампаваць у Google Play Market па наступнай спасылцы: <https://play.google.com/store/apps/details?id=by.ssrlab.krokapp&hl=be>.

У прыкладанні быў прапрацаваны беларускамоўны падраздзел. Для апісання кропак агляду дададзены фотаздымкі месцаў ці будынкаў, друкаваны тэкст і гукавы файл.

Для змяншэння памеру мабільнага прыкладання па меры дабаўлення тэкстаў апісанняў у будучым плануецца агучваць тэксты з дапамогай сэрвіса, распрацаванага на базе лабараторыі распазнавання і сінтэзу маўлення – “Сінтэз маўлення па тэксце” [3].

Адной з асноўных праблем для сінтэзатара беларускага маўлення з’яўляецца ўспрыманне і агучванне незразумелых для машыны сімвалаў (лікі, скарачэнні, даты і г.д.) [3]. Таму аўтары дадзенага артыкула лічаць неабходным распрацаваць спіс правяраных нармалізаваных мностваў колькасных выказаў, які будзе пакладзены ў аснову аўтаматызаванага працэсу нармалізацыі ўсіх невядомых сімвалаў.

Крыніцай матэрыялаў сталі тэксты апісанняў кропак агляду, падрыхтаваныя для агучвання. Сабраны матэрыял быў аформлены як спіс сказаў з ненармалізаванымі выразамі па наступных семантычных катэгорыях:

1. Час (г., гг., ст., -х).
2. Памер і габарыты (м, км, га).
3. Выпадковыя дадзеныя (скарачэнні, знак нумару, абрэвіатуры і інш.).

Важна адзначыць, што “Сінтэзатар маўлення па тэксце” (Text-to-Speech RHP-based Synthesizer) ажыццяўляе агучванне тэксту на беларускай і рускай мовах [3]. Але карэктная апрацоўка спецыфічных токенаў (лікаў, дат, скарочэнняў, адзінак вымярэння, абрэвіатур і інш.) на дадзены момант адсутнічае.

У працэсе даследвання былі вылучаны галоўныя правілы, паводле якіх адбывалася экспертная генерацыя лікаў у колькасныя і парадкавыя лічэбнікі. З правіламі можна азнаёміцца ў папярэдніх артыкулах па гэтай тэме: Гецэвіч Ю.С. і інш. “Алгарытм і лінгвістычныя рэсурсы для нармалізацыі тэкстаў геаграфічнага дамена” [4], Гецэвіч Ю.С. “Мадэляванне і распрацоўка сістэм пошуку колькасных выказаў з адзінкамі вымярэння ў электронных тэкстах на беларускай і рускай мовах” [5].

Паводле гэтага алгарытму атрымаўся спіс нармалізаваных дадзеных, размеркаваных па семантычных катэгорыях (час, памеры і габарыты, выпадковыя дадзеныя) (табл. 1):

Табліца 1

Фрагмент спіса нармалізаваных дадзеных па семантычным катэгорыям

Катэгорыя	Выраз па-беларуску	Нармалізаваны сказ на беларускай мове
Час	Мінскі чыгуначны вакзал (да 1928 г. – Віленскі вакзал) – пасажырскі тэрмінал чыгуначнай станцыі «Мінск-Пасажырскі».	Мінскі чыгуначны вакзал (да тысяча дзевяццот дваццаць васьмага года – Віленскі вакзал) – пасажырскі тэрмінал чыгуначнай станцыі «Мінск-Пасажырскі».
Памер і габарыты	Вышыня Фарнага касцёла дасягае 50 м , што робіць яго прыкметным з любой кропкі горада.	Вышыня Фарнага касцёла дасягае пятдзсят метраў , што робіць яго прыкметным з любой кропкі горада.

У будучыні плануецца дадаць лінгвадыдактычную функцыю на вывучэнне беларускіх слоў для замежнікаў, якія валодаюць англійскай мовай. Яе працу можна апісаць двума сцэнарыямі.

Сцэнарый 1. На працягу экскурсіі карыстальніку прапаноўваецца азнаёміцца з беларускімі адпаведнікамі некаторых слоў (малюнак 1). Прыкладны сцэнар узаемадзеяння карыстальніка з мабільным прыстасаваннем наступны:

1. На экране паказваецца інфармацыя пра славутасць.
2. Прапаноўваецца націснуць на назву славутасці.

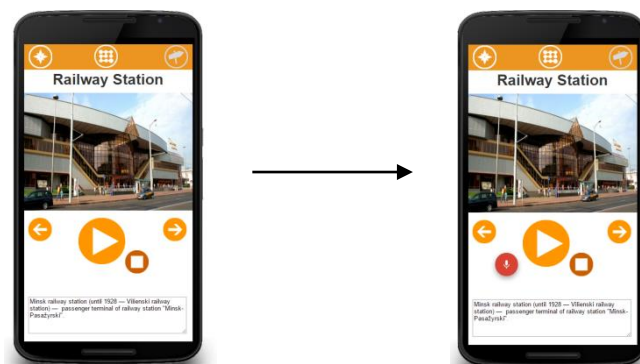
3. Пасля націскання карыстальнік слухае пераклад слова *Railway station* – ‘Вакзал’.

4. Прапаноўваецца паўтарыць за аўдыёгідам слова ‘Вакзал’. На правільнае прагаворванне даецца 3 спробы.

5. Падсумоўваюцца балы за правільна прагавораныя словы.

6. Карыстальніку выказваецца падзяка па заканчэнні экскурсіі, паказваецца агульная сума балаў, за якую ён можа атрымаць пэўны бонус.

Для аўтаматычнага агучвання беларускамоўных тэкстаў прапаноўваецца выкарыстанне сервісаў, распрацаваных лабараторыяй распазнавання і сінтэзу маўлення: Text-to-Speech PHP-Based Synthesizer (‘Сінтэзатар маўлення па тэксце’) [6].



Малюнак 1. Прыклад першага варыянта ўзаемадзеяння карыстальніка з мабільным прыкладаннем падчас вывучэння назваў беларускіх славутасцяў

Сцэнарый 2. Карыстальнік можа націснуць на назву славутасці (Загалолак) і пабачыць транскрыпцыю гэтага слова ці фразы ў міжнародным фармаце (International phonetic alphabet – IPA) (малюнак 2). Для гэтага будзе выкарыстаны сервіс “Transcription Generator” (‘Генератар транскрыпцый’), які быў створаны на базе лабараторыі распазнавання і сінтэзу маўлення [7].



Малюнак 2. Прыклад другога варыянта ўзаемадзеяння карыстальніка з мабільным прыкладаннем пры вывучэнні назваў беларускіх славутасцяў

Такім чынам, замежнік будзе мець магчымасць не толькі праслухаць экскурсію па горадзе, але і вывучыць назвы беларускіх рэалій. Экскурсія на англійскай мове можа быць прапанавана таксама і беларускім студэнтам адпаведных спецыяльнасцей для павышэння навыкаў аўдыявання ці падрыхтоўкі ўласнай экскурсіі для сваёй групы.

ЛІТАРАТУРА

1. *Zakrzewski, Cat.* Mobile Searches Surpass Desktop Searches At Google For The First Time [Electronic resource] / Cat Zakrzewski. – Mode of access : <http://techcrunch.com/2015/10/08/mobile-searches-surpass-desktop-searches-at-google-for-the-first-time/?ncid=rss#.54rwqr:fdAV>. – Date of access : 12.02.2016.
2. Государственная программа развития туризма в Республике Беларусь на 2011–2015 год [Электронны рэсурс]. – Рэжым доступу : <http://mst.gov.by/ru/programma-razvitiya-turizma>. – Дата доступу : 8.02.2016.
3. Text-to-Speech PHP-Based Synthesizer [Электронны рэсурс]. – Рэжым доступу : <http://www.corpus.by/tts3/>. – Дата доступу : 15.01.2016.
4. *Гецэвіч, Ю. С.* Алгарытм і лінгвістычныя рэсурсы для нармалізацыі тэкстаў геаграфічнага дамена / Ю. С. Гецэвіч, Я. С. Качан, С. І. Лысы [і інш.] // Открытые семантические технологии проектирования интеллектуальных систем = Open Semantic Technologies for Intelligent Systems (OSTIS-2016) : матер. VI Междунар. науч.-техн. конф., Минск, 18–20 февр. 2016 г. / БГУИР; пад рэд. В. В. Голенков (отв. ред.) [и др.]. – Минск : БГУИР, 2016. – С. 397–400.
5. *Гецэвіч, Ю. С.* Мадэляванне і распрацоўка сістэм пошуку колькасных выказаў з адзінкамі вымярэння ў электронных тэкстах на беларускай і рускай мовах / Ю. С. Гецэвіч, А. М. Скопінава, А. Ф. Есіс // Развитие информатизации и государственной системы научно-технической информации (РИНТИ-2013) : докл. XII Междунар. конф., Минск, 20 нояб. 2013 г. – Минск : ОИПИ НАН Беларуси, 2013. – С. 282–287.
6. Лабараторыя распазнавання і сінтэзу маўлення. Сэрвіс Text-to-Speech PHP-Based Synthesizer [Electronic resource]. – Mode of access : <http://corpus.by/tts3/>. – Date of access : 9.02.2016.
7. Лабараторыя распазнавання і сінтэзу маўлення. Сэрвіс “Генератар транскрыпцый” [Электронны рэсурс]. – Рэжым доступу : <http://ssrlab.by/archives/3286>. – Дата доступу : 10.02.2016.

А. В. Молоткова

СРЕДСТВА ВЫРАЖЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ И ФАКТОВ В ПИСЬМЕННОМ ТЕКСТЕ

Преобразование неструктурированной текстовой информации в хорошо организованные и структурированные данные и знания является одной из основных задач современных технологий *Data Mining, Text Mining, Information Extraction*. Указанные технологии положены в основу создания компьютерных систем, способных правильно извлекать из текстовых массивов различные объекты и факты. Например, даты каких-либо событий, имена, адреса и телефоны персон, наименования товаров и услуг, названия компаний, описания событий в новостных сводках, мнения, оставленные пользователями социальных сетей, гостевых книг или форумов, отзывы о разнообразных продуктах и услугах и т.д. Успешность функционирования подобных систем зависит как от самой извлекаемой информации, т.е. количества и типов извлекаемых объектов, фактов и связей между ними, так и от способа представления и средств обработки полученных данных и знаний,

что непосредственно определяет класс решаемых компьютером задач: идентификации объектов, выявления и анализа фактографической информации, семантического поиска, экспертных решений, ответа на запросы, выраженные на естественном языке и т.п.

Несмотря на разнообразие существующих методов извлечения информации из неструктурированных текстовых массивов и других типов данных, до сих пор не решены ключевые проблемы информационного поиска, связанные с автоматическим построением баз данных и знаний. Наименее разработанными являются такие области, как извлечение знаний из открытых информационных ресурсов, поиск локализованных во времени фактов, а также извлечение именованных сущностей, т.е. имен и фамилий людей, названий организаций, географических названий и т.п. Проблема извлечения локализованных во времени фактов состоит в автоматическом определении временных границ. Она имеет достаточно хорошие решения в случае точного указания дат, а также при наличии взаимоисключающих фактов. Однако в реальных текстах временные границы часто задаются нечетко, например, «в детстве», «в старости» и т.д. Предположение о взаимном исключении фактов даже в классической задаче *Marriage Problem* не всегда истинно (как, например, в странах, где узаконены полигамные отношения). Проблема извлечения именованных сущностей связана, прежде всего, с тем, что даже в большом текстовом массиве они могут встречаться редко. Вторая трудность заключается в открытости информационного ресурса. Если крупные географические объекты относительно полно покрываются газетами, то другие именованные сущности достаточно быстро могут появиться в корпусе текстов, однако их регистрация в списках требует значительного времени. Третья проблема извлечения именованных сущностей сводится к проблеме снятия омонимии. Таким образом, несмотря на все существующие подходы к автоматическому извлечению данных и знаний из больших массивов неструктурированных данных, до сих пор не решены важные проблемы информационного поиска.

Английский термин *Named Entity (NE)* – именованная сущность – был впервые использован в 1995 году в рамках конференции *MUC (Message Understanding Conferences)* в связи с постановкой задачи автоматического распознавания именованных сущностей [1]. Именованной сущностью считается слово или словосочетание, предназначенное для конкретного, вполне определённого предмета или явления, выделяющее этот предмет или явление из ряда однотипных предметов или явлений. Именованная сущность обязательно имеет референт, т.е. объект внеязыковой действительности, подразумеваемый автором конкретного речевого отрезка. Референт может принадлежать реальному миру, например, фамилия, имя и отчество (ФИО) конкретного человека или название конкретной организации, либо вымышленному миру, например, быть персонажем художественного произведения.

В письменном тексте именованные сущности представлены объектами и их атрибутами. Под объектами понимаются либо персоны, имеющие такие атрибуты, как фамилия, имя, отчество, должность, звание и т.д., либо органи-

зации и топонимы, характеризующиеся такими атрибутами, как представленное в полном или сокращенном виде название предприятия, его адрес, направление деятельности и т.д. Именованные сущности характеризуются следующими типами признаков:

1) признаками уровня слова (N-граммами, суффиксами, префиксами, частями речи и т.д.);

2) признаками уровня текста (акронимами, позицией слова в предложении, наличием слова в заголовке и т.д.);

3) дополнительной информацией (газетными, словами-указателями, например, *Inc.*, *Corp.*, стоп-словами, словами с капитализацией, которые не являются именованными сущностями и т.п.) [2].

Выделяют три категории атрибутов:

1) атрибуты, характерные непосредственно для текста документа или статьи, например, заголовок, дата создания документа, исполнитель документа;

2) атрибуты, характерные исключительно для лиц, например, год рождения, национальность, особенности внешности человека;

3) атрибуты, свойственные двум отмеченным выше категориям признаков, например, адрес, номер и марка машины и т.д.

Необходимо учитывать возможные варианты представления именованных сущностей в тексте. Такие типичные объекты как ФИО, дата, адрес, должность приводятся к одному стандартному виду. При выделении цепочек типа ФИО важным является отождествление формально различных ФИО в одном тексте. Исходя из предположения, что в тексте вряд ли могут упоминаться два лица с одним и тем же именем и фамилией, но являющиеся разными людьми, можно соединить неполные ФИО с более полными вариантами. Например, именованные сущности *А. В. Петров*, *Александр Петров* и *Александр Владимирович Петров* будут считаться относящимися к одному человеку [3]. Следует отметить, что выявление объектов осуществляется не только с учетом их кратких наименований (например, соотнесение отдельных фамилий или имен с полным ФИО), но также с учетом анафорических ссылок (указательных и личных местоимений, например, *этот человек*, *он*) и определений (например, *мэр Москвы Лужков* идентифицируется со словами *мэр*, *Лужков*). Такая информация выражается грамматически правильно записанными последовательностями слов и символов, которые называется значимыми компонентами текста. В состав значимых компонентов входят информационные и вспомогательные слова. Информационные слова определяют объекты и атрибуты. Здесь важную роль играют слова-классификаторы, наличие которых указывает на присутствие соответствующей информации. Например, в определенном типе текста, где важными являются приметы людей, слово *рыжий* при наличии в непосредственной близости лица указывает на то, что речь идет о внешнем виде человека. В другом типе текста слово *московский* указывает на географическое положение. К вспомогательным словам относятся те лексические единицы, без которых значимые компоненты не теряют своей сущности. К ним относятся предлоги, знаки пунктуации и так называемые шаблонные слова, указывающие на положение соответствующих информационных слов в тексте. С точки зрения их выяв-

ления значимые компоненты условно делятся на жесткие и мягкие. Жесткие значимые единицы состоят из фиксированного числа позиций или слов, например, ФИО и дата. У мягких компонентов количество позиций переменное, например, приметы человека или признаки организации. В зависимости от используемого способа выделения адрес может рассматриваться и как жесткий, и как мягкий значимый компонент. С грамматической точки зрения значимые компоненты состоят из обязательных элементов (именные слова, прописные буквы, части речи, числа) и вспомогательных элементов.

В русскоязычных новостных текстах для обозначения персон обычно используется имя (иногда также отчество) и фамилия или же просто фамилия. С учетом принципа подбора минимального языкового материала в качестве контекстуального синонима используется только фамилия. При этом необходимо проверять такие синонимы на наличие в пре- или постпозиции подходящего соответствующему объекту имени. Это может быть непосредственно имя, сочетание имени и отчества, инициалы. По умолчанию фамилия без имени идентифицируется с тем же объектом, с которым идентифицируется эта же фамилия в сочетании с именем. Более сложный анализ необходим, если в тексте упоминается несколько персон, носящих одну и ту же фамилию (например, *Мишель Обама* и *Барак Обама*). Исключение из описанного правила составляют широко известные персоны, для обозначения которых могут использоваться только фамилии. В данном случае применимо следующее правило «непротиворечивости имени»: если в тексте нет сочетания имени и фамилии-синонима, в котором имя не соответствует имени объекта, то считать данную фамилию верифицированным синонимом для данной персоны. Кроме фамилий, в качестве синонимов (которым не нужна проверка на корректность имени) могут использоваться: псевдоним, сценическое имя или прозвище человека (например, *Борис Акунин*, *Витас*); сочетание имени и фамилии (это актуально, когда существует несколько варианта правописания имени или когда у человека есть несколько имен и могут использоваться их разные, например, *Абдалла ибн Абдель Азиз Ал Сауд*, *Абдуллах бин Абдул Азиз ал Сауд*, *Абдель Азиз Аль-Сауд*); имя человека и его статус (*Королева Елизавета*, *Патриарх Московский и Всея Руси Кирилл*); имя иностранной персоны, написанное латиницей (*Britney Spears*).

В наши дни в новостных сообщениях освещаются самые разные события, относящиеся к различным сферам: политической, социальной, экономической и многим другим. В зависимости от предметной области и тематики, к которой относится текст сообщения, в нем используются определенные факты, например, «визит», «встреча», «авария», «смерть», а также их атрибуты, например, «участник встречи», «место встречи», «дата начала визита», «место происхождения». Тексты новостных сообщений о дорожно-транспортных происшествиях написаны на основе официальной информации, предоставленной полицией, и могут включать высказывания-свидетельства очевидцев и пострадавших. В отличие от русскоязычных новостных сводок об авариях в англоязычных текстах данного типа очень часто представлена личная информация о родственниках пострадавших: пол, возраст, фамилия и имя. Например, *Raul Pachecho (a 63-year-old who works nearby)*; *Wendy*

Mateo – a 39-year-old who was hurt; Three others, including 23-year-old Ilhajjam Hamid and 42-year-old Henriquez Wellington, Yuko Hopkins, a 32-year-old. Чаще всего, указателями персон являются такие атрибуты как *police officer (police, cops, officers), driver, sources, witnesses, authorities, a Muslim woman, vehicle's passenger, a driver of a BMW*. В текст новостной сводки об аварии включают данные о месте происшествия, например, название штата, города, местности, улицы (*New Jersey, Atlantic City, on Broadway, West 62nd Street*) и т.д. Как отмечалось выше, именованные сущности и факты напрямую связаны с тематикой новостного сообщения. Например, в текстах данного типа преобладают такие слова, указывающие на именованные сущности, как, например, *car accident, crash, incident, license plates, scene, suffered, car, collision, hospital, injuries*, а также слова, обозначающие их атрибуты: *chain-reaction, Bellevue, non-life threatening* и др.

Автоматическое извлечение именованных сущностей и фактов из письменных текстов может проводиться в рамках следующих подходов [4, с. 252–254].

1. Извлечение сущностей и фактов на основе признаков подразумевает наличие фиксированного набора признаков и весов использования этих признаков в контексте извлекаемых элементов текста. Для каждого извлекаемого элемента определяется характеризующий его вектор признаков. Процесс извлечения сводится к распознаванию некоторого сегмента текста на основе вероятностного анализа признаков, обнаруженных в контексте этого сегмента. Недостатками данного подхода является использование ограниченного размера контекста (как правило, не более 2–3 слов), необходимого для обеспечения требуемой точности извлечения. Учет контекста большего размера приводит к снижению полноты распознавания и к увеличению размера необходимой репрезентативной выборки, на которой собирается статистика для расчета оценок вероятностей.

2. Извлечение сущностей и фактов на основе ядер устраняет часть недостатков предыдущего подхода за счет алгоритмического определения меры подобия между представлениями сопоставляемых текстовых сегментов. Суть данного подхода сводится к замене скалярного произведения векторов, отражающих признаковое представление распознаваемых элементов, некоторой функцией – ядром. Такая функция задается алгоритмически и учитывает более сложное представление распознаваемых элементов и их контекстов, как правило, – древовидное, описывающее синтаксическую структуру текстового сегмента. Для древовидных представлений расчет ядра, чаще всего, сводится к сопоставлению всех вложенных деревьев в исходные. Недостатком данного подхода является высокая сложность расчета ядер и определения синтаксической структуры текстового сегмента.

3. В основе подхода, основанного на сопоставлении образцов, лежит понятие «образец» и правила сопоставления образцов с фрагментами текста. Образцы представляют собой цепочки ограничителей (символы, слова, части речи и семантические классы), являющиеся своего рода лексическими шаблонами текстовых сегментов. В этом отношении данный подход аналогичен ядерному подходу при условии, что текстовые сегменты имеют «плоское» представление, а не древовидное.

4. Подход, основанный на фразовых образцах, является своего рода компромиссом между подходом, основанным на ядрах, и подходом, основанным на сопоставлении текстовых образцов. Текстовые сегменты, также как и в основанном на ядрах подходе, представляются в виде деревьев, отражающих результаты синтаксического анализа предложений. Однако вместо сложного расчета ядер в данном случае выполняется более простая процедура сопоставления сегмента текста с синтаксическим шаблоном. Чаще всего, для отражения синтаксических связей между единицами текста используются контекстно-свободные грамматики, которые позволяют оценить вероятность применения того или иного правила для конкретного фрагмента текста и выбрать правило с максимальной степенью вероятности. При этом на некоторой обучающей выборке текстов, предварительно размеченной человеком, осуществляется ручное составление формальных правил и вычисление оценок вероятностей их употребления. Разметка содержит также указания на правила грамматики, которые необходимо применять при извлечении данных элементов.

ЛИТЕРАТУРА

1. Кузнецов, И. Методики выявления объектов и связей, заданных в неявном виде [Электронный ресурс] / И. Кузнецов. – Режим доступа : http://www.dialog-21.ru/digests/dialog2012/materials/pdf/%D0%9A%D1%83%-D0%B7%D0%BD%D0%B5%D1%86%D0%BE%D0%B2_%D0%98_%D0%9F.pdf. – Дата доступа : 18.03.16.
2. Котельников, Д. С. Итерационное извлечение шаблонов описания событий по новостным кластерам / Д. С. Котельников, Н. В. Лукашевич // Тр. XIV Всерос. науч. конф. «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» RCDL–2012, Переславль-Залесский, 15–18 окт. 2012 г. – Переславль-Залесский, 2012. – С. 292–298.
3. Голованова, Е. И. Соотношение естественного и искусственного начал в языке профессиональной коммуникации / Е. И. Голованова // Языки профессиональной коммуникации : матер. Междунар. науч. конф. – Челябинск, 2003. – С. 26–33.
4. Кормалев, Д. А. Обобщение и специализация при построении правил извлечения информации / Д. А. Кормалев // Тр. X нац. конф. по искусственному интеллекту «КИИ–2006», Обнинск, 25–28 сент. 2006 г. : в 3 т. – М. : Физматлит, 2006. – Т. 2. – С. 572–579.

Э. В. Нагарнович

АВТОМАТИЧЕСКАЯ КОРРЕКЦИЯ ОДНОБУКВЕННЫХ ОПЕЧАТОК ВО ФРАНКОЯЗЫЧНЫХ ЗАПРОСАХ ПОЛЬЗОВАТЕЛЕЙ ИНТЕРНЕТ

Ошибки до сих пор остаются одним из самых малоисследованных лингвистических объектов, хотя почти все ученые, занимающиеся вопросами речевой деятельности, сталкиваются с ними и фиксируют их. В письменном тексте могут быть допущены орфографические ошибки, в результате которых образуются комбинации символов, не являющиеся словами данного языка; семантические ошибки, когда правильные слова переходят в другие

правильные слова, но совсем иные по смыслу; грамматические ошибки, когда слова переходят из одной грамматической формы в другую; пропуски отдельных слов и групп слов; ошибки в знаках препинания и полиграфическом оформлении текста; формульно-цифровые ошибки. Поэтому кроме средств орфографической проверки каждая система автоматической обработки текста должна быть оснащена средствами автоматизированного или автоматического исправления выявленных ошибок.

В статье представлена процедура моделирования автоматической коррекции орфографических ошибок (опечаток) в поисковых запросах франкоязычных пользователей без учета контекста. Материалом исследования послужили однословные и многословные запросы, взятые из лога запросов франкоязычных пользователей системы *Goldfire*, предоставленного компанией *IHS Global*. Данная система ориентирована на семантический поиск технической информации, сведений и данных в области химии, физики, биологии, медицины, строительства и транспорта.

В процессе анализа лога запросов были выделены наиболее типичные ошибки, приведенные в таблице 1. Достаточно распространенным типом ошибок, имеющим вид «замена буквы», является пропуск диакритических знаков в однословных и многословных запросах: *é* – в 115 словах из 800 слов проанализированных запросов; *è* – в 9 словах, в том числе в слове *systeme* вм. *système*, *ô* – в 13 словах, в том числе в слове *côlon* вм. *côlon*, *â* – в 3 словах. Например, *declencheur magnétique* вм. *déclencheur*, *augmentation de la conductivité électrique de poly-mere* вм. *polymère*, *controle dimensionnel de courbes* вм. *contrôle*, *distance mini-male entre les cables secondaire et les barres de puissance* вм. *câbles*.

Для создания эталонного словаря из лога системы *Goldfire* были выбраны франкоязычные запросы, не содержащие ошибок. Количество единиц в эталонном словаре составило 180 запросов со средней длиной слова, равной 12 буквам. В использованный для распознавания символов алфавит были включены буквы французского алфавита (A-Z, *Éé, Ââ, Êê, Îî, Ôô, Ûû, Àà, Èè, Ùù, Êë, Ïï, Üü, Ýÿ, Çç, æ*), а также цифры (0 – 9). Размер алфавита составил 51 символ. Эталонный словарь вошел в лингвистическую базу данных системы автоматической коррекции орфографических ошибок в запросах пользователей.

Т а б л и ц а 1

Типичные ошибки, выделенные из лога запросов франкоязычных пользователей системы *Goldfire*

№ п/п	Тип ошибки	Количество слов в поисковом запросе	Пример
1.	Вставка буквы	одно	<i>operation => opération</i>
		два	<i>constraste image</i> <i>constraste => contraste</i>
		три и более	<i>élément chauffant pplat sur sol gel</i> <i>pplat => plat</i>

2.	Замена буквы	одно	<i>polyphosphotiques => polyphosphoriques</i>
		два	<i>poluution bactérienne poluution => pollution</i>
		три и более	<i>diagraphie non nucleaire nucleaire => nucléaire</i>
3.	Перестановка двух соседних букв	одно	<i>porject => project</i>
		два	<i>marjoliane grange marjoliane => marjolaine</i>
		три и более	<i>contrôle d'accès à un bâtiment par potrail potrail => portail</i>
4.	Пропуск буквы	одно	<i>maitenance => maintenance sol-el => sol-gel</i>
		два	<i>lamele colle lamele => lamelle</i>
		три и более	<i>tranformation d'un liquide en solide tranformation => transformation</i>

Предложенная в статье формальная модель системы коррекции ошибок (опечаток) в запросах франкоязычных пользователей ориентирована, в основном, на однобуквенные искажения слов. Кроме однобуквенных опечаток известен ряд часто встречающихся не однобуквенных ошибок, например перестановка согласных через гласную (*émilination*). В рамках разработанной формальной модели обрабатываются некоторые не однобуквенные опечатки, однако их конкретные типы не рассматриваются, поскольку при практической реализации системы эти типы могут выбираться произвольно. Общая стратегия выбора того или иного способа поиска и исправления ошибок зависит от количества слов в запросе пользователя. Если запрос состоит из одного слова, то оно сопоставляется с каждым словом как однословных, так и многословных запросов из базы данных. Например, если введен однословный запрос *logociel* (вм. *logiciel*), то сначала компьютер проверит вхождение этого слова в список однословных запросов. Если совпадения нет, компьютер сопоставит введенный запрос со словами списка многословных запросов. В эталонном корпусе он найдет запрос *logiciel de travail collaboratif* и исправит *logociel* на *logiciel*. В случае ввода пользователем двухсловного запроса он сопоставляется со всеми двухсловными запросами базы данных. Если совпадения не найдены, компьютер ищет наличие каждого слова запроса в словаре двусловий. Такая процедура помогает достичь более высокой точности коррекции опечаток. Для запросов, состоящих из трех и более слов, была выбрана стратегия исправления ошибок, используемая при восстановлении однословных запросов. При этом каждое слово запроса сопоставляется с каждым словом лингвистической базы данных. Например, в запросе *cooment éviter la cristallisation d'urée* ошибку в слове компьютер исправит, опираясь на эталонный запрос *comment réduire le cholestérol*.

Отмеченные выше правила были положены в основу создания алгоритма автоматического обнаружения и коррекции орфографических ошибок в запросах франкоязычных пользователей. Рассмотрим несколько конкретных примеров, иллюстрирующих работу разработанной формальной модели. Например, в процессе обработки запроса *contrôle d'accès à un bâtiment par*

potrail, последовательно сравнивая слова запроса с единицами эталонного словаря, компьютер классифицирует слово *potrail* как ошибочное, поскольку такое слово в словаре отсутствует. Он определяет длину данного слова (семь символов). Следующее действие системы связано с выделением первого и последнего символов слова. Ими будут, соответственно, *p* и *l*. Осуществляя поиск по словарю, компьютер выделяет из него слово длиной в семь символов, начинающееся и заканчивающееся указанными выше буквами. Это слово *portail*. Следующее действие системы заключается в сравнении анализируемого слова с найденным словом эталонного словаря слева направо и справа налево. При этом компьютер определяет количество несовпадающих букв. В данном случае оно равно двум. Если в словаре есть слова с таким же количеством символов и аналогичными начальной и конечной буквами, например, *parasol*, система выбирает слово, в котором количество не совпавших букв меньше. Сравнив слова *parasol* и *potrail*, где количество не совпавших символов равно пяти, компьютер определяет слово *portail* как единственно правильный вариант и заменяет им слово с ошибкой.

Рассмотрим процедуру автоматической обработки однословного запроса *lamelle*, в котором пропущена буква. Такое слово в эталонном словаре отсутствует, поэтому считается ошибочным. Длина слова равна шести символам, а первая и последняя буквы, соответственно, *l* и *e*. В словаре могут быть слова с такими начальными и конечными параметрами, но количество не совпавших букв будет больше исходного. Поэтому, уменьшив количество букв в анализируемом слове на единицу (пять символов) компьютер находит в эталонном словаре слово с аналогичными параметрами. Это слово *laque*. Оно заносится в список 1. Следующим действием системы будет увеличение количества букв в анализируемом слове на единицу (семь символов). С такими параметрами в эталонном словаре есть слово *lamelle*. Оно заносится в список 2. Система сравнивает анализируемое слово *lamelle* со словами из списка 1 и списка 2. В результате сравнения остается вариант *lamelle*, и компьютер заменяет слово с ошибкой на найденный в эталонном словаре вариант.

В следующем запросе *diagraphie non nucleaire* система выявляет ошибочное слово *nucleaire*. Определив длину этого слова (девять символов), а также первую и последнюю буквы (*n* и *e*), компьютер с опорой на данные признаки выделяет из эталонного словаря слова *nucléaire*, *normative*, *naissance*, *nombreuse*. Далее система сравнивает анализируемое слово с каждым из перечисленных выше четырех слов слева направо и справа налево, определяя при этом количество несовпавших букв. В случае со словом *nucléaire* количество несовпавших букв равно единице, со словом *normative* – шести, со словом *naissance* – шести, со словом *nombreuse* – семи буквам. Компьютер определяет слово *nucléaire* как единственно верное и заменяет им ошибочный вариант.

Разработанная формальная модель системы автоматической коррекции ошибок в поисковых запросах франкоязычных пользователей была запрограммирована на языке *Ruby*. В ходе компьютерного эксперимента в систему было введено 200 франкоязычных поисковых запросов, состоящих из разного количества слов и содержащих разные типы ошибок в разных позициях слова. Фрагменты компьютерного эксперимента представлены на рис. 1–5.

```
130 ruby_script>./main_script
predictive
corrected query: prédictive
sol-el
corrected query: sol-gel
torrefaction
corrected query: torréfaction
raccordemnt
corrected query: raccordement
```

```
130 ruby_script>./main_script
copronickel
corrected query: cupronickel
equilibre
corrected query: équilibre
maintenance
corrected query: maintenance
operation
corrected query: opération
plumero
corrected query: plumeau
```

Рис. 1. Исправление ошибок в однословных запросах

```
130 ruby_script>./main_script
achine à sérigraphie
corrected query: machine à sérigraphie
bils vibrantq
corrected query: bols vibrants
colle vynylique
corrected query: colle vinylique
contact elastique
corrected query: contact élastique
```

```
130 ruby_script>./main_script
couteaux de contatc
corrected query: couteaux de contact
panneau de controle
corrected query: panneau de contrôle
station de pompae
corrected query: station de pompage
acydes polyphosphoriques
corrected query: acides polyphosphoriques
```

Рис. 2. Исправление ошибок в двухсловных запросах

```
130 ruby_script>./main_script
assemblage modulaire tolerie
corrected query: assemblage modulaire tôlerie
comment attenuer les vibrations
corrected query: comment atténuer les vibrations
comment eviter la cavitation
corrected query: comment éviter la cavitation
conception d'une poupée hydraustatique
corrected query: conception d'une poupée hydrostatique
```

```
130 ruby_script>./main_script
marqueur du cancer du cölon
corrected query: marqueur du cancer du côlon
materiaux sandwich aluminium
corrected query: matériaux sandwich aluminium
panneaux solaires thermiqes
corrected query: panneaux solaires thermiques
synthese du propane dithiol
corrected query: synthèse du propane dithiol
```

Рис. 3. Исправление ошибок в многословных запросах

```
130 ruby_script>./main_script
oocydasion
corrected query: oxydation
cholasterol
corrected query: cholestérol
consummation
corrected query: consommation
serigrapie
corrected query: sérigraphie
```

```
130 ruby_script>./main_script
anjectio
corrected query: injection
energetique
corrected query: énergétique
vibrassions
corrected query: vibrations
declenchémant
corrected query: déclenchémant
```

Рис. 4. Исправление нескольких ошибок в одном слове

```
130 ruby_script>./main_script
instrumntation nucleaire
corrected query: instrumentation nucléaire
etagement de la combustio
corrected query: étageement de la combustion
pannneau de controle
corrected query: panneau de contrôle
poluution bacterienne
corrected query: pollution bactérienne
```

```
130 ruby_script>./main_script_2
coment reduire le cholesterol
comment réduire le cholestérol
compositio des pudres luminescentes
composition des poudres luminescentes
comment tmesurer la toxicite d'un fumigene
comment mesurer la toxicité d'un fumigène
amortisseur de presion
amortisseur de pression
cooment éviter la cristalysation d'uree
comment éviter la cristallisation d'urée
```

Рис. 5. Исправление ошибок в каждом слове многословного запроса

Основные результаты тестирования компьютерной системы сводятся к следующему.

1. 90 % запросов, содержащих орфографические ошибки, были обработаны правильно.

2. Примеры 10 % не обработанных запросов и связанные с этим действия компьютера отражены в табл. 2.

3. На поиск и исправление всех ошибок компьютер потратил почти 22 секунды, то есть приблизительно 0,1 секунды на коррекцию одного запроса.

Примеры не обработанных системой запросов, содержащих ошибки

№ п/п	Действие компьютера	Пример	Кол-во запросов	Комментарий
1.	Исправление не произведено	<i>abbatement mécanique bils vibrants conductibilité thermique conseve vitamine constraste image declencheur magnétique echauffement excessif echauffement primaire electro précipitation émission acoustique etat de l'art contrôle non destructif marjoliane grange poluution bactérienne prevenir l'arthrose recuperer l'énergie rsique chimique systeme d'écriture acydes polyphosphoriques aviasion civile</i>	19	10 из 19 запросов на восстановление диакритиков
2.	Неверная замена	<i>comment detecter un arc serie dans une installation => comment détect-er un arc servie dans une installa-tion (надо: comment detecter un arc sé-rie dans une installation)</i>	1	В словаре есть пример, который повлиял на исправление ошибки: <i>comment pétrir une pâte et retir-er les outils ay-ant servie au pé-trissage</i>

Представленная формальная модель может быть использована в качестве дополнительного модуля интеллектуальной информационно-поисковой системы для обнаружения и коррекции орфографических ошибок и опечаток в словах поисковых запросов франкоязычных пользователей.

И. И. Панова, А. Д. Коваленко

ПАРАМЕТРИЧЕСКИЕ ЗНАЧЕНИЯ ПРОСОДИЧЕСКОЙ ДИСТАНТНОСТИ ОРФОЭПИЧЕСКОГО СТАНДАРТА АНГЛИЙСКОГО ЯЗЫКА В ШОТЛАНДИИ

Описание объективной реальности функционирования современного английского языка с точки зрения единства его орфоэпически нормативной заданности, равно как и определения объема и пределов ее вариативности, как правило, опираются на всесторонний анализ разных типов языковых реализаций, в том числе в его различных этнически и культурно сложившихся сообществах.

В британском варианте современного английского языка признается несколько функционирующих литературных стандартов на основе региональных признаков их закреплённой нормативности, степень и характер вариантности которых максимально проявляется в произношении. Так, по данным фундаментальных исследований, направленных на систематизацию признаков трех основных региональных типов произношения английского языка (северно-ирландского, северно-английского, шотландского) на территории Великобритании признается наличие его четырех фонологических стандартов наряду с общенациональным типом произношения, наиболее многочисленным среди них является вариантная регионально ограниченная норма произношения английского языка в Шотландии. Согласно переписи населения 2001 года, носителями шотландской произносительной нормы (Scottish Standard English, далее SSE) являются более 5 миллионов человек.

В исследовании Ж. Дюрана, Дж. Стюарт-Смит и других лингвистов [1; 2] на сегментном уровне установлены как системные фонологические, так и ярко выраженные фонетические различия в шотландском стандарте английского языка по сравнению с общенациональной британской орфоэпической нормой (Received Pronunciation, далее RP), которые определяют его региональную нормативность, обязательность и функциональную адекватность.

Имеющиеся во всех сравнительных исследованиях указания [3; 4; 5; 6] на то, что своеобразие регионального шотландского произносительного стандарта также включает его просодическую специфику, не оставляют сомнения в важности накопления и систематизации признаков его региональной просодической нормы.

В данной статье представлены результаты сопоставительного экспериментально-фонетического изучения структурирования просодии шотландской (SSE) и английской (RP) общенациональной кодифицированной речи в ее двух жанровых разновидностях: дикторской презентации шотландских и британских теленовостей и телеинтервью с этнически разными носителями соответствующей речевой культуры.

Объектом исследования выбрана коммуникативная сфера профессионального публичного общения как наиболее репрезентативная форма речи в плане признания ее орфоэпической правильности, и которая, следовательно, представляет собой объективный источник для расширения базы описания характера и степени устойчивости шотландской просодической нормативности в ее современной реализации.

Тексты новостных сообщений дикторов BBC, канала Scottish National Heritage (SNH), Scottish TV (STV, телепрограмма STV News, Scotland Tonight) тексты телеинтервью, транслируемые BBC, Scottish Research Society и STV в 2011-2014 годах, в исполнении 12 испытуемых, были получены через сеть Интернет [7; 8; 9]. Выборка текстов для фонетического анализа, общая продолжительность звучания которых составляла 14 минут, была сделана на основе их стопроцентной идентификации носителями и неносителями английского языка с высоким уровнем его владения как типичных образцов общенационального английского (RP) и нормативного шотландского (SSE) звучания.

Аудитивный анализ текстов идентичной временной продолжительности проводился опытными аудиторами-фонетистами по признакам тона, ударения и членения на просодические единицы и был представлен в их графической транскрипции по принятой в английской фонетике системе их дистинктивной типологизации. Формат представления перцептивных портретов текстов позволил с достаточной степенью детализации выявить различия в произнесении сравниваемых образцов английской и шотландской официально-деловой речи.

На акустическом уровне их сравнение проводилось по признакам среднеслоговой длительности фраз и интонационных групп (СДС) в миллисекундах (мсек) и уровня частоты основного тона (ЧОТ) в герцах (Гц).

По перцептивным оценкам аудиторов шотландские новости обладают определенной регионально обусловленной контрастивностью их звучания относительно британских новостных сообщений.

Тексты новостных сообщений в SSE представляются дикторами в более быстром темпе, что находит себе подтверждение в большем количестве слов (179) и интонационных групп (39) за единицу времени по сравнению с RP (145 и 33 соответственно). СДС в SSE равна 250 мсек, в RP – 272 мсек. При этом следует отметить большую плотность сравниваемых признаков значений у большинства дикторов, речь которых имеет идентичные показатели.

В SSE по сравнению с RP имеет место более дробная сегментация новостных текстов на просодические единицы при относительно равном количестве фраз.

Различие в темпе речи в шотландских новостных блоках также выражено в вербальном укрупнении количественного состава интонационных групп, а именно преобладании двухакцентных интонационных групп, количество которых равно 35 % случаев, в то время как в RP их количество составляет 26 %. Здесь также необходимо указать на практически равное количество полиакцентных интонационных групп в обоих региональных вариантах, которые, на наш взгляд, являются общедикторским признаком профессионального говорения, ускоренной передачи информации.

A memorial has taken place in the Norwegian Parliament to commemorate those who died in the bombing and shootings last month. Norway's prime minister warned against the danger of the witch hunt after the deaths of 77 people mostly teenagers saying all of us have got something to learn.

Например:

BBC News:

STV News:

Различие состоит также в том, что в SSE трехакцентные интонационные группы преобладают над четырехакцентными (6:1), в RP наблюдается противоположная тенденция (4:5).

Количество паузальных перерывов звучания в новостных выпусках в обоих типах совпадает, что является закономерным в силу институционального характера дикторской речи в обоих вариантах, исключаящее нали-

чие длинных пауз в силу регламентации программного времени. Сегментация текстов в обоих вариантах осуществляется преимущественно короткими паузами (SSE – 58 %, RP – 60 %).

A recycling firm has been fined a record two hundred thousand pounds for storing hundreds of tons of waste without a license at a former perse at Westloothian. Branded Scotland's worst polluter, Dunnan Plant Ltd. has previous convictions for similar offences.

Направление движения терминальных тонов в финальных и нефинальных интонационных группах характеризуется различием значительного разброса их терминального завершения по высотно-уровневому признаку, несмотря на профессиональную идентичность представления новостей по другим тональным признакам. В чтении новостных сообщений на фоне общей тенденции к преимущественному использованию нисходящих тонов в финальных интонационных группах, отмечается разная частотность использования восходяще-нисходящего движения тона в завершении фраз (30 % в SSE и 4 % в RP соответственно) и более частая встречаемость нисходяще-восходящего тона в SSE (9 % и 4 % в RP).

Высокие уровневые разновидности как однонаправленных, так и разнонаправленных тонов преобладают в обоих вариантах новостных сообщений.

Предъядерные тональные контуры фраз в текстах SSE более вариативны и содержат большее количество кинетических тонов.

Речь дикторов BBC воспринимается аудитором как более экспрессивная за счет частого использования эмфатических тонов (12 % случаев); в шотландских новостных сообщениях их число составляет только 1 %.

На акустическом уровне максимальный и минимальный уровни ЧОТ в обоих типах произношения у всех дикторов имеют примерно равные величины: 148–276 Гц в RP для женского голоса (диапазон составляет 1,9 относительных единиц) и 109–217 Гц для мужского (диапазон – 2,0 ед.), и в SSE 145–294 Гц для женского (диапазон 2,0 ед.) и 98–205 Гц (диапазон – 2,1 ед.) для мужского голоса. Данные показатели свидетельствуют о примерно одинаковых технических ограничениях высотно-тональной подачи дикторской речи.

Совокупности наиболее релевантных просодических признаков отличия шотландских новостных сообщений представлены в виде диаграммы.

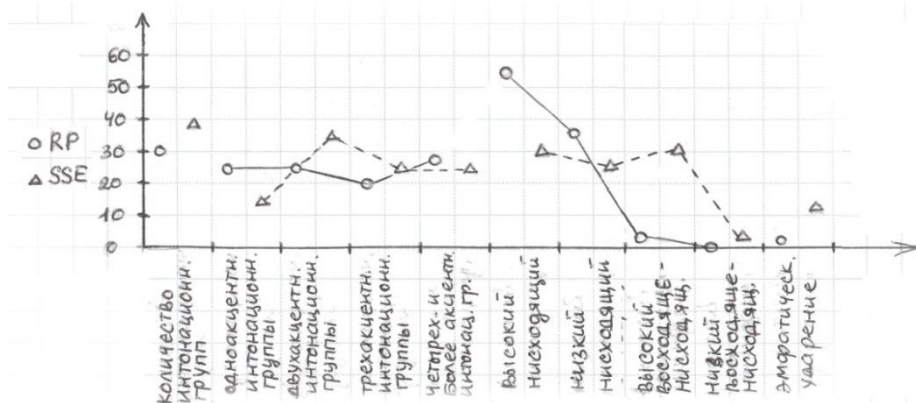


Диаграмма 1.

Отобранные для анализа телеинтервью в отличие от чтения новостей представляли собой образцы реализуемой спонтанно в теледебатах устной речи, степень подготовленности которой основывается на профессиональной компетенции говорящих.

Сравнительный анализ просодических структур фраз, составляющих тексты респондентов телеинтервью, проводился по однотипным признаковым значениям их количественных величин на воспринимаемом и акустическом уровнях. Их наиболее значимые различия показаны на диаграмме 2.

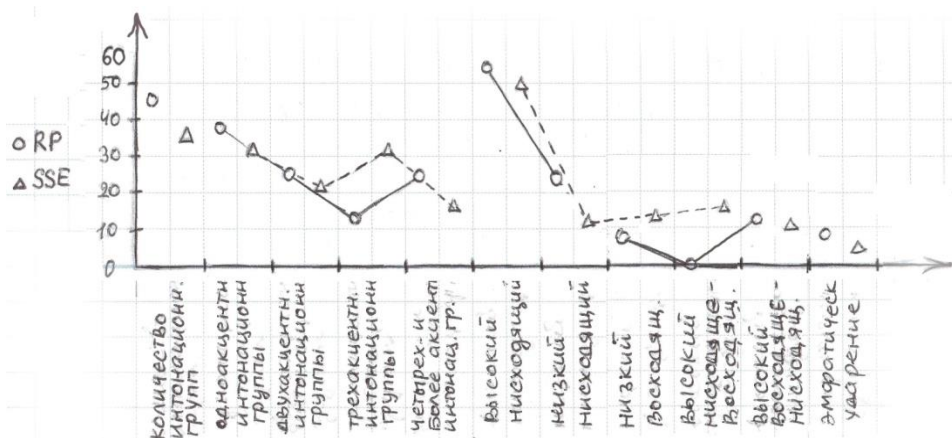


Диаграмма 2.

Тексты шотландских испытуемых-респондентов интервью в нашем материале представляли собой регионально более маркированные варианты как относительно речи носителей RP, так и по сравнению с речью шотландских дикторов. Количественные показатели величин признаков значений в шотландских интервью, представленные на диаграмме, показывают больший размах их отклонений.

Данное сопоставительное исследование позволяет утверждать, что региональная специфика просодических средств в SSE затрагивает ограниченное число отдельных просодических модификаций фраз при сохранении константных просодических признаков с точки зрения базового инвентаря структурирования английской устной официально-деловой речи.

Выбор носителями SSE тех или иных просодических средств выразительности речи предопределяется социальной и интралингвальной спецификой шотландского регионального стандарта. Просодически это проявляется в различной сегментации текста, темпе речи, мелодическом контуре предъядерной части интонационных групп, предпочтительности дистрибуции терминальных тонов в завершении нефинальных и финальных интонационных групп и в частотности использования типов фразовой слоговой выделенности. Повторяемость специфических характеристик просодического структурирования фраз в речи дикторов и респондентов телеинтервью в SSE, наряду с установленными различиями на сегментном уровне, подтверждает статус шотландского варианта английского языка как регионально маркированного орфоэпического стандарта. Анализ показал, что важным для оценки значимости сходств и различий данного регионального варианта является не сам набор тех или иных признаков просодических структур фраз, а отношения различных форм противопоставленности их выражения в устном тексте.

Наличие значительного количества тождественных просодических характеристик, как в дикторской, так и в речи образованных носителей шотландской речевой культуры убеждает в том, что из-за «выравнивающего» влияния единого орфоэпического общенационального английского стандарта, в том числе распространяемого средствами массовой информации, региональные типы произношения на территории Великобритании перестают существовать в «чистом» виде; большинство говорящих, например, в Шотландии, наряду с родным диалектом Scots пользуются литературным английским, приобретенным с высшим образованием [3]. Речевые образцы дикторских английских и шотландских теленовостных сообщений и интервью отличаются общей совокупностью их базового «орфоэпического ядра», составляющего единый коммуникативно-стилистический тип официально-деловой речи.

ЛИТЕРАТУРА

1. *Durand, J.* English in early 21st century Scotland: a phonological perspective / J. Durand // *Tribune des Langues Vivantes* 36, 2004 – P. 87–105.
2. *Stuart-Smith, J.* Scottish English Phonology / J. Stuart-Smith // *Varieties of English*. – Berlin, N. Y. : Mouton de Gruyter, 2008. – P. 47–67.
3. *Sundkvist, P.* SSE as spoken in Lerwick [Electronic resource] / P. Sundkvist. – Mode of access : <https://www.abdn.ac.uk/pfrlsu/documents/Scottish%20Standard-%20English%20as%20spoken%20in%20Lerwick.pdf>. – Date of access : 15.03.2016.
4. *Шевченко, Т. И.* Социальная дифференциация английского произношения / Т. И. Шевченко – М. : Высш. шк., 1990. – 142 с.
5. *Macafee, C.* Scots and Scottish English / C. Macafee // *Legacies of Colonial English: Studies in Transported Dialects*. – Cambridge : CUP, 2004. – P. 11–61.
6. *Scobbie, J. M.* Acquisition of Scottish English Phonology: an overview / J. M. Scobbie // *QMU Speech Science Research Centre Working Papers*, 2006. – P. 4–21.
7. Scottish National Heritage [Electronic resource]. – Mode of access : www.snh.gov.uk. – Date of access : 11.04.2016.
8. BBC News [Electronic resource]. – Mode of access : <https://www.youtube.com/user/bbcnews>. – Date of access : 11.04.2016.
9. STV News [Electronic resource]. – Mode of access : <https://www.youtube.com/user/STVNews>. – Date of access : 11.04.2016.

З. А. Сиразитдинов

О МОДЕЛИРОВАНИИ СЛОВОИЗМЕНИТЕЛЬНОЙ СИСТЕМЫ ИМЕННЫХ ЧАСТЕЙ РЕЧИ БАШКИРСКОГО ЯЗЫКА ПАРНЫМИ СОЧЕТАНИЯМИ

Башкирский язык относится к агглютинативным языкам, которые характеризуются низкой средней частотой встречаемости словоформ в текстах и большим разнообразием реализуемых от основы самих словоизменительных форм. В табл. 1 приведены средние частоты повторяемости словоформ по текстам ряда языков с сопоставимыми объемами. Данные, за исключением башкирского языка, взяты с работы К. Б. Бектаева [1, с. 47–48]. Показатели по башкирскому языку получены из корпусов башкирского языка (mbfl2.ru).

Средняя частота повторяемости словоформ в текстах

Тип языка	Язык	Подъязык	Объем текста в словоформах	F (частота повторяемости словоформы)
Флективно-аналитический	Английский	Судостроение	50000	9,94
Флективно-синтетический	Болгарский	Публицистика	51894	5,2
Агглютинативный	Казахский	Публицистика	50000	3,23
	Башкирский	Публицистика (газеты “Йэшлек”, “Киске Өфө”)	57089	3,33
			59799	3,48

Средняя частота повторяемости словоформы увеличивается при увеличении текста. Так в текстах английского языка (смешанные тексты всех подъязыков) объемом в 60 123 359 словоупотреблений этот показатель равен 234,56 [2], в текстах же башкирского языка (подъязык художественной литературы) объемом вдвое выше (12 043 546) средняя повторяемость словоформ всего лишь 30,56.

Моделирование словоизменительной системы башкирского языка показывает, что от именной основы возможна практическая реализация 611 словоизменительных форм, от глагольной основы – 665 форм [3, с. 105]. Данные близкие к нашим имеются и по казахскому языку (700 вариантов словоизменительных форм для казахского глагола) [4, с. 88].

Таким образом, формальное описание словоформы, составление моделей словоизменительной системы языка, которое является актуальной задачей прикладной лингвистики в целях разработки систем автоматической переработки текстов, многократно становится значимой для агглютинативных языков. Для решения таких задач нами была предложена псевдотензорная модель, которая дает хорошие результаты при разработке системы автоматического морфологического разбора словоформы национального языка [5, с. 20–22].

В процессе работы возникла идея представить более простую модель словоизменения через порядки расположения морфем. Существует грамматика порядков, теоретическое обоснование которой на основе упорядочения элементов языка по их позиции в правильных последовательностях принадлежит Глисону [6]. Предложенную идею он иллюстрирует на агглютинативном турецком языке. Есть работа по грамматике порядков относительно узбекского языка [7].

Несмотря на некоторые различия, в учебной и академической грамматике башкирского языка для именных частей речи выделяются 15 основных категорий словоизменительных аффиксов [8, 9]. Эти категории представлены в табл. 2.

Категории словоизменения именных частей речи

1	категория множественного числа
2	категория падежного склонения
3	категория сказуемости
4	категория принадлежности
5	категория вопросительности
6	категория неопределенности (частица неопределенности)
7	категория усиления (усилительно-утвердительная частица)
8	категория притяжательности (с аффиксом -дыкы/-деке)
9	категория уменьшительно-ласкательности (с аффиксом -кай/-кэй)
10	категория уподобления (с аффиксами -дай/-дэй, -са/-сэ)
11	усложненная конструкция (с аффиксом -тағы/-тәге)
12	категория обладательности (аффиксом -лы/-ле)
13	категория лишительности (аффиксом -һыз/-һез)
14	категория предельности (аффиксом -ғаса/-ғәсә)
15	категория сравнительной степени (аффиксом -рак/-рәк)

По Глисону понятие порядка вводится следующим образом: порядок 1 включает все морфемы которые непосредственно следуют за корнем. Порядок 2 включает морфемы, которые могут следовать непосредственно за морфемой первого порядка или непосредственно за корнем. Порядок 3 состоит из морфем, следующих за корнем или за морфемами первого или второго порядка и т.д. [6, с. 164]. Для словоизменительных аффиксов именных частей речи башкирского языка в 1 порядок войдут все 15 аффиксов из таблицы 2. Порядок 2 будет также включать те 15 аффиксов, поскольку все они входят в 1 порядок, и повторение некоторых из них. То же самое произойдет с другими порядками. Применить грамматику порядков непосредственно по Глисону для системы словоизменения башкирского языка оказывается не простой задачей. Отметим, что и Глисон, и Ревзин с Юлдашевой иллюстрировали теорию порядков на примере смешанных словоизменительных, формообразовательных и словообразовательных форм языка. Конкретно только словоизменительная система языка в полном разнообразии всех форм ими не рассматривалась.

Для именных частей речи башкирского языка существуют определенные порядки следования словоизменительных морфем между собой. Так аффикс множественности может сочетаться со всеми аффиксами, при этом данный аффикс стоит перед всеми остальными аффиксами за некоторыми исключениями. Падежные аффиксы могут сочетаться со всеми аффиксами и частицами. При этом они употребляются перед частицами, но после аффиксов, за исключением аффикса сказуемости. Подробный порядок следования аффиксов представлен нами в одной из наших работ (5, с. 14–18). На основе анализа порядка следования аффиксов словизменения нами составлены парные сочетания именных аффиксов. Результат приведен в табл. 3.

Т а б л и ц а 3

Парные сочетания именных словоизменительных аффиксов

1	2	8	10	12	4
1	3	9	1	12	5
1	4	9	2	12	6
1	5	9	3	12	7
1	6	9	4	12	8
1	7	9	5	12	9
1	8	9	6	12	10
1	11	9	7	12	11
1	13	9	8	12	15
1	14	9	11	13	1
2	3	9	12	13	2
2	5	9	13	13	3
2	6	9	14	13	4
2	7	10	1	13	5
2	15	10	2	13	6
3	5	10	3	13	7
3	6	10	4	13	8
3	7	10	5	13	9
4	2	10	6	13	10
4	3	10	7	13	11
4	5	10	8	13	15
4	6	10	11	14	5
4	7	11	1	14	6
4	8	11	2	14	7
4	10	11	4	15	1
4	11	11	5	15	2
4	14	11	6	15	3
4	15	11	7	15	4
5	6	11	8	15	5
8	1	11	10	15	6
8	2	11	15	15	7
8	3	11	35	15	8
8	5	12	1	15	10
8	6	12	2	15	35
8	7	12	3		

Как видно из табл. 3, парные сочетания образуют определенный порядок, хотя и не в понимании грамматики порядков Глисона, и могут быть использованы при моделировании в целях автоматического описания любой словоформы языка. Проиллюстрируем это на следующих примерах:

1. Словоформа *кешеларебе3зе* формально на основе списка словоизменительных аффиксов и словаря основ может быть описана как *кеш* + *е* + *лар* + *е* + *бе3* + *зе* с моделью сочетаний аффиксов 4+1+4+3+2 и как *кеше* + *лар* + *е* + *бе3* + *зе* с моделью сочетаний аффиксов 1+4+3+2. Парные сочетания позволяют сразу же отбросить не реализуемый в языке первый вариант.

2. Словоформа *балыксыга* может быть формально разбита на модели *балыксы* + *га* (2) и *балык*+*сы*+*га* (7+2). Задаваемая модель сочетаний (порядков) аффиксов именных частей речи исключает второй вариант.

Аналогичным образом работают и модели парных сочетаний глагольных форм.

При моделировании словоизменительных форм через парные сочетания необходимо учитывать и возможные исключения, типа не полного сочетания (аффикса сказуемости с аффиксом принадлежности: не реализуется сочетание принадлежности и сказуемости в одном лице) и некоторые другие.

Разработанная программа автоматического морфологического анализа башкирской словоформы на основе модели парных сочетаний и словаря основ показывает хорошие результаты. Отметим, качество и полнота анализа во многом определяется объемом словаря основ. На сегодня наш словарь включает более 80 000 единиц. Словарь основ создан на базе ранее изданных словарей и пополняется за счет издающихся в ИИЯЛ УНЦ РАН новых академических, учебных и отраслевых лексикографических разработок. Мы считаем, что предложенный подход может быть с успехом применен и для остальных агглютинативных языков народов России, представители которых сегодня заняты разработкой национальных корпусов с системой автоматической переработки и поиска информации [10; 11; 12].

ЛИТЕРАТУРА

1. Бектаев, К. Б. Статистико-информационная типология тюркского текста / К. Б. Бектаев. – Алма-Ата : Наука, 1978. – 182 с.
2. Risland, H. D. A Basic Writing Vocabulary of Elementary School Children / H. D. Risland. – New York, 1945.
3. Сиразитдинов, З. А. О моделировании словоизменительной системы и разработке программ автоматического морфологического анализа башкирского языка / З. А. Сиразитдинов, Б. З. Сиразитдинов // Современное казахское языкознание: актуальные вопросы прикладной лингвистики : матер. Междунар. науч.-теор. конф., посвящ. 75-летию проф. А. К. Жубанова. – Алматы, 2012. – С. 103–107.
4. Жубанов, А. Х. Основные принципы формализации содержания казахского текста / А. Х. Жубанов. – Алматы, 2002. – 250 с.
5. Сиразитдинов, З. А. Моделирование грамматики башкирского языка / З. А. Сиразитдинов. – Уфа: Гилем, 2006. – 159 с.
6. Глисон, Г. Введение в дескриптивную лингвистику. М.: Издательство иностранной литературы, 1959. – 485 с.
7. Ревзин, И. И. Грамматика порядков и ее использования / И. И. Ревзин, Г. Д. Юлдашева // Вопросы языкознания. – 1969. – № 1. – С. 42–56.
8. Грамматика современного башкирского литературного языка (под. ред. А. А. Юлдашева). – М. : Наука, 1981. – 795 с.
9. Хәзәрге башкорт теле (педагогия институтының башланғыс кластар факультеты студенттары өсөн дәреслек). – Өфө, 1986. – 308 с.
10. Куканова, В. В. Словоизменительные типы в калмыцком языке в свете автоматической обработки текстов (на примере имени существительного) / В. В. Куканова // Вестн. Калмыц. ин-та гуманитар. исслед. РАН. – 2012. – № 2. – С. 168–177.

11. Салчак, А. Я. Электронный корпус тувинского языка: состояние, проблемы / А. Я. Салчак, А. В. Байыроол // Мир науки, культуры, образования. – Горно-Алтайск, 2013. – № 6. – С. 408–409.
12. Бадмаева, Л. Д. Бурятский языковой корпус: создание, проблемы и перспективы / Л. Д. Бадмаева // Вестн. Бурятского науч. центра Сибирского отд. РАН. – 2013. – № 2 (10). – С. 118–122.

А. В. Скрябина

ОПРЕДЕЛЕНИЕ ОСНОВНОГО СОДЕРЖАНИЯ АНГЛОЯЗЫЧНЫХ ПУБЛИЦИСТИЧЕСКИХ ТЕКСТОВ

В настоящее время тексты на естественном языке (ЕЯ) являются основным способом хранения и передачи знаний. В связи с усиливающейся тенденцией к хранению текстов в цифровом виде и с быстрым ростом объёма текстовой информации актуальной является проблема автоматизации обработки подобной информации, в частности проблема машинного анализа текста [1].

Задаче компьютерного анализа текста на естественном языке посвящено множество теоретических и практических работ. Системы анализа текста решают сегодня целый ряд задач, такие, как формирование информационного портрета текста в терминах ключевых понятий, выявление смысловых связей между понятиями, автоматическое реферирование. Доступные сегодня вычислительные мощности позволили применить широкий класс математических методов анализа неструктурированных данных для обработки больших массивов документов, эффективно решая задачи поиска информации, классификации, кластерного анализа, выявления скрытых закономерностей и др. К сожалению, внедрение математических методов в обработку текста проходит на фоне некоторой неразработанности собственно лингвистической составляющей алгоритмов, что не позволяет достичь высокого качества работы прикладных систем [2].

В отличие от смысла, содержание у текста может быть только одно, оно постоянно и неизменно. Ввиду этого факта, представляется возможным определить основное содержание текста с помощью методов прикладной лингвистики. На сегодняшний день компьютер в состоянии определить основное содержание любого текста, хранящегося в его памяти. Что же касается его смысла, то тут проблема состоит в множественности смыслов одного текста.

Для определения основного содержания текста существует несколько критериев [3]:

- позиционный (критерий положения слова в тексте);
- статистический (критерий выделения «опорных слов»);
- логико-семантический.

В основе статистического метода лежит критерий выделения ключевых или опорных слов, которые пронизывают текст и влияют на понимание всего текста. В структуре «опорных слов» или «смысловых вех» существует определённая иерархия. Часть из них является «главными», «важными словами», «номенклатурными дескрипторами», «выступающими точками» всего предложения. Они определяют главный предмет сообщения и по существу явля-

ются свёрнутым замыслом текста. При этом один и тот же денотат может выражаться отдельными словами – контекстуальными синонимами – в тех пределах, при которых возможно их взаимное приравнивание. Ввиду линейности означающего текста, такие слова оказываются разнесёнными по разным абзацам текста.

Основным недостатком подобных систем является то, что они работают со смыслами отдельных слов и словосочетаний, реже – со связями между ключевыми словами, но структура предложения целиком не анализируется. Это можно определить как недостаток, потому что часто смысл слов зависит от контекста. В настоящее время не существует (или находятся в экспериментальной стадии) автоматизированных систем, анализирующих семантическую структуру текста на уровне выше предложения.

В настоящее время ни одна из теорий не может претендовать на полноту, хотя наиболее совершенные теории достигли удовлетворительных теоретических результатов. Подобные теории используют предложение как структурную единицу и имеют слабые средства представления связного текста. Кроме того, все они требуют дополнительной формализации для их реализации на компьютере.

Исходя из этого, задача компьютерного анализа и определения основного содержания текста на естественном языке остается одной из ведущих тем многочисленных теоретических и практических исследований в области компьютерной лингвистики. В нашей работе мы используем статистический метод для определения основного содержания публицистического текста.

В нашей работе материалом исследования послужили англоязычные новостные публицистические тексты по теме культура, взятые с Интернет-ресурсов www.bbc.com и www.theguardian.com. Для определения основного содержания каждого из текстов мы использовали статистический метод. В основе этого метода лежит критерий выделения ключевых или опорных слов, которые пронизывают текст и влияют на понимание всего текста.

Процесс выделения опорных слов каждого анализируемого текста включает в себя следующие этапы.

1. Построение частотно-алфавитного словаря для каждого анализируемого текста. В частотно-алфавитном словаре для каждого слова текста указывается, сколько раз оно встретилось в данном тексте, в каком количестве абзацев, в каких абзацах анализируемого текста, а также сколько раз в каждом из данных абзацев. В данной работе мы использовали программу «**DIST**», чтобы получить частотно-алфавитные словари по анализируемому материалу. Для начала, мы привели отобранные тексты к виду, соответствующему требованиям, предъявляемым данной компьютерной программой: обозначили начало каждого абзаца символом «*», элиминировали переносы, преобразовали строчные символы в прописные символы.

2. Выделение из частотно-алфавитного словаря потенциальных опорных слов. Далее они служат основой для деления опорных слов на главные и второстепенные, которые и определяют, в конечном счёте, основное содержание анализируемого текста. Для осуществления данной операции мы использовали компьютерную программу «**UNIFY**», которая выделяет из частотно-алфавитного словаря слова и словосочетания, имеющие частоту два и более.

Данная операция заключалась в выполнении следующей последовательности действий:

- а) удаление из полученных частотно-алфавитных словарей всех служебных слов (предлогов, артиклей, местоимений и так далее);
- б) объединение всех грамматических форм одного слова в начальную, словарную форму, т.е. все личные формы глаголов сводились к инфинитиву, а существительные – к форме единственного числа именительного падежа;
- с) пополнение базы данных потенциальных опорных слов словами схожими по тематике;
- д) указание в базе данных возможных словоформ опорных слов.

С учетом того, что английский язык – аналитический язык, является целесообразным построить словарь словоформ для всех входящих в базу данных опорных слов.

Рассмотрим ход выполнения данных операций на примере одного англоязычного публицистического текста.

Итак, для текста «Andrew Garfield and Michael Shannon to team up for 99 Homes movie» с помощью программы «**DICT**» был построен частотно-алфавитный словарь, который показал, что текст состоит из 6 абзацев и содержит 247 слов. Для каждого слова указана частота его встречаемости в тексте и в каждом абзаце. Таким образом, в данном тексте имеются слова с частотой от 1 до 9. С помощью программы «**UNIFY**» данный частотно-алфавитный словарь был сведён к потенциальному словарю опорных слов, при этом из частотно-алфавитного словаря были удалены следующие служебные слова: артикли – *the, a, an*, местоимения – *he, you, him*, союз *and*, предлоги – *of, about, with, in, to*, вспомогательные глаголы *is, was*, имена собственные *Garfield, Michael Shannon, General Zod*, числительные *30, 2*. Все глагольные формы были приведены к форме инфинитива, существительные – к форме единственного числа именительного падежа. В итоге, полученные опорные слова были сведены в таблицы: «Главные опорные слова к тексту «Andrew Garfield and Michael Shannon to team up for 99 Homes movie» (табл. 1), «Второстепенные опорные слова к тексту Andrew Garfield and Michael Shannon to team up for 99 Homes movie» (табл. 2).

Т а б л и ц а 1

Главные опорные слова к тексту «Andrew Garfield and Michael Shannon to team up for 99 Homes movie»

Потенциальное опорное слово	F-Частота слова в тексте	m-число абзацев, в которых встретилось слово	0	1	2	3	4	5	6
			A	A	A	A	A	A	A
			B	B	B	B	B	B	B
			3	3	3	3	3	3	3
			A	A	A	A	A	A	A
			Ц	Ц	Ц	Ц	Ц	Ц	Ц
DRAMA	4	4	1	1	1	0	0	1	0
FILM	4	2	0	0	1	0	3	0	0
PLAY	3	3	1	0	1	0	1	0	0
STAR	3	3	1	1	1	0	0	0	0
TAKE	3	3	0	0	1	1	0	1	0
WORK	2	2	1	1	0	0	0	0	0

**Второстепенные опорные слова к тексту «Andrew Garfield
and Michael Shannon to team up for 99 Homes movie»**

Потенциальное опорное слово	F-Частота слова в тексте	m-число абзацев, в которых встретилось слово	0 А	1 А	2 А	3 А	4 А	5 А	6 А
			Б	Б	Б	Б	Б	Б	Б
			З	З	З	З	З	З	З
			А	А	А	А	А	А	А
			Ц	Ц	Ц	Ц	Ц	Ц	Ц
CONTRACTOR	2	2	1	1	0	0	0	0	0
MAN	2	2	1	0	1	0	0	0	0
EVICT	2	2	1	1	0	0	0	0	0
DEADLINE	2	2	0	1	0	1	0	0	0
FINANCIAL	2	2	1	0	1	0	0	0	0
CRISIS	2	2	1	0	1	0	0	0	0
OPPORTUNITY	2	2	0	0	0	1	1	0	0
AMAZING	2	2	0	0	0	0	1	1	0
HOME	2	2	0	1	0	1	0	0	0

ЛИТЕРАТУРА

1. *Захаров, В. П.* Информационные системы (документальный поиск) / В. П. Захаров. – Санкт-Петербург, 2002. – 71 с.
2. *Шемакин, Ю. И.* Тезаурус в автоматизированных системах управления и обработки информации / Ю. И. Шемакин. – М., 2006. – 266 с.
3. *Зубов, А. В.* Статистический аспект содержания текста и его формальное представление / А. В. Зубов // Квантитативная лингвистика и автоматический анализ текстов: уч. зап. Тартуск. гос. ун-та. – Тарту, 1986. – № 745. – С. 75–91.

А. И. Чапля

ИСПОЛЬЗОВАНИЕ СТАТИСТИЧЕСКИХ МЕТОДОВ В СТИЛИСТИЧЕСКИХ ИССЛЕДОВАНИЯХ

Стилеметрия – это прикладная филологическая дисциплина, занимающаяся измерением стилистических характеристик текстов с целью систематизации и упорядочения (типологии, атрибуции, датировки, диагностики, реконструкции и т.п.) текстов и их частей [1, с. 420].

Для решения любой из перечисленных выше задач необходимо отобрать определенное число числовых параметров, вычисляемых по исследуемым текстам. Число таких параметров, предлагаемых разными исследователями, составляет не одну сотню. Их простейшая классификация приведена в работе [2, с. 209]. Ниже детально рассмотрен статистический метод принадлежности текстов к некоторому подязыку или же языку в целом.

Для проведения исследования были использованы частотные словари, полученные по французским текстам нефтехимии, нефти и газу, электроники и публицистики. Каждый такой частотный словарь был получен из текстов

длиной в 100 000 словоупотреблений. В табл. 1 приведены данные о величинах абсолютной частоты (F) 520 словоформ, взятых из упомянутых частотных словарей.

Т а б л и ц а 1

№ п/п	Словоформа	Абсолютные частоты			
		нефтехимия	нефть и газ	электроника	публицистика
1	de	5592	6388	6445	6462
5	les	1907	1936	1867	2052
15	dans	907	1000	977	1115
37	production	163	257	0	54
46	forage	16	184	0	0
52	produits	213	168	0	0
68	pression	153	133	0	0
70	raffinerie	30	132	0	0
91	partie	76	95	70	55
108	hydrocarbures	376	79	0	0
139	pays	17	69	0	152
147	temps	46	65	141	108
171	prix	31	57	0	113
187	fonction	46	54	110	0
220	profondeur	8	49	0	0
227	appareil	4	7	49	0
520	vapeur	41	22	0	0

Отвлекаясь на время от лингвистического содержания задачи, представим себе, что нам заданы некоторые объекты $C_1, C_2, \dots, C_k$ (слова) и их признаки $F_1, F_2, \dots, F_k$, т.е. частоты этих объектов (слов и словоформ) в разных подязыках (отраслях, функциональных стилях).

Для объекта (слова) C_i частоты $F_1, F_2, \dots, F_k$ в разных K частотных словарях являются его признаками, т.е.

$$\begin{aligned} &C_1(F_1^1, F_2^1, \dots, F_k^1) \\ &C_2(F_1^2, F_2^2, \dots, F_k^2) \\ &\dots\dots\dots \\ &C_i(F_1^{(i)}, F_2^{(i)}, \dots, F_k^{(i)}) \end{aligned}$$

Для каждого объекта C_i (с соответствующим ранговым номером), т.е. для каждого слова частотных словарей соответствующего ранга, можно вычислить среднее арифметическое значение признака F_i [3, с. 102].

Если обозначить среднее арифметическое значение признака F_i через $\bar{X}_i$, то

$$\bar{X}_i = \frac{F_1^{(i)} + F_2^{(i)} + \dots + F_k^{(i)}}{K}, \quad (1)$$

т.е. среднее арифметическое значение F_i равно сумме всех признаков данного объекта, деленной на количество сравниваемых частотных словарей.

Каждую словоформу с частотами, употребленными в подязыках, можно рассматривать как дискретный вариационный ряд [4, с. 131].

Тогда для него можно определить среднее линейное отклонение от среднего арифметического значения по формуле [5, с. 221]

$$\bar{e} = \frac{\sum_{i=1}^n (F_i - \bar{X})}{n} \quad (2)$$

и дисперсию

$$\sigma^2 = \frac{\sum_{i=1}^n (F_i - \bar{X})^2}{n} \quad (3)$$

Извлекая квадратный корень из величины σ^2 , найдем среднее квадратическое отклонение от среднего арифметического значения

$$\sigma = \sqrt{\frac{\sum_{i=1}^n (F_i - \bar{X})^2}{n}} \quad (4)$$

Величина σ показывает, на сколько, в среднем отклоняются конкретные частоты словоформ (слов) от среднего их значения.

Для практических исследований представляется более удобным вычислять величину коэффициента вариации V , которая обычно выражается в процентах [3, с. 122]

$$V = \frac{\sigma}{\bar{X}} \cdot 100\% \quad (5)$$

Величины σ и V являются мерами степени равномерности частот словоформ (слов). Чем меньше σ и V , тем однороднее изучаемая совокупность признаков $F_1, F_2, \dots, F_k$ [5, с. 223].

Принимая во внимание, что в работах по лингвостатистике за максимальную ошибку принимают величину $\pm 30\%$ [6, с. 89], находим по формулам (1)–(5) для каждой словоформы (для каждого слова) все необходимые величины.

Если величина V получается менее чем 30% , то частоты данной словоформы распределены в языке равномерно, т.е. можно считать, что такая словоформа (слово) относится в языку в целом.

Если же величина V оказывается больше 30% , то можно предположить, что распределение частот словоформы (слова) неравномерно и такая словоформа специфична для какого-то конкретного подязыка.

Для выяснения вопроса, к какому именно подязыку относится данная словоформа (или данное слово), применяется мода [3, с. 109] D – значение наибольшего признака F_i . Тот подязык, который содержит эту наибольшую частоту данной словоформы (слова), может считаться искомым.

Однако может случиться, что в двух подязыках значение частот очень близки. Для учета таких случаев вводится дополнительная проверка степени близости двух наибольших частот F_1 и F_2 . нами было принято априорно, что, если эти два значения отличаются друг от друга не более чем на 5% , то они считаются близкими. Такая проверка осуществляется по следующей процедуре:

1. Находится величина ΔF , равная 5% от F_1 :

$$\Delta F = \frac{F_1}{100} \cdot 5 \quad (6)$$

2. Находится величина F_2^1 :

$$F_2^1 = F_1 - \Delta F \quad (7)$$

3. Определяется величина разности $F_2^1 - F_2$

Если $F_2^1 - F_2 \leq 0$, то F_1 и F_2 – близки.

Если же $F_2^1 - F_2 > 0$, то F_1 и F_2 – не близки.

Фрагменты и рекомендации по итогам вычисления всех этих величин представлены в табл. 2.

Т а б л и ц а 2

Фрагменты результатов машинного анализа ключевых слов (словоформ)

№ п/п	Словоформа	Статистические характеристики			Результат анализа и рекомендация
		$\bar{X}$	σ	V	
1	2	3	4	5	6
1	de	6221,7501	364,6179	5,86037	Слово принадлежит языку в целом
5	les	1940,5000	68,8785	3,54952	Слово принадлежит языку в целом
37	production	119,5000	99,6656	83,40225	Слово может быть ключевым в подязыке нефти и газа
46	forage	50,0000	77,6401	155,28039	Слово может быть ключевым в подязыке нефти и газа
52	produits	95,0000	96,3820	101,45479	Слово может быть ключевым в подязыке нефтехимии
68	pression	72,0000	72,4189	100,58183	Слово может быть ключевым в подязыке нефтехимии
70	raffinerie	40,4999	54,2286	133,89798	Слово может быть ключевым в подязыке нефти и газа
91	partie	74,0000	14,3352	19,37198	Слово принадлежит языку в целом
108	hydrocarbures	113,7500	154,8069	136,09401	Слово может быть ключевым в подязыке нефтехимии
139	pays	59,5000	59,1460	99,40505	Слово может быть ключевым в подязыке публицистики
147	temps	97,0000	37,0337	41,14863	Слово может быть ключевым в подязыке электроники
171	prix	50,2499	41,4691	82,52560	Слово может быть ключевым в подязыке публицистики
187	fonotion	52,5000	39,0736	74,42599	Слово может быть ключевым в подязыке электроники
220	profondeur	14,2499	20,3770	142,64571	Слово может быть ключевым в подязыке нефти и газа
227	appareil	25,0000	23,0542	92,21713	Слово может быть ключевым в подязыке электроники
236	pipe-line	14,0000	18,9208	135,14920	Слово может быть ключевым в подязыке нефти и газа

299	monde	35,5000	39,8528	112,26156	Слово может быть ключевым в подъязыке публицистики
341	vitesse	49,7500	35,4920	71,34085	Слово может быть ключевым в подъязыке электроники
388	reforming	19,7499	21,0994	106,83274	Слово может быть ключевым в подъязыке нефтехимии
433	etat	30,2500	20,0795	66,37861	Слово может быть ключевым в подъязыке публицистики
512	nombreux	21,0000	2,9999	14,28571	Слово принадлежит языку в целом
520	vapeur	15,7500	17,1227	108,71568	Слово может быть ключевым в подъязыке нефтехимии

Вычисление величин V и D дает возможность ответить более или менее точно на ряд таких вопросов, как:

1. Принадлежит ли слово языку в целом, или определенному функциональному стилю?

2. Характерно ли слово для всего периода творчества определенного автора или оно специфично для отдельного периода его творчества?

3. Принадлежит ли слово языку в целом или оно специфично для текстов определенной специальности? и т.д.

Таким образом, использованный в настоящей работе метод может быть также применен при изучении стилей разных авторов, при изучении особенностей лексики одного автора в разные периоды его жизни, при изучении различных функциональных стилей (прозы, поэзии, драматургии), наконец, в изучении словарей различных специальностей и тем. Этот же метод может быть использован и при дешифровке анонимных текстов.

ЛИТЕРАТУРА

1. Мартыненко, Г. Я. Стилеметрия / Г. Я. Мартыненко, С. В. Чебанов // Прикладное языкознание. – Санкт–Петербург : СПбГУ, 1996. – С. 420–435.
2. Мальцева, Г. Ф. Некоторые количественные приемы описания индивидуального авторского стиля / Г. Ф. Мальцева // Статистика текста: сб. статей. – Минск : Изд-во БГУ, 1969. – Т. 1. – С. 206–248.
3. Шторм, Р. Теория вероятностей. Математическая статистика. Статистический контроль качества / Р. Шторм. – М. : Мир, 1970.
4. Маринеску, И. Основы математической статистики и ее применение / И. Маринеску, Ч. Мойнягу, Р. Никулеску, Н. Ранку, В. Урсяну. – М. : Статистика, 1970.
5. Суслов, И. П. Общая теория статистики / И. П. Суслов. – М. : Статистика, 1970.
6. Фрумкина, Р. М. Статистические методы изучения лексики / Р. М. Фрумкина. – М. : Наука, 1964.

ЧИСЛЕННАЯ ОЦЕНКА КОЛЛИЗИЙНОСТИ ТЕКСТОВ

Текстоаналитическая деятельность (как и любая экспертная) должна опираться на научные экспериментально апробированные численные (количественные) методики анализа. Представление результатов анализа (экспертизы) в числовом виде усиливает их доказательность, что немаловажно для практики, поскольку на сегодня текстоаналитическая и лингвоэкспертная деятельность на Украине законодательно и нормативно не регламентированы.

Предметом исследования, отдельные результаты которого приведены в этой статье, выступают естественные языковые тексты (*ЕЯТ*), находящиеся в отношении *коллизии* – смыслового несоответствия (далее – *коллизийные тексты*).

Несмотря на значительный научный и практический интерес к коллизийным *ЕЯТ* (особенно – юридическим, следственно-судебным, публицистическим, историческим, каноническим религиозным и художественным), уровень научной разработанности проблемы коллизийного отношения *ЕЯТ* в прикладной лингвистике и смежных науках недостаточен. Так, в настоящее время отсутствуют хотя бы минимально формализованные правила, по которым устанавливаются и анализируются *текстовые коллизии* (далее – *ТК*). Заметим, что в силу стереотипности рационального мышления, воспринимающего и сравнивающего тексты, отсутствие означенных правил не сказывается на воспроизводимости (повторяемости) результатов поиска *ТК*. Однако образ несоответствия, создаваемый сознанием разных субъектов, может отличаться, вследствие чего коллизийное отношение *ЕЯТ* интерпретируется весьма широко. Отсюда – многочисленные речевые обороты, его констатирующие. Например, в правоприменительной практике коллизия между правовыми нормами, нормативно-правовыми предписаниями или положениями юридических документов констатируется конструкциями «не соответствуют», «противоречат», «не согласованы», «имеет место расхождение». Также распространены конструкции с модификаторами, усиливающими или ослабляющими «силу проявления» коллизии («в незначительной мере противоречат», «имеет место незначительное расхождение», «не соответствуют или в значительной степени противоречат», «противоречат или в значительной мере не соответствуют»).

Очевидно, что множественность выражений суть следствие множественности форм *ТК*, однако в настоящее время какие-либо подходы к численному анализу *ТК* и коллизийных *ЕЯТ*, по данным нашего многолетнего мониторинга отечественной и зарубежной научной периодики, отсутствуют.

Вышесказанное обуславливает **актуальность** темы статьи.

Экспериментальным материалом исследования послужили коллизийные тексты разных литературных стилей в количестве свыше 1000 *оппозиций* (сопоставленных *фрагментов* – отрывков текста, в пределах которого лексические, синтаксические, грамматические, семантические, логические, прагматические и др.. связи достаточны для того, чтобы приписать *фрагменту* *свойство смысла*).

Исследование проводилось в Украинском языково-информационном фонде НАН Украины (г. Киев) в рамках выполнения НИР «Когнитивно-коммуникативные аспекты функционирования коллекций электронных текстов» (2013–2016), № госрегистрации 0112U004483.

«Диапазон» и оценочная «шкала» коллизийного отношения ЕЯТ. Под «диапазоном» коллизийного отношения ЕЯТ понимаем упорядоченную совокупность всех установленных форм (значений) ТК. Отправной точкой наших размышлений послужила структура противоречия, предложенная современным русским философом Л.Е. Балашовым [1], согласно которому сторонами противоречия выступают *тождество* и *противоположность*, а «пунктами перехода», «опосредующими звеньями» (термины Э. В. Ильенкова) – разное соотношение этих сторон. Таким образом, «диапазон» отношения *фрагментов* в коллизийном аспекте с одной стороны ограничен нулем (тождество, полное соответствие, адекватность), с другой – единицей (противоречие). Понятно, что область рассуждений (*предметная область, множество ситуаций*) коллизийного отношения ЕЯТ будет эже, поскольку ноль в нее не входит: $[0 < ; 1]$.

Между граничными точками «диапазона» ТК заключены все возможные ее формы. С целью квантования области рассуждений коллизийного отношения ЕЯТ установим лингвистические термы – «Противоречие», «Сильная противоположность», «Противоположность», «Отличие», «Слабое тождество» и соответствующие каждому терму числовые значения функции принадлежности отношения *коллизийных фрагментов* f_i и f_j ко множеству «Противоречие» C (в символьной записи $MC(f_i, f_j)/\Lambda$).

- «Противоречие»: $MC(f_i, f_j)/1$;
- «Сильная противоположность»: $MC(f_i, f_j)/0,8$;
- «Противоположность»: $MC(f_i, f_j)/0,6$;
- «Отличие»: $MC(f_i, f_j)/0,4$;
- «Слабое тождество»: $MC(f_i, f_j)/0,2$.

Таким образом, коллизийное отношение ЕЯТ представляется конъюнкцией вида:

$$K = \{MC(f_i, f_j)/\Lambda_i\}, MC(f_i, f_j) \in [0,2] \dot{\vee} [0,4] \dot{\vee} [0,6] \dot{\vee} [0,8] \dot{\vee} [1].$$

В целях лингвистической экономии с числовым значением Λ будем отождествлять «мощность» коллизии – абстракцию, позаимствованную из юридических наук, где она обозначает «степень столкновения нормативно-правовых предписаний» (см. работы по юридическим коллизиям и коллизионному праву Ю. А. Тихомирова, Н. А. Власенка, С. С. Алексеева, В. В. Ершова и др.)¹.

¹ Ввиду ограниченного объема статьи работы упомянутых и других правоведов-коллизонистов (термин) в списке литературы (а также ссылки на их электронные версии) не приводим. Интересующихся отсылаем к обзорной статье Патраковой И.И. «Классификации юридических коллизий: история и современность» [Электронный ресурс]. URL: http://www.juristlib.ru/book_9587.html, 2010.

Общие методические рекомендации к определению формы ТК:

1. Определение формы ТК производится аналитиком (экспертом). Коллизии ЕЯТ присваивается форма «Противоречие» и функция принадлежности $\Lambda = 1$ (соответственно, «мощность» = 1), если один из фрагментов-оппозиций содержит выраженное любым способом и любыми средствами¹ *контрадикторное* общее (полное/ фразовое/ предикатное) отрицание, сфера действия которого распространяется на весь текст. Посредством общего отрицания выражается идея о том, что некоторое положение вещей (ситуация, действие, результат), описанное в оппозиционном фрагменте, по мнению автора-оппонента, не имело, не имеет, и не будет когда-либо иметь места. Противоречием также считаем отрицание модальностей (м. долженствования *разрешено – не разрешено (запрещено)*, м. возможности *может – не может*, м. необходимости *должен – не должен*, м. желания *хочет – не хочет* и т.д.).

Пример:

О. Ляшко, лидер «Радикальной партии», народный депутат Украины (URL: http://dt.ua/)	Экс-министр экологии и природных ресурсов Украины И. Шевченко (URL: https://www.facebook.com/igor.shevchenko/posts/10100933161947774)
Против отправленного в отставку за коррупцию министра экологии Шевченко начато уголовное производство за незаконное обогащение.	Ляшко, как всегда, нагло врет! Никакого уголовного производства против меня нет!

Комментарий: в оппозиции к утверждению – общее отрицание, усиленное речевой агрессией.

2. ТК присваивается форма «Сильная противоположность» ($\Lambda = 0,8$), если в одной из оппозиций имеет место противопоставительное отрицание с утверждением (уточнением), в котором значатся антонимические (полярные, противопоставленные), относящиеся к предмету мысли: признаки, свойства, действия, состояния; явления, положения, направления и т.д. По Ю. Д. Апресяну и И. М. Богуславскому – «противоположность с оттенком умеренности» (*не закрытый, а открытый перелом*).

Пример:

Материал веб-сайта Наша-природа.ру (URL: http://ours-nature.ru/lib/b/book/.../14)	Материал веб-сайта Фактрум.ру (URL: http://www.factroom.ru/life/20-mysteries-of-world-history)
Неандертальцы – дикие псы Ледникового периода	Вопреки расхожему мнению о том, что неандертальцы были дикими или полудикими, они таковыми не являлись. Это была достаточно развитая эволюционная ветвь, они даже умели делать музыкальные инструменты.

Комментарий: в оппозиции к утверждению – противопоставительное отрицание с утверждением.

¹ Способы и средства выражения разных видов отрицания в русском языке достаточно полно освещены в работе Е.В. Падучевой «Русское отрицательное предложение», изд-во «Языки славянских культур», 2013.

3. *ТК* присваивается форма «Противоположность» ($\Lambda = 0,6$), если в одной из оппозиций имеет место *контрарное* частичное (относящееся к некоторому атрибуту ситуации) отрицание (при имени, присловное и др.), в том числе, допускающее исключенное третье (*плохо – неплохо – хорошо*).

Пример (материал сайта политических новостей ПОЛИТ.РУ <http://www.polit.ru/news/2010/12/16/podmoskovie/>, публикация от 16.12.2010):

<i>Сотрудник правоохранительных органов</i>	<i>Администрация Солнечногорска</i>
Большая группа радикально настроенных молодых людей прошла по улицам подмосковного Солнечногорска, <u>выкрикивая националистические лозунги</u> и <u>избивая людей</u> кавказской и азиатской внешности, попадавших им по дороге.	Некое скопление молодежи имеет место быть, но никаких <u>правонарушений</u> с их стороны <u>не наблюдалось</u> .

Комментарий: в оппозиции к утверждению – частичное отрицание.

4. *ТК* присваивается форма «Отличие» ($\Lambda = 0,4$), если в оппозициях значатся разные обстоятельства (атрибуты) ситуации (качества, места, причины, времени, цели, меры, условия и т.д.), называются разные акторы, участникам ситуации приписываются разные имена, несовместимые функции, состояния, свойства, характеристики, положения и т.д..

Пример: Контекст: В Иерусалиме (Израиль) взорван рейсовый автобус – как минимум 25 человек пострадали (сообщение ИА TOP NEWS от 23.03.11 (URL:http://topnews.ru/news_id_41557.html)).

<i>Свидетельство одной группы очевидцев</i>	<i>Свидетельство другой группы очевидцев</i>
Взрывное устройство привел в действие смертник, находившийся <u>в автобусе</u> .	Взрыв прогремел <u>на автобусной остановке</u> возле Дворца нации («Биньяней ха-Ума»).

Комментарий: в оппозициях – разные обстоятельства места.

5. *ТК* присваивается форма «Слабое тождество» ($\Lambda = 0,2$) если качественное/ количественное расхождение атрибутов незначительно, или/ и ситуация имеет низкую социальную значимость (например, в силу давности, спорности, очевидного отсутствия какого-либо влияния на жизнедеятельность и развитие общества, отдельных его членов, предмета мысли и т.д.).

Пример: Контекст: ради кортежа Авакова, который летает по Киеву с бешеной скоростью, перекрывают улицы (URL:<http://sled.net.ua/node/13064>).

<i>Подзаголовок веб-публикации</i>	<i>В тексте</i>
Яценюк <u>отчитал</u> Авакова и выразил надежду, что это был последний сюжет о езде кортежами	Премьер-министр Украины А. Яценюк <u>пожурил</u> министра внутренних дел А. Авакова из-за передвижения кортежем.

Комментарий: в оппозициях – синонимы с разными оттенками значения.

Количественная оценка коллизийности ЕЯТ в случае множественности ТК. Между сопоставленными *ЕЯТ* может иметь место множественность *ТК* (две и более). Здесь возможны два варианта: I) все *ТК* одной формы; II) *ТК* разных форм.

Пример: Контекст: По поводу личности самого Чонкина толки были самые противоречивые. (цит. по В. Войнович «Жизнь и необычайные приключения солдата Ивана Чонкина»).

<i>Группа 1</i>	<i>Группа 2</i>	<i>Группа 3</i>
Чонкин – это <u>уголовник</u> , бежавший из тюрьмы вместе со своими товарищами.	Чонкин – <u>белый генерал</u> , который в последнее время жил в Китае, а теперь вот напал на Советский Союз	Под фамилией Чонкин скрывается <u>сам Сталин</u> , <u>бежавший от немцев</u> .

Форма всех *ТК* в примере – «отличие».

Пример: Контекст: Из Туркмении поступает противоречивая информация про *ЧП* на складе (URL: <http://www.rosbalt.ru/main/2011/07/08/867233.html>).

	<i>МИД Туркмении</i>	<i>Представители местных правозащитных организаций</i>
I	жертв <u>нет</u>	<u>Очень много</u> жертв
II	«особых разрушений» <u>нет</u>	вокруг <u>пожары</u> . Тучи <u>взрывов</u> и зарева пожаров были видны из Ашхабада
III	<u>ситуация стабильная</u>	в Абадане <u>отключены</u> водо-, газоснабжение и электроэнергия. Возникла <u>паника</u> среди населения. Началось <u>массовое мародерство</u> .
IV	все случилось на специальных <u>складах пиротехнических средств</u> , предназначенных для <u>фейерверков</u>	взрыв прогремел на <u>территории военного арсенала</u> , где хранится <u>оружие и боеприпасы</u> .

В примере: оппозиции I, II – «противоречие»; III – «противоположность»; IV – «отличие».

Очевидно, что «уровень» («степень») коллизийности *ЕЯТ* зависит от формы и числа *ТК*. В случае множественности *ТК* одной формы проявляется эффект суперпозиции (наложения), в силу которого «суммарный эффект» двух *ТК* больше, чем одной, трех, больше, чем двух и т.д.

В случае множественности *ТК* разных форм, коллизийность *ЕЯТ* обозначается по названию *ТК* более «мощной» формы, что расцениваем как проявление эффекта поглощения (более «мощная» *ТК* «поглощает» «слабую»).

Если K_m и K_n – разные по форме *ТК*, причем $K_m < K_n$, для любых текстов T_i и T_j :

$$(K_m(T_i, T_j) \wedge K_n(T_i, T_j)) \rightarrow K_n(T_i, T_j).$$

Дальнейшие размышления сводятся к следующему. Основным принципом количественной оценки коллизийности *ЕЯТ*, безусловно, является системность, предполагающая единую методическую основу. Ее реализация требует наличия: 1) правил идентификации *ТК* (приведены выше); 2) шкалы баллов (показателей) коллизийности *ЕЯТ*. Показатели коллизийности *ЕЯТ* должны быть сравнительными, наглядными, легко выводиться, отображать эффекты суперпозиции и поглощения, при этом численно не превышая 1 – значения Λ *ТК* в форме «Противоречие» – высшей «степени» коллизийности.

Числовые значения Λ , установленные выше, определяют нижнюю и верхнюю границы «диапазона» *ТК* каждой формы. Для *ТК* в форме «Слабое

тождество» «диапазон» 0,2 ч меньше 0,4; «Отличие» 0,4 ч меньше 0,6; «Противоположность» 0,6 ч меньше 0,8, «Сильная противоположность» 0,8 ч меньше 1.

Как было сказано, в силу эффекта суперпозиции, каждая последующая *ТК* той же самой формы увеличивает коллизийность текстов в пределах соответствующего «диапазона». Если между текстами, наряду с коллизиями других форм, есть хотя бы одна в форме «Противоречие», в силу эффекта поглощения, тексты считаются такими, что противоречат, соответственно, показатель (индикатор) коллизийности (обозначим через *КИ*) будет равен 1. Для остальных случаев *КИ* меньше 1.

Числовое значение *КИ* сопоставленных *ЕЯТ* для случая множественности *ТК* соответствующих форм сведем в таблицу:

Т а б л и ц а

Количество <i>ТК</i>	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Показатель коллизийности <i>ЕЯТ</i>																			
Сильная противоположность	0,81	0,82	0,83	0,84	0,85	0,86	0,87	0,88	0,89	0,90	0,91	0,92	0,93	0,94	0,95	0,96	0,97	0,98	0,99
Противоположность	0,61	0,62	0,63	0,64	0,65	0,66	0,67	0,68	0,69	0,70	0,71	0,72	0,73	0,74	0,75	0,76	0,77	0,78	0,79
Отличие	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,4	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5	0,5
Слабое тождество	0,21	0,22	0,23	0,24	0,25	0,26	0,27	0,28	0,29	0,30	0,31	0,32	0,33	0,34	0,35	0,36	0,37	0,38	0,39

Для учета эффекта поглощения при множественности *ТК* разных форм применим позиционный подход, связывающий значение числового знака в записи и позицию (разряд). *КИ* в общем случае будет иметь вид:

$$0, \begin{matrix} \text{I} & \text{II} & \text{III} & \text{IV} \end{matrix} \begin{matrix} [**] & [**] & [**] & [**] \end{matrix}$$

I – *КИ*, соответствующий количеству *ТК* в форме «Сильная противоположность»; II – *КИ*, соответствующий количеству *ТК* в форме «Противоположность»; III – *КИ*, соответствующий количеству *ТК* в форме «Отличие»; IV – *КИ*, соответствующий количеству *ТК* в форме «Слабое тождество».

Например, между текстами установлено 3 *ТК* в форме «Противоположность», 2 – «Отличие», 5 – «Слабое тождество». Тогда *КИ* таких текстов составит 0,624124.

Предложенный подход имеет ограничение – максимальное количество *ТК* каждой формы между двумя анализируемыми *ЕЯТ* (фрагментами одного *ЕЯТ*) – 20. Однако для решения практических задач в рамках современной лингвоэкспертной и текстоаналитической деятельности этого более чем достаточно.

Выводы и перспективы дальнейших исследований. Следующий шаг – переход от количественной оценки к качественной. Числовые показатели коллизийности *ЕЯТ* возможно однозначно «разбить» на «диапазоны»

(от ... до) и проинтерпретировать, используя широкий набор имеющихся в языке инструментов (квалификаторы, квантификаторы, модификаторы и др.). Такой подход позволит выстроить лингвистическую шкалу порядка и исключает преференциальное употребление выражений для оценки коллизийности текстов, о чем было сказано выше.

ЛИТЕРАТУРА

1. Балашов, Л. Е. Новая метафизика (Категориальная картина мира или Основы категориальной логики) / Л. Е. Балашов. – М., 2003 – 868 с.

V. S. Yakovishin

FORMAL PROTO-LANGUAGE IN A TEXT PROCESSING SYSTEM

Introduction. The presented text processing system is based on the use of a formal language that represents a dibasic ring-like algebra with a set of words over an alphabet and a set of sentences on which one unary operation and two binary operations are defined. So in the formal language, text sentences are expressed as semantic formulas using both words and special operation signs: the words can represent all lexical meanings; unary operation symbols can represent all general-sentence grammatical meanings (such as modality, negation, question, etc.); and symbols of binary operations can represent functional meaning of words [1].

Thus, in the proposed formal description, every natural language sentence assumes the shape of an algebraic expression – *semantic formulas*. A set of such algebraic expressions, representing “reconstructed” deep structures of the given natural language, can be considered as a *formal proto-language* (FpL) or, specifically, a *formal proto-English*, *formal proto-Russian*, etc. Then one can say that any natural language (NL) differs basically from various artificial formal languages (FL) only in having its natural (non-formal) surface manifestations.

It is obvious that a full description of the natural language must include not only sets of *grammatical* rules (syntax and semantics), that generate a formal proto-language, but also sets of *algorithmic* rules realizing an interpretation of generated formal language, i.e., its conversion into surface manifestation. Or, in other words: a realization of the full description of natural language can be considered as creating a text processing system in which the formal proto-language is used as a semantic representation of text information. As will be shown below, the presented formal proto-language can be used for the knowledge representation in the text processing system.

Initial syntax. We assume that any language description contains a set of rules of *initial* (algebraic) syntax. The initial syntax of text sentences can represent one unary operation, denoting the syntactically independent words, and two binary operations, denoting known syntactic connections – the determination (“grammatical multiplication”) and coordination (“grammatical addition”). The determination is non-associative, non-commutative, and distributive over the coordination; and the coordination is associative and commutative [1; 2].

According to the said operation properties, the initial syntax of natural language can be specified as the following set of rules:

$$\begin{aligned} S &\rightarrow \sqrt{S}[(S\Delta X)|(X\Delta S)|(X\nabla X)]X \\ X &\rightarrow X\nabla X, \end{aligned}$$

where $\sqrt{}$ is an unary operation; Δ (“determination”) and ∇ (“coordination”) are binary operation. That is, the associativity and commutativity are specified as parentheses-free rules in which the same symbol is used for both members: $S\nabla S$, $X\nabla X$; the parentheses in $(X\nabla X)$ are necessary to express the distributivity of determination over summand strings like $(X\nabla X\nabla \dots \nabla X)$; the non-commutative property of determination is specified as a possibility to use combinations of two different elements: $S\Delta X$ and $X\Delta S$; the parentheses must be used to indicate the non-associative property of determination: $(S\Delta X)$, $(X\Delta S)$.

Generated by the initial syntax sentences have the form of abstract parenthesis expressions indicating various *syntagmatic structures* (“dependency structure”). The elementary syntagmatic structure is called the *syntagme* (aka the “syntagma” or “syntagm”). The syntagm can now be represented as a two-part string in the form of $(\alpha\Delta\beta)$, where $\alpha \in A^*$, $\beta \in A^+$ (it can be an *elliptical* syntagme, if α is the empty word). The first part α is a *head* (independent member) and the second part $\Delta\beta$ is its *modifier* (dependent member); the determination symbol Δ represents a dependency meaning and plays the role of a relation feature in syntagmatic oppositions between the marked modifier and the unmarked head.

Note that sentence structures can be expressed not only in the standard algebraic form, using the usual means of syntactic connections, i.e., parenthesis with the fixed order. For the purposes of representation of sentences in computing systems, it is convenient to use a parenthesis-free form. In the proposed here formal language notation, syntactic structures are expressed by the set-theoretical form in which all given syntagmes are expressed as certain set’s elements [3]. So the known types of syntactic structures can be given by the following both forms:

- (a) $((X_1\Delta_1X_2)\Delta_2X_3) \Leftrightarrow \{X_1, X_1\Delta_1X_2, X_1\Delta_2X_3\}$,
- (b) $(X_1\Delta_1(X_2\Delta_2X_3)) \Leftrightarrow \{X_1, X_1\Delta_1X_2, X_2\Delta_2X_3\}$,
- (c) $(X_1\Delta(X_2\nabla X_3)) \Leftrightarrow \{X_1, X_1\Delta X_2, X_1\Delta X_3\}$,
- (d) $((X_1\Delta_1(X_2\Delta_2X_3))\Delta_3X_4) \Leftrightarrow \{X_1, \Delta_3X_4, X_1\Delta_1X_3, X_1\Delta_3X_4\}$,

where (a) is the *collateral subordination*, or the parallel connection, in which all modifiers Δ_1X_2 and Δ_2X_3 are subordinated to a single head X_1 ; (b) is the *consecutive subordination*, or the stepwise connection, in which the modifier of preceding syntagme X_2 serves as the head of the succeeding syntagme $X_2\Delta_2X_3$; (c) is a sentence structure with *homogeneous* (homofunctional) parts ΔX_2 and ΔX_3 ; (d) is a sentence structure with an extreme modifier Δ_3X_4 , i.e., emphasized part.

Thus, in the generated structures, two extreme parts of the sentence can be expressed. These are the extreme head and the extreme modifier. The extreme head, known as “subject”, is a part of the sentence that occupies the absolutely independence position (the root node) of the syntagmatic structure. As an absolutely independent member the extreme head is expressed in languages by a

zero form, i.e., a syntactically neutral (unmarked) noun form (known as the “nominative”, “indefinite case”, “casus rectus”). The extreme modifier, known as “emphasized part”, “determinant”, “theme”, ‘topic’, is a part of the sentence that relates as a modifier to the whole sentence and not associates with any of its individual word-head.

Secondary syntax. Semantics. Combinatory capabilities of real words are specified by a *secondary syntax*. The secondary syntax can be induced on a postulated algebraic syntax using special symbols to denote various word categories, *parts of speech*, and functional word forms, *parts of the sentence*.

The parts of speech will be specified in terms of the existing dependency relation. That is, the belonging of various words to one and the same part of speech is grounded on possibility of each of these words as a head to attach the proper modifier as the agreed inflectional word-form. So the *noun* X_n can have the adjective modifier a_nS (*the big boy*); the *verb* X_v can have the adverbial modifier a_vS (*to run quickly*); and the *quality* X_q can have the ad-quality modifier a_qS , i.e., the “adverb of degree” (*unusually big*). It seems that only these three syntactic parts of speech exist in languages. The other word classes (such as the pronoun and the numeral) will be considered as *semantic* categories.

Thus, we can say that the given word represents a part of speech X_k , if it has (as a head word) the special *agreed* modifier a_kS . It means that on the basis of the initial algebraic syntagme ($X\Delta S$) we can obtain an induced syntagme ($X_k a_kS$) for $k \in \{n, v, q\}$. Similarly, we can get the categories of words that play the role of heads in constructions with *governed* modifiers. Such syntactic categories are the valence words, in particular the univalent X_{v1} and bivalent X_{v2} transitive verbs.

Using the rules of algebraic syntax together with the obtained secondary rules, we can construct various secondary syntactic systems. So we have received a multilevel hierarchy of syntagmes with the following types of modification.

Free modification: realized by $S \rightarrow (S\Delta X)$, where ΔX (modifying a whole sentence) can be considered as an emphasized part (determinant).

Predicativity: realized by $X \rightarrow (X_n \Delta_p S)$, where $\Delta_p S$ (predicates) modify a noun phrase representing a subject.

Attributivity: realized by rules in the form $X_k \rightarrow (X_k a_k S) | (X_k a_k \Delta S)$, where $a_k S$ or $a_k \Delta S$ (attributes) can modify the noun (adjective attributivity), verb (adverbial attributivity), or quality (attributivity of the quality).

Valence modification: realized by $X_v \rightarrow (X_{v1} o_1 X_n)$ and $X_{v1} \rightarrow (X_{v2} o_2 X_n)$, where the modifiers are distinguished as the (direct) *object* ($o_1 X_n$) and *indirect object* ($o_2 X_n$).

In the stage of semantic interpretation, the operation symbols are replaced by special designations denoting the various semantic elements, known as “grammatical meanings”. The unary operation can be interpreted by means of the usual logical modal operators (“possibly”, “necessarily”, etc.) and usual punctuation symbols (e.g., question and exclamation marks); it can also be represented by usual independent words or their abbreviations. The coordination symbol can be replaced using the special words (grammatical codes), for example: $\nabla \rightarrow \nabla_c | \nabla_{dj} | \nabla_{adv} | \dots$, where the special words (∇ with subscripts) denote the conjunctive (*and*), disjunction (*or*), adversative (*but*), etc. Similarly, the

determination symbol is replaced by special words used to link the subordinate adverbial clauses (formed from ΔS): $\Delta \rightarrow \Delta_{cd} | \Delta_{cs} | \Delta_{pl} | \dots$, where the special words denote the condition (*if*), cause (*why?*), place (*where?*), etc.

The initial determination used to link the subordinate words can be replaced by meanings of the free (adverbial) modifiers (formed from ΔX). In the text, these meanings are expressed by the case endings and prepositions; and the lexical part (a word stem) of these subordinate words (as adverbial modifiers) can be expressed either by nouns or modifying root words (“adverbs”), i.e.

$$\Delta X \Rightarrow \Delta_m X_n | mX$$

where $\Delta_m X_n$ is a modifying noun (*in the evening, on the table*); mX is a modifying root word (*yesterday, here*). The further concretization of Δ_m can be realized using the signs (prefixes) denoting the various case meanings of nouns. For example:

$$\Delta_m \rightarrow abs. | in. | inl. | md. | sb. | sp. | trs. | \dots$$

where *abs.* is a meaning of the abessive case (*without whom/what?*); *in.* is a meaning of the inessive case (*inside whom/what?*); *inl.* is a meaning of the illative case (*into whom/what?*); *md.* is a meaning of the mediative/instrumental case (*by whom/what?*); *sb.* is a meaning of the subessive case (*under whom/what?*); *sp.* is a meaning of the superessive case (*on whom/what?*); *trs.* is a meaning of the transitive case (*during whom/what?*).

The predicativity is concretized as combinations of tense, aspect, and mood:

$$\Delta_p \rightarrow p. | pt. | pf. | pCt. | p_pCt. | p_fCt. | \dots$$

where *p.*, *pt.*, and *pf.* denote the present, past, and future indefinite tense; *pCt.*, *p_pCt.*, and *p_fCt.* denote the present, past, and future continuous tense, etc.

All abbreviations denoting the grammatical (relational) meanings of the modifying words can be represented as prefixes attached to usual word stems:

abs.glass (*without glasses*); *in.garden* (*in the garden*); *inl.house* (*into the house*); *md.train* (*by train*); *sb.table* (*under the table*); *sp.* (*on the table*); *trs.forest* (*through the forest*); *pt.write* (*wrote*), *pCt.write* (*to be writing*); *a_vquick* (*quickly*); *o_1book* (*[to read] a book*).

The lexical meanings can be represented by the following word-formative rules:

$$\begin{aligned} X_n &\rightarrow R_n | RD_n | R_c | R_p, \\ X_v &\rightarrow R_v | RD_v | R_c, \\ X_q &\rightarrow R_q | R_q D_q, \\ R &\rightarrow R_n | R_v | R_q | R_c, \end{aligned}$$

where R_n (“noun”), R_v (“verb”), R_q (“quality”), R_c (“numeral”), and R_p (“pronoun”) are the categories of primary meanings (semantic parts of speech); D_n , D_v , and D_q are the categories of grammatical (derivational) meanings of the respective word categories.

The primary meanings are denoted by usual root morphemes: $R_n \rightarrow \text{book} | \dots$, $R_v \rightarrow \text{go} | R_{v1} | \dots$, $R_{v1} \rightarrow \text{read} | R_{v2} | \dots$, etc. The derivational meanings are denoted by special postfixes: $D_n \rightarrow .n | .d | .pl | \dots$, $D_v \rightarrow .v | .rv | \dots$, $D_q \rightarrow .q | .cp | .sp | \dots$, where *.n*

(“noun”), *.d* (“definiteness”), *.pl* (“plural”), etc. are derivational meanings of the noun; *.v* (“verb”), *.rp* (“repeating”), *.rv* (“reverse”), etc. are derivational meanings of the verb; *.q* (“quality”), *.cp* (“comparative”), *.sp* (“superlative”), etc. are derivational meanings of the quality. For example:

big.n (*bigness*); *big.cp* (*bigger*); *read.n* (*reading*); *slow.n* (*slowness*); *five.n* (*quintuple*); **modern.v** (**modernize**); *glory.v* (*glorify*); *drink.q* (*drinkable*); *man.pl* (*men*); *turn.rv* (*return*).

Note that the known two types of “grammatical meanings” become now the purely formal definition: the *relational* meaning is a sign of the word dependency (and hence it cannot be used in the position of the absolute independent word); the *derivational* meaning is a word modifier (i.e., it can be used in any syntactic position, including the absolute independent part of the sentence). It is known that the distinctions between the relational and the **derivational** meanings are observed at the surface level as distinctions of morphological means: the relational meanings are expressed by case endings or prepositional cases; the derivational meanings are expressed by word-formative affixes or particles and articles. But these distinctions are not always necessarily. For example, the word-formative affix **-ly** (that changes adjectives into adverbs) must be considered as a sign of the relational meaning: *a_vslow* “**slowly**”.

Using the above syntactic and semantic rules, we can represent sentence descriptions on syntactic and semantic levels. For example:

$$S \Rightarrow_1 ((X_n p \Delta_n ((X_{v2} o_2 X) o_1 (X_n \Delta_n X_q) \Delta_n X)) \Delta X_n)$$

$$S \Rightarrow_1 ((\text{author}.d p_p ((\text{give } o_2 \text{we}) o_1 (\text{book } a_n \text{new}) a_n \text{he})) \text{in.meet}.d)$$

In the meeting, the author gave us his new book.

where $\Rightarrow_1$ is the deducibility on the syntactic level; $\Rightarrow_2$ is the deducibility on the semantic level.

Knowledge representation. The recognizable subject (as a noun phrase denoting “the something or someone that the sentence is about”) can be served as a basis for integration of input sentences into their common knowledge representation. In the text processing system, input text sentences can be transformed into set-theoretical form, and then the resulting formal sentence structures are selected and united into growing knowledge representations. The integration of the sentences that have one and the same subject (a noun phrase contained in user’s request) is considered as a subject knowledge representation. And any collection of the subject knowledge representations produced in the knowledge formation process can be considered as a user-oriented knowledge base [4].

So all sentences that contain one and the same subject N_0 , indicated in the user’s request, can be integrated into a user-oriented *subject knowledge representation* $\sigma(N_0)$, i.e.,

$$\sigma(N_0) = \{S_i | N_i \supseteq N_0\},$$

where $S_i \supseteq N_i$ ($i \geq 1$) is a sentence of the incoming text. Then any collection of the subject knowledge representations $\{\sigma(N_0)_1, \sigma(N_0)_2, \dots\}$, produced in the formation process, is a *field of subject knowledge representation*, i. e., a user-oriented

knowledge base. The integration of input sentences into a common knowledge representation is based on the use of set-theoretical relations and operations. That is, the knowledge forming process can be realized by the following integration rules:

- $$(1) S_0 \times S_i \rightarrow S_i, \text{ if } S_0 \subseteq S_i;$$
- $$(2) S_i \times S_j \rightarrow S_i \cup S_j, \text{ if } N_i = N_j,$$

where $\times$ is the integration sign; S_0 is a user's request; S_i and S_j are any sentences obtained as a result of use of (1).

Let us suppose that the user's request is $S_0 = \{\text{man, man } p.\text{read}\}$. Then, as a result of the use of (1), we can obtain a set of sentences like

$S_1 = \{\text{man, man } p.\text{read, read } o.\text{book}\}$

The man reads a book;

$S_2 = \{\text{man, man } p.\text{read, read } o.\text{book, read } in.\text{library}\}$

The man reads a book in the library;

$S_3 = \{\text{man, man } a.\text{young, man } p.\text{read, read } in.\text{park}\}$

The young man reads in the park,

And as a result of the use of (2), we get the subject knowledge representation:

$\sigma(\text{man}) = \{\text{man, man } p.\text{read, read } o.\text{book, read } in.\text{library; man } a.\text{young, man } p.\text{read, read } in.\text{park}\}$

The man reads a book in the library; the young man reads in the park.

Thus, it can be obtained an on-line (user-oriented) knowledge base. The forming of the knowledge base is a two-stage process: in the first stage (data search), the usual information retrieval is realized (by drawing text information from all sources that contain the user's request); in the second stage (knowledge extraction), the obtained result is processed to extract the more specific information.

REFERENCES

1. *Yakovishin, V. S.* Algebraic representation of syntagmatic structures [Electronic resource] / V. S. Yakovishin // Web journal of formal, computational & cognitive linguistics, Issue 11. – Mode of access : fccl.ksu.ru/issue11. – Date of access : 12.04.2016.
2. *Yakovishin, V. S.* An algebraic basis of formal language description / V. S. Yakovishin // Problems of modern applied linguistics. – Minsk : MSLU, 2014. – P. 59–64.
3. *Yakovishin, V. S.* Formation of object descriptions on the basis of discrete syntagmatic representation / V. S. Yakovishin // Interactive systems and technologies. The problem of human-computer interaction : Proceedings of the 4th International conference, Ulyanovsk, 26–30 Sept. 2005. – Ulyanovsk : UISTU, 2005. – P. 16–18.
4. *Shibut, M. S.* Selection and Aggregation of Sentences in the Knowledge Formation Process / M. S. Shibut, V. S. Yakovishin // Proceedings of the 6th International Conference on Intelligent Data Acquisition and Advanced Computing Systems (IEEE'11). – Prague, Czech Republic, 2011. – Vol. 2. – P. 647–650.

**‘SCORAPICE’: ADAPTATION OF RUSSIAN SHORTHAND
FOR THE ENGLISH LANGUAGE**

At present there are computer programmes which transcribe sound speech into a written text without the need for a stenographer. In the near future computers with amazing accuracy will be able to transcribe human speech without any pre-tuning for a speaker's idiosyncrasy.

The digital Sky wifi smartpen launched by a company called Livescribe makes it possible to record an interview as well as register the writing with a miniature webcam at the tip of the pen.

At the same time it will suffice to browse in the Internet for the words *shorthand* or *стенография* to find hundreds of references to various courses and numerous sources of shorthand applications.

Shorthand allows to take rapidly written notes anywhere. It is well known that in some countries any recording gadgets, including mobile phones, are forbidden at formal meetings and conferences. The only recording devices allowed for the tech-savvy audience are those of the pen and paper kind. Moreover, it is a fascinating and creative hobby that trains your memory and reaction time or serves as a ready tool for hiding your scripts from the uninitiated.

For centuries shorthand has been used as a cultural instrument. George Bernard Shaw used to write his plays by means of shorthand. Charles Dickens, when serving as a court secretary and later as a news reporter, enlivened his shorthand records with new artistic content in his novels. Shorthand was not alien to Fiodor Dostoyevsky and Leo Tolstoy. Some of Shakespeare's plays and Martin Luther King's prayers were all preserved due to shorthand records done by unknown stenographers. No marvel, Sir Isaac Pitman (1813–97), an educator who advocated spelling reform, was knighted by Queen Victoria for his contributions to shorthand.

Pitman shorthand, which is based on the phonetic principles, is still widely used in the UK. It consists of 25 single consonants, 24 double consonants, and 16 vowel sounds. Phonetically similar, related sounds are expressed by similar signs, the shortest signs are used to represent the shortest vowels, and single strokes are used to represent single consonants. It also makes use of *positioning on the line* to express omitted vowels, i.e. writing the signs above, on, or below the line of writing. Additionally, different stroke thicknesses (thick and thin) are used to distinguish related sounds. Pitman shorthand has been adapted to many languages including Arabic, French, German, Hebrew, Hindi, Italian, Japanese, Spanish et al.

The Irish-born John Robert Gregg (1867–1948) invented his own system. Like Pitman shorthand, Gregg's shorthand is based on the phonetic system, i.e. both of them rely on the sounds not the spellings of the words for their transcription. It can be written with the same stroke thickness and uses slopes and curves for rapid writing. The length of the stroke differentiates some letters. Based on the ellipse or oval and on the slope, all characters are in one position, i.e. written on a single line of writing. Gregg shorthand is highly cursive and fast but only marginally readable. It has been published in English, French, Spanish, Portuguese, Hebrew and Russian among others.

If you write something and then immediately transcribe it, as secretaries tend to do, then there is no major problem. But if you try to read something you wrote last year, then much greater effort may be needed to decipher it, unless you have mastered sight reading thousands of abbreviated forms.

While Gregg shorthand style is the most used one in North America, other countries including the United Kingdom, Germany, Australia and New Zealand use the Teeline shorthand system. First published in 1968 by James Hill (1908–71), it was rapidly put into use as it is quick to learn and easy to use. Teeline is an alphabetic system which is rather awkward in terms of handwriting. So it retains the inadequacies of the standard alphabet. Vowels are mostly omitted for speed at the expense of readability. It is intended to aid in taking dictation by creating word outlines, and so needs to be transcribed as soon as possible. It is, therefore, more suited to professional rather than personal use.

Teeline Shorthand is taught to journalists in the UK and some Commonwealth countries, but is little known elsewhere. Speed building in Teeline is an important part of the Diploma in Journalism.

When you compare it to Gregg you see many similar shapes such as the signs for O, T, M. Vowels are eliminated unless they are the first letter of the word or the last one. When you miss out the vowels you speed up writing but you lose on precision and readability as well.

The present course called **Scorapice** is the adaptation of the Russian shorthand system which was thought out, developed and popularized by Professor Nickolai Sokolov (1900–90) and his numerous followers throughout many years [1]. He grounded his system on the research of biomechanics of hand movements in writing as well as the Russian language statistics. Later it became a unified national system of shorthand which was also taught as a correspondence course (GZOS) [2; 3].

Back in the 60s the author of the project took a GZOS specialist course. Since then I have been using the system in my linguistic career. From the very beginning, the course impressed me by its consistency and forethought, which facilitated successful and staunch acquisition of writing skills. It has become a starting point for the present course of *Scorapice* that is an English transliteration of the Russian word *скоропись* for ‘speedwriting’.

However, unlike the Russian system, *Scorapice* depends mainly on the phonetic principles rather than orthographic rules. Although the learner is supposed to know what the prefix or suffix is, it relies primarily on the sounding of a word.

In general terms, the system of Scorapice may be defined as cursive (slanting), positional and combinatory speedwriting based on phonetic principles.

The vowel sounds are expressed mostly by the relative positioning of the following consonants on the line of writing. Diacritics, such as dashes and dots for vowels, play only a subordinate role to exactly specify the true spelling of a term or a rare name. Familiar methods for contraction of words and phrases are being applied, with special characters for affixes, – i.e. morphemes, endings, prefixes, suffixes, – sound / word combinations and conventional signs being assigned on the basis of the British National Corpus and language statistics [4; 5; 6].

The material in the course is arranged in such a way that new elements are introduced gradually, being based on the principles covered earlier in the textbook. By the end of the course the expectable speed may reach more than 100 wpm and would solely depend on the practice time the learner can afford.

Each of the six units contains some guidelines and advice as to the rules and material under study. The units end with a test to estimate learner's progress of training and to encourage further studies. The concise vocabulary *English Scorapice: Frequent Words and Phrases* contains more than 1,500 entries and is meant both for reference and for further consolidation of optimal word / phrase reduction and improving writing skills.

Table 1 shows that the English set of consonants for *Scorapice*, with rare exception, matches that for the Russian shorthand style invented by N. Sokolov and his followers [2].

Table 1

The signs for consonant phonemes in *Scorapice*

[p] _____	[k] _____	[ð/ʈ] _____	[m] _____	[n/ɳ] _____
[b] _____	[g] _____	[C] _____	[h] _____	[w] _____
[t] _____	[f] _____	[ʃ] _____	[s] _____	[r] _____
[d] _____	[v] _____	[G] / [Z] _____	[z] _____	[l] _____

Certain signs also have specific meanings as well as their traditional alphabetic value, as shown in Table 2.

Table 2

Fragment of a list of basic common words represented by initial letters

Letter	Meaning	Letter	Meaning
_____	last/like	_____	good/girl/go/get
_____	be/but	_____	have/high/how/who
_____	do/day	_____	must
_____	can	_____	on
_____	for	_____	say/so
_____	general/gist/job/just	_____	shall/show

As shown in Table 3, alphabetic signs may acquire new meaning due to an altered position as to the line of writing. Many common words are represented by initial characters above, on, or below the line of writing.

Shorthand systems can hardly make do without multiple meanings of some signs which are defined by context.

Table 3

Fragment of a list of common words denoted by a single sign raise above the base-line or written through the line

Letter	Meaning	Letter	Meaning
_____	about/or	_____	machine/mechanic
_____	necessity/necessary	_____	techn(ique) /term
_____	due/duty	_____	knew/new/neo/non-
_____	to/too	_____	fact
_____	pose/post/past	_____	are/anti-

Frequently occurring clusters of consonants are indicated by special symbols – *blends*, e.g. ____taste, ____ask, ____rent, ____sweet. Blends are also used for denoting certain words: ____state, ____school, ____Sweden.

Vowels are expressed by the relative positioning of the consonant to follow on the writing line. Table 3 lists some expedients for marking vowels within a word. The diacritics (dots, dashes and apostrophe) are occasionally used to pinpoint vowel sounds in unfamiliar or isolated words, but otherwise they are hardly ever indicated. Yet, there are special signs for some diphthongs: [qʋ] ____ , [aʋ] ____ , [ɛq] ____ , and [lq] ____: ____own, ____bout, ____bear, ____beer.

Table 4

Some techniques to render vowel phonemes in *Scorapice*

Vowels	Examples	Vowels	Examples
[ʀ], [ʀ], [q], [aɪ] – omitted between consonants	bar ____ but ____ abet ____ bite ____	[ʀ], [ɪ], [ɔɪ] – writing the following sign through the line or shifting it half-size	shock ____ bought ____ boiler ____
[j], [ɪ] – raising the following symbol half-size	bit ____ beat ____	semivowel [w] – indicated by the sign for vowel [u]	____we ____whistle
[ɛ], [ɛ:], [x], [ɛɪ] – shifting the following sign to the right or raising half-size	bet ____ berth ____ bag ____ bail ____	[ʋ], [u], [ɪɪ], [jɪɪ] – writing the following sign through the line and shifting it to the right	book ____ boob ____ tube ____

In fact, the semivowel [w] is expressed in two ways – either by a sign for [u] or by an extra long vertical stroke: ____work, ____wish, ____within, ____white, ____when.

A number of suffixes and terminations deserve special attention such as ____-tion/sion, ____-(t)ute, ____-ance/ence/nounce, e.g.: ____attention, ____gratitude, ____assistance, ____chance, ____denounce.

There are special signs for prefixes which make up a most numerous array: ____under-, ____inter-, ____over- in such words as ____understand, ____interval, ____overcharge et al.

Certain frequently used roots have also received special treatment: ____ject, ____-prov/proof, ____-tain in the following words: ____object, ____approve, ____obtain. As seen from the latter example, signs may intersect each other as an expedient for new sign-formation.

Table 5 illustrates several contracting techniques widely used in *Scorapice*.

Table 5

Abbreviation techniques applied in *Scorapice*

Abbreviation by:	Examples		
the word's beginning	_____ candidate	_____ colleague	_____ literature
the initial and final signs	_____ opportunity	_____ manager	_____ method
by conventional symbols	_____ circum-/circuit	_____ surround/circle	_____ from my point of view
signs for prefixes, roots, suffixes, and word terminations	_____ precious _____ pretext	_____ analogy _____ childless	_____ emergence _____ dispute
joining two or more words	_____ full-time job _____ as much as possible	_____ social work _____ as a matter of fact	_____ year-round _____ once a year _____ global crisis

In compound words both roots are usually written in full, while some vowels may be omitted without sacrificing the word readability: _____ daydream, _____ str(aw)berry.

Particular attention has been given to the contraction of conjunctions and phrases as exemplified by Table 6.

Table 6

Contraction of some conjunctions and phrases

Outline	Meaning	Outline	Meaning
_____	at (the) top of (the)	_____	ow(ing) to
_____	(in the) course of the	_____	(it goes) without saying
_____	due to	_____	at (the) back of (the)

As Table 6 suggests, many grammatical elements are being left out in many-worded phrases. These and the other techniques mentioned above secure the achievement of a high speedwriting within half a year.

A pilot self-testing gives promise to believe that *Scorapice* may come in useful for a wide range of specialists concerned with writing:

- Non contradictory system of simple rules based on the previously consolidated material.
- The adequate choice of exercises with keys facilitating the control and profound practice for self-study.
- Low distortion of signs in speed writing which makes *Scorapice* records sufficiently readable after a long period of time.
- The hand follows habitual movements thanks to cursive writing with minimal backward movements of the pen in the course of speed writing.
- The full course of training may take up to six months to secure a speed of 100 wpm and more.
- Writing from dictation at various speed, which is available on the internet, will facilitate attaining a high level performance.

It would be a good idea to take a second look at the course of *Scorapice* after its practical evaluation by learners.

An example of text written in *Scorapice* is shown below.

I love you, citadel of Peter's,	_____
I love your elegant austerity of line,	_____
Your broad Neva, whose gracious waters	_____
Mid granite-clad embankment shine.	_____
I love your tracery iron of gates and railing,	_____
The moonless brilliance of the light	_____
That sheds a beauty never paling	_____
Upon your meditative night.	_____
When, neither lamp nor candle needing	_____
I sit composing verse or reading,	_____
And slumbering mansions towering high,	_____
Stand clear against the lucid sky.	_____
A. Pushkin, <i>The Bronze Horseman</i> _____	

ЛИТЕРАТУРА

1. Соколов, Н. Н. Теоретические основы государственной единой системы стенографии / Н. Н. Соколов. – М. : Учпедгиз, 1949. – 195 с.
2. Прохорская, Н. Ф. Задания для заочного обучения стенографии. Основной курс / Н. Ф. Прохорская, Н. П. Скородумова, А. М. Юрковский. – М., 1963. – 242 с.
3. Юрковский, А. М. Стенография сквозь века / А. М. Юрковский. – М. : Изд-во «Книга», 1969. – 163 с.
4. Leech, G. *Word Frequencies in Written and Spoken English: based on the British National Corpus* / Geoffrey Leech, Paul Rayson, Andrew Wilson. – London : Longman, 2001. – 320 p.
5. Частотный список 5-членных словосочетаний в английском языке [Электронный ресурс]. – Режим доступа : <http://web.archive.org/web/2007082-8004603/http://www.unitrier.de/uni/fb2/anglistik/Projekte/stubbs/icame-2004.htm>. – Дата доступа : 02.06.2016.
6. The British National Corpus, version 3 (BNC XML Edition) [Electronic resource] // Distributed by Oxford University Computing Services on behalf of the BNC Consortium. – Mode of access : <http://www.natcorp.ox.ac.uk>; http://ucrel.lancs.ac.uk/bncfreq/lists/2_3_writtenspoken.txt; BNCweb (CQP-Edition) Version 4.3, November 2013 <http://bncweb.lancs.ac.uk/cgi-bin/bnc-XML/BNCquery.pl?theQuery=search&urlTest=yes>. – Data of access : 10.06.2016.

СОЗДАНИЕ СЛОВАРЕЙ РАЗЛИЧНЫХ ТИПОВ

Л. Н. Беляева, К. В. Новикова

ПЕРЕВОДНОЙ СЛОВАРЬ В СИНЕРГЕТИЧЕСКОМ АСПЕКТЕ: ПОТЕНЦИАЛ И ПРИМЕНЕНИЕ

Синергетический подход к исследованию текста, вполне принятый в современной лингвистике и дающий интересные результаты, предполагает специальное исследование механизмов организации и самоорганизации текста. Основным условием его применения является понимание текста как сложной системы со своими компонентами, внешней организацией и структурой связей между компонентами, которая отражается на лексическом, синтаксическом и семантическом уровнях. Внешними симптомами работы синергетических механизмов, определяющих существование текста как равновесной системы, являются баланс и зависимости между лексической структурой текста, количественными параметрами отдельных лексических единиц и характеристиками энтропии текста [1].

При этом текст рассматривается как результат решения задачи передачи информации и источник (отправную точку) ее извлечения. Наши знания не только фиксируются и представляются в виде текстов, но и порождаются в языковом смысле как текст [2]. Соответственно, содержание научного текста, то есть 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 ситуации перевода возможность извлечения информации из научного текста определяется для читателя корректностью передачи терминов – имен объектов. При этом адекватность понимания текста на лексическом уровне определяется насыщенностью текста именными единицами, степенью компрессии и/или развернутостью номинации объектов, точностью отражения связей между ними.

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 затрудняется или даже становится невозможным даже для специалиста в соответствующей предметной области. Учет сложности восприятия и понимания подобных лексических образований, их упрощение и уточнение позволяет создавать адекватный текст перевода, то есть от хаоса нагромождения терминологических единиц перейти к новому состоянию, отражающему структуру денотатного пространства текста и терминосистему конкретного языка в том числе [3].

Сегодня можно утверждать как некоторую аксиому, что объективное исследование терминологии конкретной предметной области требует ее предварительного моделирования для того, чтобы установить систему понятий и связей между ними. При этом принято различать терминопole как системное образование плана содержания, представляющее собой организованную совокупность специальных понятий и связей, и терминосистему, т.е. совокупность языковых средств, отражающих исследуемое терминопole. Наиболее традиционным подходом к изучению соотношения терминопoleй и отражающих их терминосистем является структурное моделирование на основе тезауруса, создаваемого на основе лексики, определяемой современным состоянием рассматриваемого терминопoleа и соответствующей ему терминосистемы. При этом терминосистема должна извлекаться из словарного и текстового материала на одном или нескольких языках с учетом постоянного развития различных научных и технических сфер деятельности человека и его отражения в языке. В современном понимании терминопoleа как концептуальная основа терминосистемы может рассматриваться как онтология [4].

Онтология отличается от тезауруса тем, что содержит только не зависящую от языка информацию и множество семантических отношений, кроме того, она содержит таксономические отношения. Онтология должна задавать понятия для представления значений слова в лексиконе [5] и ее можно рассматривать как пространственное представление семантической сети – сферу, в узлах внешней сетчатой поверхности которой располагаются терминологические единицы разных языков. Эти термины связаны с узлами–понятиями, располагающимися внутри (на следующем уровне) сферы. Узлы–понятия, в свою очередь, могут быть связаны с концептами, отражающими факты внешнего мира и/или нашими представлениями о них и хранящимися на более глубоких уровнях (слоях) сферы.

Синергетический подход к процессу развития терминологии позволяет утверждать, что при активном развитии предметной области в терминосистеме конкретного языка возникают точки дифуркации, которые требуют пересмотра зафиксированной терминосистемы для того, чтобы она реально отражала соответствующую систему понятий. Будем исходить из того, что лексика языка представляет собой систему, что подтверждается наличием структуры связей между ее компонентами и изменением этой структуры при введении или удалении некоторого отдельного компонента или набора компонентов. Тогда лексика языка для специальных целей представляет собой подсистему этой макросистемы, в которой заданы терминологические единицы этого языка (универбы и словосочетания) и отношения между ними, формирующие структуру связей между компонентами.

При создании переводного словаря, ориентированного на узкую предметную область, происходит слияние двух систем, в основе которого лежит единство системы номинируемых понятий, т.е. онтология. Практический смысл создания онтологии заключается в разработке систем для хранения и

поиска фрагментов знания; так, например, онтологическое понимание предмета (сущности) позволяет понять взаимозависимость определенных базовых понятий, а также связанных с ними терминов [6].

При таком подходе переводной словарь также представляет собой сложную систему, которая развивается, отражая изменения в исходных терминосистемах, и достигает точки диффракции, предшествующей его распаду и разбалансировке, то есть приводит к несоответствию задачам переводчика и/или специалиста. Такое состояние словаря характеризуется несколькими параметрами, степень реализации которых должна устанавливаться экспериментально на основе анализа комплекса словарей, оцениваемых специалистами как несоответствующие задачам перевода. Предварительно значения таких параметров можно описать следующим образом:

- отсутствие в словаре определенного, экспериментально устанавливаемого процента слов, встречающихся в текстах предметной области (не менее 20 % терминологических единиц);
- наличие в словаре или в отраслевых глоссариях составных терминов, имеющих более 1 перевода (не менее 10 % состава словосочетаний);
- наличие в словаре или в отраслевых глоссариях терминов-универбов, имеющих более 2 переводов (не менее 10 % состава терминов-универбов).

Предложенные выше параметры и их ориентировочные значения отражают нарушение унификации терминов и терминосистем, что является одним из основных направлений прикладного терминоведения, основной задачей которого и является стандартизация, упорядочение и гармонизация терминологий на различных уровнях описания и фиксации [7]. Результатом такой унификации является создание терминологических переводных словарей, ориентированных на узкие предметные области и языки для специальных целей. Такие словари, как правило, являются бинарными, хотя существуют и многоязычные, для нашей страны основную массу переводных словарей составляют те, в которых исходным языком является английский, словари с русским языком в качестве исходного создаются реже. Теоретически двуязычные переводные словари необратимы, поскольку лексические системы языков, рассматриваемые как множества лексических единиц (ЛЕ), несимметричны. Отношение симметричности работает только на множествах номенов. В случае языков для специальных целей вариативность переводов ЛЕ резко уменьшается и можно говорить о введении локальной симметричности отношения между множествами терминов двух языков. Эта локальная симметричность позволяет разрабатывать методы преобразования словарей. Особое значение при этом приобретают процедуры, позволяющие «перевернуть» словарь, в принятой сегодня терминологии – конвертировать [8].

Современное лексикографическое исследование в области создания переводных словарей (как бумажных, так и электронных) предполагает проведение предварительной терминологической работы для отбора и описания

терминологии на разных языках, осуществление гармонизации этих описаний и согласование терминологических систем разных языков. Однако, в области создания и ведения переводных словарей, представляющих собой естественную основу профессиональной компетентности переводчика, ситуация остается сложной. Дело в том, что в различных областях знаний и переводчиками, и специалистами создается огромное количество глоссариев, которые никак не соотносятся друг с другом, фиксируют разные переводы для одних и тех же терминов, а в качестве терминов используют словосочетания самой разной длины и структуры. Поэтому достижение словарем, глоссарием или системой отраслевых глоссариев точки дифуркации является показателем необходимости коренной перестройки переводного лексикографического ресурса.

Различие терминосистем разных языков, в частности, исходного языка и языка перевода определяет необходимость установления и изучения именно пар терминов вида исходный термин – переводной эквивалент, что позволяет выявить расхождения в терминопole и терминосистемах соответствующей предметной области. Установление таких расхождений и определяет дальнейшее упорядочивание, стандартизацию и унификацию терминологий разных языков, гармонизацию терминосистем этих языков, что и обеспечивает решение проблем перевода терминов и эффективность межъязыковой коммуникации [9].

Международная стандартизация в специальных областях знаний определяет требования к уточнению значений терминов, более того, их точных определений, которые могут использоваться в структуре терминов и соответствующих описаний с учетом номинируемых понятий. Поэтому сегодня в различных областях знаний и переводчиками, и специалистами создается огромное количество глоссариев, которые отражают представления переводчиков и специалистов, но часто не соотносятся друг с другом, а в качестве терминов используют словосочетания самой разной длины и структуры.

Гармонизация терминологии особенно важна для областей повышенного риска, к которым, например, относятся предметные области сейсмобезопасности и сейсмостойкого строительства. В соответствии с ситуацией в этих предметных областях переводные словари должны разрабатываться с учетом терминологии международного стандарта на проектирование конструкций в сейсмоопасной зоне Еврокод8 (Eurocode8: Design of structures for earthquake resistance). Стандарт был утвержден Европейским комитетом по стандартизации (CEN) 23 апреля 2004 года и включает основные положения и требования предыдущих стандартов.

Для решения задачи гармонизации этой терминологии рядом ведущих исследовательских и проектных институтов нашей страны были разработаны и изданы словари, глоссарии и правила.

- Проектирование и строительство. Понятийно-терминологический словарь к Еврокодам EN 1992 ÷ EN 1996, EN 1998, EN 1999. Рекомендации национального объединения строителей. М.: Р НОСТРОЙ, 2014. – 102 с.

- Англо-русский словарь по проектированию строительных конструкций (в порядке возрастания номеров Еврокодов). М.: Межгосударственная научно-техническая комиссия по стандартизации, техническому нормированию и оценке соответствия в строительстве (МНТКС), 2011. – 35 с.

- Англо-русский словарь по проектированию строительных конструкций (в алфавитном порядке). М.: Межгосударственная научно-техническая комиссия по стандартизации, техническому нормированию и оценке соответствия в строительстве (МНТКС), 2011. – 29 с.

- Свод правил. Здания сейсмостойкие и сейсмоизолированные. Правила проектирования. Издание официальное. М.: Министерство регионального развития Российской Федерации, 2014. – 51 с.

- Свод правил. Строительство в сейсмических районах. Правила проектирования. Издание официальное. М.: Министерство регионального развития Российской Федерации, 2014. – 85 с.

- Терминологический словарь для национальных нормативных документов, реализующих Еврокоды. М.: ЗАО «ЦНИИПСК им. Мельникова», 2014. – 199 с.

Рассматриваемые материалы использовались как база для создания исследовательского корпуса текстов, в котором выравнивание осуществляется по ЛЕ, рассматриваемым авторами перечисленных изданий как термины, а также по толкованиям терминов и их переводам. На основе такого выровненного корпуса могут извлекаться и исследоваться кандидаты в термины, которые при последующем анализе могут вводиться в действительно гармонизированный переводной словарь.

Для создания словарей была применена следующая процедура.

- 1) Построение параллельного корпуса на материале объединения глоссариев и терминологических словарей. В основе такого объединения лежат базовые словари, разработанные Межгосударственной научно-технической комиссией по стандартизации, техническому нормированию и оценке соответствия в строительстве (МНТКС), эти словари были дополнены информацией из всех остальных словарей (табл. 1).

- 2) Составление единого словарного массива (индекса) на основе сопоставления английских и русских лексических единиц (табл. 2), выявление кандидатов в английские термины на основе использования программы выделения простых именных групп.

- 3) Получение русско-английского словаря на основе конвертации англо-русского словаря, дополненного кандидатами в термины, дополнение словаря кандидатами в термины – простыми именными группами.

- 4) Работа с экспертами и уточнение списка исходных терминов и их переводов, фиксирование отношений синонимии и нормализация исходных терминов и переводов.

Англо-русский объединенный выровненный словарь (фрагмент)

Термин на англий- ском языке	Номер Еврокода, обозначе- ние	Определение термина на английском языке	Термин на русском языке	Определение термина на русском языке	Источник
adhesion:	EN 1996-1-1	the effect of mortar developing a tensile and shear resistance at the contact surface of masonry units	адгезия: адгезионная прочность сцепления кладки (адгезия)	Сопротивление растяжению или срезу (сдвигу), возникающее на контактной поверхности между раствором и элементами кладки Сопротивление растяжению или срезу (сдвигу), возникающее между раствором и контактной поверхностью кладки из камней или блоков.	Р НОСТРОЙ МНТКС
aerodynamic device:	EN 1993-3-2	A device fitted to the chimney to reduce vortex excitation without increasing the structural damping	аэродинамическое устройство:	Устройство, которое устанавливается на дымовую трубу для уменьшения вихревого возбуждения без увеличения конструкционного демпфирования.	Р НОСТРОЙ
age at loading	EN 1994-1-1, t_0		момент нагружения	Момент нагружения	МНТКС
age of concrete at the time of loading	EN 1992-2, t_0		возраст бетона на момент нагружения	Возраст бетона в момент нагружения	МНТКС
	EN 1992-1-1, t_0		возраст бетона в момент приложения нагрузки	Возраст бетона в момент приложения нагрузки	МНТКС
along-wind vibrations:	EN 1993-3-2		колебания вдоль направления ветрового воздействия:	Продольные колебания дымовой трубы, вызванные эффектом галопирования	Р НОСТРОЙ

Англо-русский словарь (фрагмент)

acceleration of gravity	ускорение свободного падения
	ускорение силы тяжести
accessory masonry unit	дополнительный элемент кладки
accidence	превышение
accidental eccentric	максимально разрешенные случайные эксцентриситеты
accidental eccentricity	случайный эксцентриситет
accidental eccentricity of the mass of one storey from its nominal location	случайный эксцентриситет массы одного этажа от его номинального положения
accidental eccentrics	случайный эксцентриситет
accidental mass eccentricity	случайный эксцентриситет масс
acting shear force	расчетная горизонтальная сила
	действующее усилие сдвига
action	воздействие
	действие
	воздействие, вызывающее изменение напряженно-деформированного состояния конструкции
action effect	эффект расчетного воздействия
	эффект воздействия
	результат расчетного воздействия
action in the seismic design situation	воздействие в расчетной сейсмической ситуации

При анализе словарей установлено большое число случаев, когда в перевод включается экстралингвистическая информация, известная специалисту в конкретной области, но отсутствующая в структуре исходного термина. Так, например, словосочетание *clearance height from roadway surfacing* переводится как *габаритная высота от поверхности дорожного полотна до нижней кромки конструкции моста*. При этом перевод термина *roadway surfacing* дополняется уточнением *до нижней кромки конструкции моста*, что связано с особенностями строительства именно мостовых и/или арочных строений. Для английского словосочетания это является имплицитной информацией, не требующей экспликации. Подобная словарная экспликация терминов, взятых вне контекста, может оказаться некорректной. Так, например, перевод терминологического словосочетания *coefficient of earth pressure at rest*, уточненный как *коэффициент бокового давления грунта в состоянии покоя*, верен только для случая описания объекта, не полностью погруженного в грунт, во всех иных случаях это давление не будут боковым и, следовательно, перевод термина не будет адекватным.

Еще более сложный случай возникает в случае не перевода, а передачи коллокации, описывающей конкретное важное явление или сложный объект, см., например следующее соотношение исходного описания и его перевода на русский язык: *correlation factors to derive the pile resistance from ground investigation results, not being pile load tests* – *неправочные коэффициенты для оценки результатов испытаний свай статической нагрузкой*. При этом в рассматриваемых глоссариях не зафиксированы терминологические слово-

сочетания *pile resistance, ground investigation, pile load, pile load test*, введение которых диктуется особенностями их перевода. Предлагаемый в глоссарии перевод может объясняться тем, что подобные грунтовые испытания свай обычно проводятся именно под статической нагрузкой. Аналогично *LE diameter of confined core in a circular column* переводится как *диаметр бетонного сечения, замкнутого поперечной арматурой в круглой колонне*.

Количественно такие термины составляют не менее 10 % от объема словаря, что характеризует недостаточную лингвистическую компетентность их составителей. В то же время подобные глоссарии представляют собой важный источник для разработки соответствующего англо-русского переводного словаря, поскольку фиксируют не только современную терминологию, но и экспертное знание.

Исходный вариант русско-английского словаря автоматически сгенерирован из дополненного и отредатированного экспертами англо-русского словаря с использованием основных средств операционных систем, в частности, средств работы с таблицами. Полученный словарь (индекс) представляет собой таблицу русских терминов и их переводов, отсортированных в алфавитном порядке. Процесс редактирования состоял в изменении, удалении или добавлении заглавных единиц, соотнесении и переупорядочении их связей с английскими эквивалентами. При этом (как и в случае работы с англо-русским вариантом) модифицировались конструкции, представляющие собой дефиниции, толкования, описания ситуации, и устанавливались термины, из которых они состоят, определялся весь комплекс переводов конкретных терминов и выявлялась синонимия переводов или многозначность исходного термина (см. табл. 3).

Т а б л и ц а 3

Русско-английский словарь (фрагмент)

Термин на русском языке	Термин на английском языке
поверхность подпорного сооружения	retained earth surface
поворот в пластическом шарнире	plastic hinge rotation
повышенная температура	elevate temperature
повышенная температура	elevated temperature
погружаемость	driveability
погружение	driving
податливые элементы	ductile members
податливый бетонный элемент	ductile concrete member
податливый элемент	ductile member
подкос	strut
подкрепленная пластина	stiffened plate
подкрепляющая стена	stiffening wall
подошва	footing
подошва фундамента	footing
подпорная конструкция или элемент	propped structure or member
подпорная стена	retaining wall
подпорное сооружение	retaining works
поверхность подпорного сооружения	retained earth surface
поворот в пластическом шарнире	plastic hinge rotation
повышенная температура	elevated temperature
погружаемость	driveability
погружение	driving

Полученный комплекс словарей исследовался с точки зрения установленной степени многозначности словарных статей (см. табл. 4). Из таблицы следует, что уровень многозначности выше для англо-русского варианта, который построен на базе исходного объединенного словаря с применением процедуры добавления простых именных групп, выделяемых из многокомпонентных словосочетаний. Русско-английский словарь, полученный в результате конвертации, содержит меньше многозначных многокомпонентных словосочетаний. Результаты показывают, что англо-русский словарь требует дополнительного анализа и обработки.

Т а б л и ц а 4

Количественные характеристики исследуемого комплекса отраслевых словарей

Словарь	Объем словаря	Количество терминов-универбов	Количество многозначных универбов	Количество словосочетаний	Количество многозначных словосочетаний
Англо-русский	2830	382	88 (23,3%)	2448	290 (11,4%)
Русско-английский	3105	304	56 (18,4%)	2801	90 (3,21%)

Полученные таким образом переводные словари должны вручную тщательно редактироваться, но объем этой работы не сопоставим с трудоемкостью создания переводного словаря вручную. При этом следует учитывать, что даже самая изоцированная система извлечения терминов не дает окончательного варианта переводного словаря, а предоставляет лишь удобно организованный и оперативно получаемый ресурс для работы терминоведа или лексикографа.

Моделирование динамических систем, к которым можно отнести и сам язык, и речемыслительную деятельность в целом, и терминологические системы языков для специальных целей, представляет собой сложную задачу, требующую решения задач категоризации, а также применения качественных и количественных (вероятностных) оценок создаваемых моделей. Кроме того, моделирование требует проверки адекватности разработанных моделей при условии использования при их создании субъективного опыта исследователя.

ЛИТЕРАТУРА

1. *Пиотровский, Р. Г.* Лингвистическая синергетика: исходные положения, первые результаты, перспективы / Р. Г. Пиотровский – СПб : Филологический ф-т СПбГУ, 2006. – 160 с.
2. *Чернявская, В. Е.* Лингвистика текста. Лингвистика дискурса / В. Е. Чернявская. – М. : УРСС-Либроком, 2013. – 224 с.
3. *Беляева, Л. Н.* Синергетика текста и адекватность перевода терминологии / Л. Н. Беляева, О. А. Бородин // Синергетическая лингвистика vs лингвистическая синергетика : матер. Междунар. науч.-практич. конф. – Пермь : Изд-во Пермского гос. технич. ун-та, 2010. – С. 119–126.

4. *Cimiano, P.* Linked Terminologies: Applying Linked Data Principles to Terminological Resources / P. Cimiano [etc.] // Proc. of the fourth biennial conference on electronic lexicography, eLex 2015 : Linking lexical data in the digital age. Sussex, United Kingdom, from 11–13 Aug. 2015. – United Kingdom, 2015. – P. 504–517
5. *Biebow, B.* TERMINAE : a method and a tool to build a domain ontology / B. Biebow, S. Szulman // Benjamins, V.R., Fensel, D. and Gomez-Perez, A. (eds) Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure – Dagstuhl Castle, Germany, 1999. – P. 25–30.
6. *Gillam, L.* Terminology and the construction of ontology / L. Gillam, M. Tariq, K. Ahmad // Application-Driven Terminology Engineering. Edited by Fidelia Ibekwe-SanJuan, Anne Condamines and M. Teresa Cabré Castellví. – Terminology, 2005. – P. 55–81.
7. *Лейчик, В. М.* Прикладное терминоведение и его направления / В. М. Лейчик // Прикладное языкознание: учебник; под. ред. А. С. Герда. – СПб. : Изд-во С.-Петерб. ун-та, 1996. – С. 276–286.
8. *Egorova, K.* Editing an automatically-generated index with K Index Editing Tool / K. Egorova // Proc. of the fourth biennial conference on electronic lexicography, eLex 2015 : Linking lexical data in the digital age. Sussex, United Kingdom, from 11–13 Aug. 2015. – United Kingdom, 2015. – P. 268–280.
9. *Беляева, Л. Н.* Лексикографический потенциал современных лингвистических технологий : монография / Беляева Л. Н. [и др.]. – СПб : ООО «Книжный дом», 2014. – 168 с.

**В. А. Василевская, А. Е. Жданович,
Е. И. Ильина, Е. В. Марковская**

СПЕЦИФИКА ОРГАНИЗАЦИИ ИНФОРМАЦИИ В АНГЛО-РУССКИХ ТЕРМИНОЛОГИЧЕСКИХ СЛОВАРЯХ ПО ИНФОРМАЦИОННЫМ ТЕХНОЛОГИЯМ

В работе “Компьютерные языковые ресурсы...” предложен единый шаблон структурированного описания терминологических словарей, содержащий 9 конкретных характеристик словаря и его составляющих [1, с. 81–86]. Для выявления специфики организации информации в англо-русских терминологических словарях по информационным технологиям нами были выбраны следующие характеристики:

1. Полные выходные данные о словаре (полное библиографическое описание с указанием издательства и количества страниц).
2. Общая характеристика словаря, включающая:
 - 2.1. Предметную область, нашедшую отражение в словаре.
 - 2.2. Языки, нашедшие отражение в словаре.
 - 2.3. Тип словаря (энциклопедический, толковый, переводной).
 - 2.4. Объем словаря (количество словарных статей).
3. Характеристики словарной статьи:
 - 3.1. Заглавный элемент (однословный термин, составной термин, имя собственное, сокращение).
 - 3.2. Наиболее полная структура словарной статьи.
 - 3.3. Наличие толкования (дефиниции) (развернутого описания понятия, выражаемого в словарной статье).

По этой схеме в процессе создания англо-белорусского терминологического словаря по информационным технологиям были рассмотрены следующие опубликованные англо-русские терминологические словари по информатике, информационным технологиям и вычислительной технике.

А. 1. Вычислительная техника и новые информационные технологии. Русско-английский терминологический словарь. – М., 1992.

2. Общая характеристика словаря.

2.1. Предметная область – вычислительная техника и новые информационные технологии.

2.2. Языки – русский, английский.

2.3. Тип словаря – переводной.

2.4. Объем словаря – 1 400 терминов.

3. Характеристики словарной статьи.

3.1. Заглавный элемент – термин, составной термин или терминологическое словосочетание на русском языке и их перевод на английский язык.

3.2. В ряде случаев может быть использована краткая форма заглавного термина (к.ф.), а также его синонимы.

3.3. В словаре имеется «Англо-русский алфавитный указатель», предназначенный для поиска термина по английскому эквиваленту.

Например:

алгоритм

algorithm

Предписанная совокупность
четко определенных правил
для решения задачи за конечное
число шагов

Б. 1. Англо-русский словарь по информационным технологиям. – М.: ИП «РадиоСофт», 2010. – 640 с.

2. Общая характеристика словаря.

2.1. Предметная область – информационные технологии.

2.2. Языки – английский, русский.

2.3. Тип словаря – переводной.

2.4. Объем словаря – 60 000 терминов.

3. Характеристики словарной статьи.

3.1. Заглавный элемент – термин, составной термин или сокращение на английском языке и их перевод на русский язык.

3.2. В некоторых случаях указываются русские переводные синонимы.

3.3. Расшифровка английских сокращений приводится в Приложении.

Например:

abbreviated text сокращенный текст

abbreviation сокращение; сокращенное наименование; аббревиатура

ABC 1. классы построения приложений; 2. функционально-стоимостной анализ; 3. оплата на базе фактического использования; вычисление расходов на основе выполненных действий; 4. аналоговый периферийный элемент

ABCA Американская ассоциация по коммуникациям в сфере бизнеса

В. 1. Борисевич М. Н., Лукина Л. В. Экспресс-словарь компьютерных терминов и сокращений. Уч.-методическое пособие. – Витебск: УО ВГАВМ, 2008.

2. Общая характеристика словаря.

2.1. Предметная область – понятия, связанные с практикой компьютерной деятельности (использование различных компьютерных программ, Интернет).

2.2. Языки – русский, английский.

2.3. Тип словаря – переводной.

2.4. Объем словаря – 200 терминов и сокращений.

3. Характеристик словарной статьи.

3.1. Заглавный элемент – русский термин или составной термин, сокращение.

3.2. В некоторых случаях указываются русские переводные синонимы.

3.3. Расшифровка английских сокращений приводится в Приложении.

Например:

Адоб Фотошоп

Adobe Photoshop – один из наиболее популярных растровых графических редакторов. Многофункционален. используется во всех дизайнерских студиях и конторах для работы, поэтому, если вы захотите посвятить себя профессии дизайнера рекламы или веб-дизайн, то знание Adobe Photoshop вам необходимо.

Аккаунт

1. Account – запись о пользователе, которая создается на сервере какой-либо он-лайн службы, при вашей там регистрации.

2. Надпись «введите свой аккаунт» или «аккаунт» подразумевает ваш логин (имя пользователя), которое вам необходимо ввести, чтобы зайти на сайт он-лайн службы, допустим, для чтения писем в своем почтовом ящике (e-mail).

Г. 1. Борковский А. Б., Зайчик Б. И., Боровикова Л. И. Словарь по программированию (английский, русский, немецкий, французский). – М.: Рус. язык, 1001. – 286 с.

2. Общая характеристика словаря.

2.1. Предметная область – программирование.

2.2. Языки – английский, русский, немецкий, французский.

2.3. Тип словаря – переводной.

2.4. Объем словаря – 5 000 терминов.

3. Характеристики словарной статьи.

3.1. Заглавный элемент – термины или составные термины на английском языке.

3.2. Все русские, немецкие и французские термины имеют указание рода (*m*, *f*, *n*) и, при необходимости, множественного числа (*pl*). Все английские термины пронумерованы в пределах каждой из букв английского алфавита для возможности отыскания русских, немецких и французских терминов-эквивалентов по указателям.

В переводах принята следующая система разделительных знаков: близкие по значению эквиваленты отделены запятой, более далекие – точкой с запятой, разные значения – цифрами.

Синонимичные варианты части перевода заключены в квадратные скобки, например: дескриптор [паспорт] массива. Термин следует читать: «дескриптор массива», «паспорт массива».

Факультативная часть термина дается в круглых скобках, например: маркер начала (магнитной) ленты. Термин следует читать: «маркер начала магнитной ленты», «маркер начала ленты».

Пояснения к терминам-эквивалентам набраны курсивом и заключены в круглые скобки, например: ответ (*в протоколе передачи данных*).

- 3.3. Для перевода с русского, немецкого и французского языков следует пользоваться помещенными в конце словаря указателями, в которых все термины имеют буквенно-цифровое обозначение, соответствующее обозначению при ведущем английском термине.

Например:

A2	<i>e</i>	abort	A11
	<i>r</i>	прерывать выполнение программы; выбрасывать задачу из решения	
	<i>d</i>	abbrechen (<i>von laufendem Programm wegen Programmfehler oder Fehler in der Datenverarbeitungsanlage</i>), aussteigen	A12
	<i>f</i>	avorter	
A3	<i>e</i>	absolute address	
	<i>r</i>	1. абсолютный адрес <i>m</i> 2. физический адрес <i>m</i> 3. машинный адрес <i>m</i>	A13
	<i>d</i>	absolute Adresse <i>f</i> , wirkliche Adresse <i>f</i> , Maschinenadresse <i>f</i>	
	<i>f</i>	adresse <i>f</i> absolue	

Д. 1. Roman Hajczuk, Altxander Zubov. English-Polish-Russian-Belorusian dictionary on informatics and computer technology. – Białystok, 2003. – 835 p.

2. Общая характеристика словаря.

- 2.1. Предметная область – все разделы информатики и информационных технологий, аппаратные и программные средства.
- 2.2. Языки – английский, польский, русский, белорусский.
- 2.3. Тип словаря – переводной.
- 2.4. Объем словаря – 11 000 терминов, составных терминов, терминологических сочетаний и аббревиатур.

3. Характеристики словарной статьи.

- 3.1. Заглавный элемент – термин, составной термин, терминологическое сочетание, сокращение.
- 3.2. Наиболее полная структура словарной статьи.

Словарные статьи словаря начинаются с английских слов (терминов, составных терминов, словосочетаний, аббревиатур), расположенных в алфавитном порядке. Следующая строка статьи представляет собой один или несколько переводных эквивалентов

английских единиц на польский язык. Синонимичные переводные эквиваленты при этом отделяются друг от друга знаком «точка с запятой». В следующей строке представлены переводы английских единиц на русский язык. И, наконец, последняя строка словарной статьи дает белорусские переводные эквиваленты английских терминов, терминологических словосочетаний или аббревиатур. При наличии среди русских и белорусских переводных эквивалентов синонимичных выражений они отделяются друг от друга также знаком «точка с запятой».

- 3.3. В конце словаря для всех польских, русских и белорусских терминов, составных терминов, терминологических сочетаний и аббревиатур приведены числовые индексы, показывающие порядковый номер словарной статьи, содержащей соответствующую английскую единицу.

Например:

- 16. **abnormal**
niezwykły, nienormalny
ненормальный; аномальный
ненармальные; анамальные
- 17. **abnormal end = abend**
zakończenie nieprawidłowe (awaryjne)
аварийное завершение
аварыйнае спыненне
- 18. **abnormal termination**
nieprawidłowe zakończenie
аварийное завершение; непредусмотренное
завершение
аварыйнае заканчэнне; неспадзяванае
заканчэнне

ЛИТЕРАТУРА

1. Компьютерные языковые ресурсы на материале специальной лексики белорусского языка / Л. В. Рычкова [и др.]. – Гродно : ГрГУ им. Я. Купалы, 2013. – 285 с.

**Е. С. Василенко, Н. В. Зыгмантович,
Т. А. Филимонова, Н. В. Шалимо**

СПЕЦИФИКА ОРГАНИЗАЦИИ ИНФОРМАЦИИ В РУССКО-БЕЛОРУССКИХ ТЕРМИНОЛОГИЧЕСКИХ СЛОВАРЯХ ПО ИНФОРМАЦИОННЫМ ТЕХНОЛОГИЯМ

На кафедре информатики и прикладной лингвистики по заданию НАН Республики Беларусь создается автоматический англо-белорусский терминологический словарь по информационным технологиям. Так как в настоящее время создано на разных языках достаточно много терминологических сло-

варей по информационным технологиям было принято решение терминологической лексики создаваемого англо-белорусского словаря использовать лексику, включенную в англо-русские и русско-белорусские терминологические словари по информационным технологиям. Для этого необходимо детально изучить принципы организации информации в существующих терминологических словарях по информационным технологиям. В работе “Компьютерные языковые ресурсы...” предложен достаточно удобный шаблон структурированного описания таких словарей. Он включает 9 конкретных характеристик словаря и его составляющих [1, с. 81–86].

Для выявления специфики организации информации в русско-белорусских терминологических словарях по информационным технологиям из предложенных девяти характеристик нами были использованы следующие три характеристики:

1. Полные выходные данные о словаре (полное библиографическое описание с указанием издательства и количества страниц).
2. Общая характеристика словаря, включающая:
 - 2.1. предметную область, нашедшую отражение в словаре;
 - 2.2. языки, нашедшие отражение в словаре;
 - 2.3. тип словаря (энциклопедический, толковый, переводной);
 - 2.4. объем словаря (количество словарных статей).
3. Характеристики словарной статьи:
 - 3.1. заглавный элемент (однословный термин, составной термин, имя собственное, сокращение);
 - 3.2. наиболее полная структура словарной статьи;
 - 3.3. наличие толкования (дефиниции, развернутого описания понятия, выражаемого в словарной статье).

По такой схеме ниже приводятся данные о нескольких опубликованных русско-белорусских и белорусских словарях по информационным технологиям.

А. 1. Кабайла А. С., Савіцкая Н. Я. Англа-руска-беларускі слоўнік асноўных скарачэнняў у галіне інфарматыкі. – Мінск: БДТУ, 2009. – 70 с.

2. Общая характеристика словаря.
 - 2.1. Предметная область: информационные технологии.
 - 2.2. Языки: английский, белорусский, русский.
 - 2.3. Тип словаря – переводной.
 - 2.4. Объем словаря – около 1 000 аббревиатур.
3. Характеристика словарной статьи.
 - 3.1. Заглавный элемент – сокращение на английском языке.
 - 3.2. Наиболее полная структура словарной статьи.

Вслед за сокращением на английском языке приводится ее расшифровка на английском языке (курсивом) и перевод на русском и белорусском языках.

- 3.3. В отдельных случаях приводится альтернативный вариант перевода в квадратных скобках, при необходимости – разъяснения или комментарии в скобках на русском и белорусском языках.

Например:

ABI *Application Binary Interface*,
двоичный интерфейс приложений

двайковы шнэртэрфейс прыкладных
[дастасоўных] праграм (дадаткаў [даста-
саванняў])

ACD

1. *automatic call distribution*
автоматическое распределение вызо-
вов (в системах телефонной связи)

аўтаматычнае размеркаванне выклікаў
(у сістэмах тэлефоннай сувязі)

2. *automatic call distributor*
устройство автоматического распе-
деления вызовов

устройства [уладжанне] аўтаматычнага
размеркавання выклікаў

Б. 1. Савіцкі М. І. Тлумачальны слоўнік па інфарматыцы. – Мінск: “МЕДЫЯЛ”, 2014. – 418 с.

2. Общая характеристика словаря.

2.1. Предметная область – информатика и вычислительная техника.

2.2. Языки, нашедшие отражение в словаре – белорусский и английский.

2.3. Тип словаря – толковый.

2.4. Объем словаря – более 3 000 наиболее употребляемых терминов на белорусском языке.

3. Характеристика словарной статьи.

3.1. Заглавный элемент – термины или составные термины на бело-
русском языке.

3.2. Наиболее полная структура словарной статьи.

Вслед за термином на белорусском языке (в круглых скобках)
приводится его перевод на английском языке.

3.3. Значение термина разъясняется на белорусском языке после
английского перевода.

Например:

Абязязковае аб’яўленне (mandatory declaration) – аб’яўленне зменных,
якое нельга апускаць.

Абанэнт (abonent, user, subscriber) – юрыдычная альбо фізічная асоба, якая мае
права на карыстаньне паслугамі вылічальнай сеткі ў якасьці яе вонкавых альбо
канцавых карытсальнікаў.

В. 1. Толковый русско-англо-белорусский словарь по информатике. Под
общ. ред. профессора М. К. Бузы. – Минск: Вышэйшая школа, 1994. – 166 с.

2. Общая характеристика словаря.

2.1. Предметная область – информатика и вычислительная техника.

2.2. Языки, нашедшие отражение в словаре – русский, английский,
белорусский.

2.3. Тип словаря – толковый.

2.4. Объем словаря – около 2 000 терминов.

3. Характеристика словарной статьи.

3.1. Заглавный элемент – русский термин или составной термин.

3.2. Наиболее полная структура словарной статьи.

Вслед за русским термином указывается его перевод на
английский язык, затем – на белорусский язык.

3.3. За беларускім пераводам даецца тлумачэнне беларускага тэрміна на беларускай мове.

Напрыклад:

А. 015. Автоматизация – automation – аўтаматызацыя выкарыстанне ЭВМ і іншых аўтаматычных прыбораў для выканання рознастайных аперацый у вытворчай дзейнасці чалавека.

А. 016. Автономность – autonomy – аўтаномнасць незалежнасць аб'екта ад іншых аб'ектаў у складзе пэўнай сістэмы аб'ектаў або магчымасць яго незалежнага функцыявання.

А. 017. Агрегат – aggregate – агрэгат сукупнасць неабавязкова аднародных аб'ектаў, якія састаўляюць адзінае цэлае.

А. 018. Агрегат данных – data aggregate – агрэгат даных іменаваная сукупнасць лагічна звязаных аб'ектаў, апэратыўна даступных, як і ўся сукупнасць; састаўны элемент даных; масіў аўто запіс.

Г. 1. Roman Hajczuk, Altxander Zubov. English-Polish-Russian-Belorusian dictionary on informatics and computer technology. – Białystok, 2003. – 835 p.

2. Общая характеристика словаря.

2.1. Предметная область – все разделы информатики и информационных технологий, аппаратные и программные средства.

2.2. Языки – английский, польский, русский, белорусский.

2.3. Тип словаря – переводной.

2.4. Объем словаря – 11 000 терминов, составных терминов, терминологических сочетаний и аббревиатур.

3. Характеристика словарной статьи.

3.1. Заглавный элемент – термин, составной термин, терминологическое сочетание, сокращение.

3.2. Наиболее полная структура словарной статьи.

Словарные статьи словаря начинаются с английских слов (терминов, составных терминов, словосочетаний, аббревиатур), расположенных в алфавитном порядке. Следующая строка статьи представляет собой один или несколько переводных эквивалентов английских единиц на польский язык. Синонимичные переводные эквиваленты при этом отделяются друг от друга знаком «точка с запятой». В следующей строке представлены переводы английских единиц на русский язык. И, наконец, последняя строка словарной статьи дает белорусские переводные эквиваленты английских терминов, терминологических словосочетаний или аббревиатур. При наличии среди русских и белорусских переводных эквивалентов синонимичных выражений они отделяются друг от друга также знаком «точка с запятой».

3.3. В конце словаря для всех польских, русских и белорусских терминов, составных терминов, терминологических сочетаний и аббревиатур приведены числовые индексы, показывающие порядковый номер словарной статьи, содержащей соответствующую английскую единицу.

Например:

16. **abnormal**
niezwykły, nienormalny
ненормальный; аномальный
ненармальные; анамальные
17. **abnormal end = abend**
zakończenie nieprawidłowe (awaryjne)
аварийное завершение
аварыйнае спыненне
18. **abnormal termination**
nieprawidłowe zakończenie
аварийное завершение; непредусмотренное
завершение
аварыйнае заканчэнне; неспадзяванае
заканчэнне

ЛИТЕРАТУРА

1. Компьютерные языковые ресурсы на материале специальной лексики белорусского языка / Л. В. Рычкова [и др.]. – Гродно : ГрГУ им. Я. Купалы, 2013. – 285 с.

Т. А. Грибач

ПОСТРОЕНИЕ ЛЕКСИЧЕСКИХ СЛОВАРЕЙ ДЛЯ СИСТЕМЫ АВТОМАТИЧЕСКОГО СОЗДАНИЯ ТРАНСПОРТНЫХ ДОКУМЕНТОВ

В данной статье рассматривается вариант построения многоязычных словарей для базы данных, которая является частью системы автоматического порождения транспортных договоров.

На основе статистического анализа текстов русскоязычных, немецкоязычных и англоязычных транспортных документов для каждого вида договора на оказание логистических и экспедиционных услуг и для каждого выше указанного языка были построены словари системы создания транспортных документов. В словаре представлены восемь классов слов: имена существительные, глаголы, имена прилагательные, наречия, местоимения, порядковые и количественные числительные, предлоги и союзы. Словарная статья автоматического словаря системы порождения включает две зоны:

- грамматических сведений (содержит информацию о части речи, к которой относится лексическая единица);
- семантических сведений (включает семантический признак лексической единицы, т.е. ее принадлежность к определенной тематической группе).

Лексика представляет собой системное образование, поэтому при создании словаря лексических единиц текстов транспортных договоров в основу был положен групповой анализ лексики.

В исследованиях [1; 2; 3], посвященных изучению лексической системности, объединение слов в группы происходит на основе определенного сходства или различия, противопоставления их значений. В современной

лингвистике нет однозначного подхода к проблеме классификации значений. Однако большинство ученых рассматривает лексическое значение как «определенную структуру, состоящую из набора мельчайших, далее неделимых семантических компонентов» [4; 5]:

- денотативный компонент является основным для большинства слов и представляет собой предметно-понятийную информацию, связанную с отражением внеязыковой действительности, объективной или субъективной;
- функционально-стилистический компонент характеризует принадлежность слова к тому или иному функциональному стилю речи;
- сигнификативный компонент включает в себе общие существенные признаки, которые являются основанием для объединения отдельных, единичных предметов в определенные классы;
- коннотативный компонент выражает отношение говорящего к предмету номинации в форме эмоции или оценки.

Семантические признаки могут быть разной степени абстракции и обобщенности, вследствие чего слово, обладающее тем или иным семантическим признаком, включается в семантические объединения разного уровня, охватывающие разное количество слов: семантическое поле (СП), лексико-семантическую группу (ЛСГ), тематическую группу (ТГ) и т.д.

В работе Л. А. Новикова семантическое поле определяется как «множество языковых единиц, объединенных общим (инвариантным) значением», а лексико-семантическая группа – как «элементарное семантическое (микро)поле, представляющее собой относительно замкнутый ряд лексических единиц одной и той же части речи, объединенных общей семой, а именно архисемой более конкретного содержания и классификационно более низкого порядка, чем архисема поля» [6]. При этом семантическое поле выступает как лексическая категория высшего порядка, синтезирующая в себе более частные категории (синонимии, антонимии, гипонимии и др.), включающая как однословные лексические единицы, причем различных частей речи, так и словосочетания. Таким образом, ЛСГ являются важнейшими составными компонентами семантического поля.

В отличие от СП и ЛСГ тематические группы слов объединяются не на основе лексико-семантических связей, а на основе экстралингвистических факторов, т.е. на связи с внеязыковой действительностью. Под тематической группой понимается «лексическая группа слов, объединенных общим семантическим содержанием и отражающих связи и отношения между различными предметами и явлениями экстралингвистической действительности и принадлежащих к одной и той же части речи» [7, с. 97].

Классификация лексического материала по тематическим группам – один из наиболее распространенных приемов описания лексики. Выделение ТГ любого языка не представляет особых трудностей, они могут быть сравнительно легко получены на основе сплошной выборки из текстов.

В качестве классификационного признака для выделения ТГ большинство исследователей [1; 3] указывают номинативную общность лексических единиц (в отличие от сигнификативной общности у лексико-семанти-

ческих групп); в работах [3; 7] ее описывают по-разному: как предметно-логическую общность, общность денотативных функций названий, наличие ономаσιологического инварианта.

Таким образом, ТГ своим содержанием в наибольшей степени отражает внеязыковые связи входящих в них слов. Однако это не означает, что тематическая классификация лексики не имеет никакого отношения к лексическому значению слова. Семантическая структура слова включает не только указание на денотат, но и информацию о группе слов, в которую входит данный предмет или явление. Тематические объединения слов наряду с СП и ЛСГ определяются общим характером семантических отношений, присущих каждому конкретному языку, поэтому любая семантическая группа слов, в том числе и ТГ, объединяет лексические единицы, которые вступают во взаимные внутриязыковые, внешнеязыковые и внеязыковые отношения.

Рассмотрение лексики в тематическом аспекте позволяет установить связи между словами и обозначаемыми ими явлениями, определить удельный вес этих групп в словарном составе языка, расширение и сужение значений, установить стилистическую дифференциацию в пределах ТГ, а также выявить некоторую общность их словообразовательных и семантических примет (например, особенности их лексической сочетаемости). Поэтому в нашем исследовании классификация ЛЕ производилась по денотативно-сигнификативному принципу на основании предметно-логической общности обозначаемых ими предметов и явлений. Для отнесения лексических единиц, встречающихся в текстах деловых электронных писем, к той или иной ТГ был использован контекстологический и статистический анализ, а также анализ словарных дефиниций из словарей толковых словарей Oxford Dictionary, Duden и словаря Ожегова.

ЛИТЕРАТУРА

1. Кобозев, И. М. Лингвистическая семантика / И. М. Кобозева – М. : Эдиториал УРСС, 2000. – 352 с.
2. Мартинович, Г. А. Текст и эксперимент: исследование коммуникативно-тематического поля в русском языке / Г. А. Мартинович. – СПб. : СПбГУП, 2008. – 256 с.
3. Штукарева, Е. Б. О соотношении терминов *лексико-семантическое поле* – *лексико-семантическая группа* – *тематическая группа* в лингвистической литературе / Е. Б. Штукарева // Вопросы филологических наук. – 2009. – № 4. – С. 90–93.
4. Васильев, Л. М. Современная лингвистическая семантика : учеб. пособие / Л. М. Васильев. – 2-е изд., доп. – М. : ЛИБРИКОМ, 2009. – 192 с.
5. Левицкий, Ю. А. Проблемы лингвистической семантики / Ю. А. Левицкий. – 2-е изд. – М. : ЛИБРИКОМ, 2009. – 152 с.
6. Новиков, Л. А. Художественный текст и его анализ / Л. А. Новиков. – 2-е изд., испр. – М. : Едиториал УРСС, 2003. – 302 с.
7. Таховский, Н. И. Системность лексики в тематических группах наряду с другими семантическими группировками слов / Н. И. Таховский // Лингвистические единицы разных уровней и их функционал, характеристики : межвуз. сб. науч. тр. / Кубанский гос. ун-т; редкол. : Т. М. Беляева [и др.]. – Краснодар : КГУ, 1982. – С. 94–100.

**ИНТЕГРИРОВАННАЯ ЛЕКСИКОГРАФИЧЕСКАЯ СИСТЕМА
КАК ИНСТРУМЕНТ АВТОМАТИЗИРОВАННОГО ФОРМИРОВАНИЯ
НОВЫХ ЭЛЕКТРОННЫХ СЛОВАРЕЙ И КАК БАЗА ЛИНГВИСТИЧЕСКИХ
ИССЛЕДОВАНИЙ В ОБЛАСТИ ЛЕКСИЧЕСКОЙ СЕМАНТИКИ**

Одной из основных задач интегрированной лексикографической системы (ИЛС) украинского языка, разрабатываемой в Украинском информационно-языковом фонде НАН Украины, является создание надежных информационно-языковых технологий, обеспечивающих исследования в области автоматической переработки языковой информации.

Предполагаемое включение в систему электронных версий существующих словарей современного украинского языка¹, программное интегрирование их между собой, а также создание программных средств экстракции из этих словарей содержащейся в них семантической информации дает право рассматривать ИЛС в качестве надежного базиса для построения лингвистических процессоров автоматического распознавания смысла обрабатываемой информации, а задачу формирования интегрированных лексикографических систем включить в круг задач, связанных с проблемой создания искусственного интеллекта.

Кроме того, программно-технологический комплекс, которым снабжена ИЛС, обеспечивает выполнение ею функции автоматизированного рабочего места (АРМ) для создания новых автоматизированных словарей, в том числе автоматических.

В основе ИЛС лежит теория лексикографических систем В. А. Широкова, в частности ее стержневое положение о наличии определенного лексикографического эффекта в любой информационной системе [1].

Техническим обеспечением ИЛС являются разработанные в Фонде программно-лингвистические технологии автоматизации интеграционных процессов [2].

В качестве базовой лексикографической системы в ИЛС выбрана электронная версия 20-томного Толкового словаря украинского языка (ЭТС), поскольку этот словарь является наиболее репрезентативным источником информации о семантических свойствах лексических единиц украинского языка и представляет собой обобщение опыта работы лексикографов многих поколений в области исследований лексической семантики.

Форма представления данных в ЛБД словаря, строгая формализация описания его макро- и микроструктуры, обеспечивающие возможность обработки данных компьютерными программами, обусловили многофункциональность электронного толкового словаря. С одной стороны, ЭТС является информационно-поисковой системой, принимающей участие в коммуника-

¹ В настоящее время в систему входят электронные версии: Украинского толкового словаря в 20 томах, грамматического, орфоэпического, этимологического, нескольких переводных и терминологических словарей.

ционном процессе человек – компьютер. В этом случае словарь может быть записан на диске, и тогда обращение к нему происходит по требованию пользователя, или он является рабочим компьютерным словарем, и обращение к нему осуществляется через интернет. С другой стороны, ЭТС выполняет функцию резидентного словаря, представляющего собой часть автоматизированного рабочего места (АРМ) лингвиста в процессе формирования ЛБД новых электронных словарей, описывающих различные лексико-грамматические, семантико-синтаксические и семантические контексты лексических единиц украинского языка. То есть речь идет об автоматических электронных словарях – грамматическом (ЭГС) и синтаксическом глагольного управления (ЭГУ), а также автоматизированных электронных семантических словарях синонимов (ЭСС) и паронимов (ЭСП).¹

Предполагается, что в дальнейшем после интеграции в ИЛС вновь созданных электронных словарей и сам ЭТС будет параметризован семантическими характеристиками, экстрагируемыми из этих словарей. Это в значительной мере повысит информационный потенциал электронного толкового словаря в плане расширения и специализации последовательного представления в нем семантических свойств лексической подсистемы украинского языка.

Параметризация однозначных единиц и конкретных значений многозначных единиц словника ЭТС с помощью грамматических словарей (ЭГС и ЭСУ) обеспечивает системное описание лексико-грамматических характеристик слов украинского языка с точки зрения

- частеречной принадлежности слова;
- зависимости между конкретным значением многозначного слова и полнотой/ неполнотой его словоизменительной парадигмы;
- принадлежности к семантико-грамматическим категориям вида и переходности/ непереходности глаголов (для многозначного глагола его конкретных значений);
- принадлежности к семантико-морфологической категории рода и к категории одушевленности/неодушевленности существительных (конкретных значений многозначного существительного);
- описания моделей глагольного и именного управления.

Параметризация ЭТС с помощью семантических словарей ЭСС и ЭПС вводит в описание конкретных значений многозначного слова семантические контексты (синонимический и паронимический) в качестве их дифференциального признака.

При определении первоочередности ввода указанных электронных словарей в ИЛС учитывалась степень важности параметров, представленных в этих словарях, для разработки лингвистического обеспечения систем АОТ. С этой точки зрения приоритет электронного грамматического словаря, обеспечивающего лексико-грамматический и семантико-грамматический

¹ Описание понятий *электронная версия бумажного словаря, автоматизированный словарь и автоматический словарь* см. в работе Л. Н. Беляевой [3].

лексикографические эффекты, очевиден. Лексико-грамматический параметр является таковым, без которого не может обойтись ни один из модулей лингвистического процессора, поскольку информация о принадлежности слова к лексико-грамматическому классу является исходной в синтаксическом, лексическом и в семантическом анализаторах. Представленная в параметре «парадигматический класс» информация о типе словоизменения слов внутри одного грамматического класса делает этот параметр определяющим в программе лемматизации (при переходе от текстовой словоформы к канонической, фиксируемой в словаре) и в программе синтеза (при построении конкретной словоформы по заданной исходной).

В ИЛС грамматический словарь представлен в двух ипостасях – как информационно-справочная система, содержащая данные о словоизменительной подсистеме украинского языка, и как автоматический словарь, ориентированный на выполнение резидентной функции в программном обеспечении систем АОТ.

Целесообразность интегрирования в ИЛС синтаксического словаря управления обусловлена тем, что выбор в качестве единицы словника ЭСУ не лексемы, а слова в его конкретном значении делает возможным представление в эксплицитном виде существующих корреляций между конкретными значениями многозначного слова и их синтаксическими контекстами (схемами управления). Осуществление параметризации электронного толкового словаря с помощью ЭСУ превращает ЭТС в репрезентативный массив языкового материала для исследований еще одного аспекта синтаксической семантики.

Автоматические словари управления, прежде всего, ориентированы на использование их в автоматических редакторах для выявления в тексте ситуаций неправильного употребления схем управления («дякувати [кого]»: «дякувати волонтерів» вместо правильного в украинском языке «дякувати [кому]»: «дякувати волонтерам»). Интеграция ЭСУ с электронным толковым словарем позволяет вводить в правила автоматического редактора подсказки о необходимости изменения формы зависимого слова. Подключение в этом случае ЭГС, в свою очередь, позволит автоматически синтезировать по словарю управления соответствующую форму.

В ходе автоматического морфологического анализа текста обращение к ЭСУ оказывается очень эффективным на этапе снятия грамматической омонимии именных форм, находящихся в тексте в постпозиции к глаголам, зафиксированным в ЭСУ. Составление с помощью операции логического умножения валентных свойств глагольной формы, заданных в словаре управления, и морфологических характеристик именной формы, установленных в результате автоматического морфологического анализа, приводит к полному снятию омонимии или уменьшению длины цепочки омонимичного кода:

замовити – {кого, що, кому} U кулі {род., дав., місц. одн., наз.,
знах. множ., іменник жін. роду} → кулі { знах. мн., іменник жін.
роду, наз., знах. мн., іменник чол. роду}.

В системах МП с украинским компонентом (в частности, украинско-русского перевода) установление в тексте для переводимого глагола реализуемой им схемы управления может сокращать количество переводных эквивалентов (ПЭ). Например, глагол «вистрибнути» имеет два русских ПЭ – «выпрыгнуть» и «запрыгнуть». Если в тексте алгоритмически устанавливается реализация схемы управления «з/ зі/ із + род.пад.», то будет выбран первый ПЭ – «выпрыгнуть», если – схема управления «на + вин. пад.» – второй ПЭ «запрыгнуть».

Программы интеграции электронных словарей в ИЛС и организация АРМ лингвиста, работающего над составлением словаря, обеспечивают автоматическое заполнение полей ЛБД ЭСУ информацией, необходимой для:

- формирования словника словаря, основной единицей которого является слово в его конкретном значении;
- индексации единиц словника номерами значений/ оттенков значения, совпадающими с соответствующими им номерами в ЭТС;
- заполнения полей «переходность» и «вид»;
- заполнения полей «схема предложного управления», «схема беспредложного управления», экстрагированными из поля синтаксических признаков в ЭТС, представленными там относительными местоимениями «кого», «що», «кому» и т.п. или предложными сочетаниями типа «з ким», «для чого»;
- заполнения поля «сфера употребления» в виде стилистических помет, экстрагированных из соответствующего поля ЭТС;
- заполнения рабочего поля «иллюстрации»;
- формирования поля толкования значений из ЭТС. Речь идет не только о простом перенесении соответствующей информации из поля ЭТС в поле ЭСУ, но и о восстановлении в эксплицитном виде толкований, заданных в ЭТС схемами «доконаний до» (ПРОДИБАТИ – док. до **дибати**), «однократный до» (АГАКНУТИ – однокр. до **ага́кати**), «пассивный до» (АБОНУВАТИСЯ – пас. до **абонува́ти**), «безособовий» или формулами отсылок к толкованиям других слов: «те саме, що» (ПЕРЕГАТА, и, ж. Те саме, що **зага́та** 1. ЗАГАТА, и, ж. 1. Споруда для затримання руху води в річці, потоці і т. ін.), «дивитися» (ПЕРЕГАТИТИ див. **перега́чувати**).

Поле лексико-грамматических признаков заполняется путем параметризации словника с помощью электронного грамматического словаря.

Сервисные программы АРМ обеспечивают возможность составителям словаря в автоматизированном режиме осуществлять обращение к конкретным полям, редактирование информации (удаление, внесение новой, исправление).

В будущем в рамках АРМ предусматривается создание в автоматизированном режиме текстовой корпусной поддержки путем снабжения единиц словника, подчиненных одной лексеме, списком коллокаций, извлеченных из корпуса [4].

Для перехода от автоматизированного словаря управления к автоматическому, выполняющему резидентную функцию внутри программного обеспечения автоматических систем обработки языковой информации

разработаны специальные программы перекодирования содержащейся в них информации, предусмотрено внесение определенных модификаций во внутреннюю структуру словаря.

Интегрирование электронного словаря синонимов и электронного словаря паронимов¹ в ИЛС обусловлено тем, что синонимия и паронимия принадлежат к широко распространенным в языке явлениям, которые из-за своей причастности к плану содержания языковых знаков требуют особого внимания к себе в системах автоматической обработки текстовой информации, в частности это касается систем машинного перевода, автоматического редактирования, систем информационного поиска. Если синонимы играют положительную роль в системах МП (они могут быть использованы как инструмент для уточнения и сокращения количества вариантов перевода), а также в системах информационного поиска (проиндексированные синонимы обеспечивают повышение показателя полноты поиска), то паронимия негативно влияет на качество работы этих систем. Неправильное употребление компонента паронимической пары в переводимом тексте может приводить к полному искажению его смысла и поэтому крайне негативно влиять на качество перевода. При информационном поиске выбор неправильного паронима в качестве ключевого слова увеличивает показатели шума (находится лишнее) или молчания (ничего не находится) информационно-поисковой системы.

Паронимическая параметризация ИЛС может стать полезной при составлении семантических правил модуля автоматического редактирования исходных текстов в системах машинного перевода, а также в информационно-поисковых системах. Стало очевидным, что степень эффективности этих систем находится в прямой зависимости от их способности распознавать и исправлять в обрабатываемом тексте ситуации с неправильным употреблением одного из компонентов паронимической пары.

При параметризации электронного толкового словаря с помощью ЭСС индексируются конкретные значения слов соответствующими этим значениям синонимическими группами. Внутри группы синонимы проиндексированы номерами значений, взятыми из ЭТС. При параметризации текстовой информации индексируемым словоформам ставятся в соответствие все синонимические группы, сформированные по конкретным значениям лексемы данной словоформы с указанием соответствующего значения. Аналогичным образом происходит и параметризация ЭТС и по электронному словарю паронимов.

В рамках корпусной лингвистики электронный грамматический словарь и автоматический словарь управления выполняют функцию инструмента соответственно морфологической и синтаксической разметок Украинского национального лингвистического корпуса. Электронные словари синонимов и паронимов предлагается использовать в качестве основного инструмента семантической разметки.

¹ Принятое определение понятий синонимов и паронимов, внешние и внутренние структуры словарей ЭСС и ЭСП, принципы формирования описаны в статьях [5–9].

ЛИТЕРАТУРА

1. *Широков, В. А.* Інформаційна теорія лексикографічних систем / В. А. Широков. – Київ : Довіра, 1998. – 331 с.
2. *Широков, В. А.* Лінгвістичні та технологічні основи тлумачної лексикографії / В. А. Широков, В. М. Білоноженко, О. В. Бугаков [та ін.]. – Київ : Довіра, 2010. – 293 с.
3. *Беляева, Л. Н.* Формирование переводного словаря средствами информационных технологий / Л. Н. Беляева // Прикладна лінгвістика та лінгвістичні технології : MegaLing-2012 : зб. наук. пр.; НАН України, Укр. мовно-інформ. фонд; редкол. : В. А. Широков [та ін.]. – Київ : УМІФ, 2013. – С. 22–36.
4. *Широков, В. А.* Корпусна лінгвістика / В. А. Широков, О. В. Бугаков, Т. О. Грязнухіна [та ін.]; під ред. В. А. Широкова. – Київ : Довіра, 2005. – 471 с.
5. *Грязнухіна, Т. А.* Индексирование паронимов в лексикографических системах / Т. А. Грязнухіна, Т. П. Любченко, В. А. Широков // Прикладна лінгвістика та лінгвістичні технології : MegaLing'2009 : зб. наук. пр.; за ред. В. А. Широкова. – Київ : Довіра, 2009. – С. 171–179.
6. *Грязнухіна, Т. А.* Технология и программно-инструментальные средства формирования электронного словаря паронимов украинского языка / Т. А. Грязнухіна, Т. П. Любченко, В. А. Широков, К. Н. Якименко // Прикладна лінгвістика та лінгвістичні технології : MegaLing'2010 : зб. наук. пр.; за ред. В. А. Широкова. – Київ : Довіра, 2010. – С. 45–54.
7. *Грязнухіна, Т. О.* Словник української мови у 20-ти томах як інструмент для створення Електронного словника паронімів / Т. О. Грязнухіна, Т. П. Любченко // Лексикографічний бюлетень : зб. наук. пр.; відпов. ред. к. філол. н. І. С. Гнатюк. – Київ : Видавничий дім Дмитра Бураго, 2011. – Вип. 20. – С. 28–34.
8. *Устимець, О. В.* Формування електронного словника синонімів української мови / О. В. Устимець // Учён. зап. Таврического нац. ун-та им. В. И. Вернадского. Серия «Филология» : в 59 т. – Симферополь, 2007. – № 4. – Т. 20. – С. 57 – 61.
9. *Грязнухіна, Т.* Операційне визначення критерію семантичної подібності синонімів / Т. Грязнухіна, О. Устимець, В. Широков // Прикладна лінгвістика та лінгвістичні технології : MegaLing-2008 : зб. наук. пр. / НАН України, Укр. мовно-інформ. фонд; редкол. : Ю. Д. Апресян [та ін.]. – Київ : Довіра, 2009. – С. 42–57.

А. В. Зубов

ОБЩИЕ ПРИНЦИПЫ СОЗДАНИЯ ТЕРМИНОЛОГИЧЕСКИХ СЛОВАРЕЙ ДЛЯ АВТОМАТИЧЕСКОГО ПЕРЕВОДА СПЕЦИАЛИЗИРОВАННЫХ ТЕКСТОВ

Одним из наиболее важных типологических параметров любого создаваемого словаря является назначение этого словаря. С этой точки зрения выделяются 4 основных типа терминологических словарей: переводные, справочные, учебные и информационные. В процессе создания терминологических словарей выделяют четыре основных этапа [1, с. 182].

На первом этапе происходит проектирование словаря, т.е. определяется его тип, читательский адрес и основные характеристики. При этом используется систематизированный перечень возможных параметров в качестве

своеобразного вопросника, на основе которого принимаются предварительные решения по всем элементам методической установки и композиции словаря, определяется состав отражаемой в нем информации о специальной лексике. В настоящее время этому этапу не всегда уделяется должное внимание, в результате чего в ходе разработки словаря приходится вносить непредвиденные изменения, что отражается как на трудоемкости его изготовления, так и на его качестве. Результатом этого этапа является проект словаря и методика его создания.

На втором этапе производится отбор лексического материала и составление словника будущего словаря.

На третьем этапе (основном) проводится непосредственный анализ и описание специальной лексики, включенной в словник. В итоге формируются основной и вспомогательный указатели.

На четвертом этапе происходит подготовка словаря к изданию (редактирование, уточнение и проверка взаимности ссылок и т.д.).

Рассматривая более детально первый этап задачи создания переводного терминологического словаря, необходимо отметить, что в январе 2016 года в рамках подпрограммы «Белорусский язык и литература» государственной программы научных исследований «Экономика и гуманитарное развитие белорусского общества» НАН Республики Беларусь на 2016–2020 годы кафедре информатики и прикладной лингвистики МГЛУ было поручено задание «Создание автоматического англо-белорусского терминологического словаря по информационным технологиям и разработка методики его использования для перевода текстов».

Это связано с тем, что в наши дни через Интернет идет активный обмен сведениями о новейших компьютерных программах автоматической обработки текстов, новейших двуязычных и многоязычных словарях, самые последние данные о возможностях перевода текстов с одного языка на другой.

Но такие данные пока могут активно использовать пользователи, хорошо владеющие английским и русским языками. Очень мало такой информации на белорусском языке.

Из более 300 различных терминологических словарей, созданных в Республике Беларусь [2] для автоматической обработки текстов на белорусском языке, используется менее 10 словарей.

Поэтому конечными целями полученного кафедрой информатики и прикладной лингвистики задания является:

1. Создание автоматического англо-белорусского терминологического словаря по современным информационным технологиям.
2. Разработка способов использования этого словаря для перевода на белорусский язык английских слов и словосочетаний по информационным технологиям.
3. Разработка способов использования создаваемого словаря для автоматического перевода английских текстов по информационным технологиям на белорусский язык.

Рассматривая второй этап создания терминологического англо-белорусского словаря по современным информационным технологиям, необходимо отметить, что в Республике Беларусь это научное направление характерно следующими особенностями:

1. Неоднородностью состава предметной области «Информационные технологии».

2. Неравномерностью темпов развития отдельных направлений информационных технологий.

3. Отсутствием необходимой информации (текстов, словарей) на белорусском языке.

Эти факты говорят о том, что в создаваемом терминологическом словаре кроме терминов должны присутствовать английские и белорусские общеупотребительные стилистически нейтральные слова, общенаучно-техническая лексика, служебные и строевые слова [3].

Все вышесказанное свидетельствует о том, что проблемы отбора лексики в создаваемые словари не решены окончательно. Поэтому при создании запланированного англо-белорусского автоматического словаря будет использована лексика, уже примененная в существующих английских, англо-русских, русских, русско-белорусских терминологических словарях по информатике, вычислительной технике и информационным технологиям.

Третий этап создания автоматического англо-белорусского терминологического словаря будет выполнен в процессе детального анализа лексики, включенной в упомянутые выше словари.

ЛИТЕРАТУРА

1. Хайчук, Роман. Основы создания терминологических словарей / Роман Хайчук, Александр Зубов. – Białystok : Trans Human, 2004.
2. Беларуская тэрміналогія. Тэрміналагічныя слоўнікі (асобныя выданні). 1918–1998. Бібліяграфічны даведнік. – Мінск : Беларускі кнігазбор, 2000. – Вып. 1.
3. Бакая, Р. Критерии отбора лексических единиц и методика обработки русско-английского словаря–минимума для чтения литературы на русском языке / Р. Бакая. – АКД. – М., 1973.

О. М. Карпова

ТЕРМИНОЛОГИЧЕСКИЕ СЛОВАРИ ШЕКСПИРА НОВОГО ПОКОЛЕНИЯ

У Шекспира есть «темные» места, над толкованием которых на протяжении многих столетий работали и потратили немало сил отечественные и зарубежные ученые. Это язык цветов, игра слов, клятвенные речения, эротическая лексика, имена собственные, уничижительная лексика и многие другие группы слов и словосочетаний, лексикографическому описанию которых посвящено более ста лингвистических и энциклопедических словарей языка писателя, разъясняющих тайны шекспировского словоупотребления [1].

В словарном богатстве Шекспира значительную часть занимают термины различных предметных областей, где писатель является общепризнанным профессионалом [2, с. 37–47]. Вот почему современные лексикографы обратились к созданию специальных терминологических справочников, чтобы избежать трудностей и ошибок (например, в выборе правильного реквизита: оружия, одежды, блюд и т.п. артефактов) при постановке шекспировских пьес в различных театрах мира [3]. Более того, появление многочисленных справочников такого рода объясняется и значительной хронологической удаленностью от нас произведений Шекспира [4, р. 158–178].

Выбранные лексикографами для описания предметные области, связаны с:

- **военным делом:** «Shakespeare's Military Language. A Dictionary» [5];
 - **религией:** «Shakespeare's Religious Language. A Dictionary» [6];
 - **музыкой:** «Music in Shakespeare: A Dictionary» [7];
 - **юриспруденцией:** «Shakespeare's Legal Language: A Dictionary» [8];
 - **политикой и обществом:** «Class and Society in Shakespeare» [9], «Shakespeare's Political and Economic Language» [10];
 - **медициной:** «A Doctor's Shakespeare: A Collection of Medical Quotations from Shakespeare» [11], «Shakespeare's Medical Language: A Dictionary» [12];
 - **ботаникой** «Shakespeare's Plants and Gardens: A Dictionary» [13];
- и другими предметными областями [14; 15; 16].

Особого внимания, на наш взгляд, заслуживает опубликованный в 2004 г. справочник, впервые составленный театральным режиссером Х. Ричмондом «Shakespeare's Theatre: A Dictionary of His Stage Context» [17]. Этот словарь открыл целую серию терминологических шекспировских словарей, известных в лексикографии как *Shakespeare Continuum Dictionaries*. Следует отметить, что его автор сам выступил экспертом при составлении терминологического словаря, поскольку испытывал определенные трудности при постановке пьес Шекспира в Австралии.

Словник включает театральную лексику, собранную из всех пьес драматурга. Входные единицы регистрируются в алфавитном порядке, а для удобства пользователя они дублируются в специальном индексе в конце издания, что значительно облегчает задачу составителя и создает удобный указатель для всех ключевых слов. Среди них: *action, art, battles, ballads, boy actors, blank verse, chorus, comedy, costumes, dialogues, duels, entertain, festivals, galleries, Jacobean drama* и т.п. слова и словосочетания, которые вполне можно отнести к театральным терминам. Однако часть словника, как и в большинстве обычных терминологических словарей, составляют общеупотребительные слова [18, р. 393–407], например: *disease, discover, element, English, interest, intention, influence, horses, kiss* и т.п.

Словарная статья не содержит каких-либо помет, принятых в лингвистических словарях языка Шекспира (грамматических, хронологических, этимологических и т.п.), однако имеет подробное толкование каждой входной единицы. Кроме того, в микроструктуре регистрируется иллюстрация употребления слова в шекспировском тексте с точным указанием ее адреса, а именно: акт, сцена и страница для быстрого нахождения необходимой справки:

accent The distinctive intonation of voice reflecting a character's region of origin, social class or education is something that Shakespeare characteristically uses to help actors to distinguish their roles. Mercutio is very sarcastic about the pretentiousness of 'these new tuners of accent' (*Romeo*, 3.4.29), while Kent, acting as a servant, talks of his dialect as the speech of 'a plain knave' with 'a plain accent' (*Lear*, 2.2.93–113). Later, Edgar adopts a broad regional accent as a disguise from his blind father (4.6.7–8) and from Oswald (4.6.231–43). Shakespeare elaborately exploits national accents throughout *Henry V*, and is aware of the divergent intonations of foreign languages alluded to in *Caesar* as 'accents yet unknown' (3.1.113).

К сожалению, цитаты присутствуют не во всех статьях, а только там, где (по мнению автора словаря), термины в шекспировских пьесах несут важную смысловую нагрузку, а иногда имеют ярко выраженное эстетическое значение:

discharge This verb sometimes has the sense in Shakespeare of 'enact on the stage', as when the workmen say of the missing Bottom in the role of Pyramus: 'You have not a man in all Athens able to discharge Pyramus but he' (*Dream*, 4.2.7–8). The implication is of meeting a challenge or completing an obligation: Ferdinand says of his log carrying: 'The sun will set before I shall discharge What I must strive do' (*Tempest*, 3.1.22–3).

Прекрасным дополнением к словарю театральных терминов стал упомянутый выше справочник «Music in Shakespeare», включающий более трехсот входных единиц [7]. Его значение для постановщиков шекспировских пьес трудно переоценить, поскольку словник фиксирует как редко встречающиеся у Шекспира термины (*degree*, *ravan* и т.п.), так и термины с высокой частотой употребления, например, *music* (143 раза зарегистрирован в произведениях писателя).

Удачным представляется принцип построения словарных статей. Они организованы таким образом, что в трех составляющих микроструктуры (А, В, С) отражаются: общезыковое значение термина, его индивидуально-авторское понимание и ссылки на источники, например:

Bagpipe (A) A wind instrument with its distinctive air bag, originating in Ancient Greece and Arabia....

(B) Throughout the Middle Ages and Renaissance, the bagpipe was generally an outdoor folk instrument, used to accompany dancing and eating at social occasions. This allusion is clear in the servant's comment preceding the Autolycus's entry:

O master! If you did but hear the pedlar at the door, you would never dance again after a tabor and pipe; no, the bagpipe could not move you

(WT 4, 4, 181–31)

The wailing sound of the bagpipe and its continuous drone has led writers to allude to its melancholic character. Such melancholy appears in *The Merchant of Venice*. "And laugh like parrots at a bagpiper" (1.1.53)....

Falstaff is as melancholy as 'the drone of a Lincolnshire bagpipe' (1H4 1.2.76)...

Особую ценность в микроструктуре словаря, несомненно, имеют цитаты из пьес Шекспира, позволяющие лексикографу адекватно интерпретировать значение каждого термина в его авторском понимании.

Начиная с 2012 г., серия терминологических словарей *Shakespeare Continuum Dictionaries* меняет свое название на *Arden Shakespeare Dictionaries* и получает в качестве главного редактора известного шекспироведа С. Кларк, автора многих специальных словарей языка Шекспира. С этого времени все словари серии направлены на сохранение индивидуально-авторского значения терминов, относящихся к различным предметным областям. Более того, составители словарей новой серии начинают постепенно вводить в их микроструктуру ссылки на полезные страницы в Интернет. *Этот прием существенно расширяет возможности справочника и дает пользователю уникальную возможность прослушать музыкальное произведение, просмотреть отрывок из постановки пьесы или кинофильма по произведениям Шекспира, делая словарь не только полезным, но и весьма привлекательным для современного пользователя. Иными словами, XXI век дает новый импульс к развитию и применению информационных технологий не только в лингвистике вообще, и в писательской лексикографии, в частности* [19].

В этот период наряду с печатными терминологическими словарями языка Шекспира появляются интересные и весьма оригинальные справочные ресурсы в Интернет. Они имеют черты как ранее опубликованных лексикографических трудов, так и новые характеристики, присущие лишь справочникам нового поколения. В силу своей специфики, новые информационные ресурсы позволяют читателю быстро найти необходимую справку и выполнить многие действия, которыми обладают только словари в Интернет.

В этой связи внимание шекспироведов и, в первую очередь, постановщиков шекспировских пьес, привлекает сайт, созданный А. Мэбилард *Shakespeare's Drinking* [20].

Известно, что у Шекспира встречается большое количество слов, связанных с напитками, которые нередко играли особенную роль в его драмах, комедиях и трагедиях. А. Мэбилард не случайно начинает свой справочник с цитирования отрывка из комедии «Двенадцатая ночь»:

OLIVIA What's a drunken man like, fool? **CLOWN** Like a drowned man, a fool, and a madman: one draught above heat makes him a fool, the second mads him, and a third drowns him. *Twelfth Night* (1.5.127–30).

Справочный ресурс включает две группы входных единиц:

1) названия напитков (их состав, предназначение), питейных заведений и т.п. информацию о тех ключевых словах, которые присущи Елизаветинской эпохе и характеризуют их с точки зрения времени, когда жил и творил Шекспир, а также

2) цитаты с этими словами из произведений Шекспира с указанием имени героя произведения и адреса цитаты.

Первая группа слов составляет большинство в корпусе словаря и называется *What Did Shakespeare Drink?*, например:

Ale and Beer

Ale (beer made with top fermenting yeast) was the drink of choice in Shakespeare's day. Everyone from the poorest farmer to the Queen herself drank the brew made from malt, and a mini brewery was an essential part of every

household. Shakespeare's own father was an official ale taster in Stratford – an important and respected job which involved monitoring the ingredients used by professional brewers and ensuring they sold their ale at Crown regulated prices. Beer, however, eventually became more popular than ale. "In 1574, there were still 58 ale brewers to 33 beer brewers in the City [London], but beer gradually replaced ale as the national drink over the course of the century" (Picard, 187).

В конце словарной статьи нередко делается ссылка на фамилию и работу ученого, исследовавшего значение того или иного слова.

Микроструктура справочника содержит подробную информацию о происхождении слова, а также ссылки на произведения писателя (акт, сцену, страницу), в которых они употребляются (см., например):

Malmsey

A rich and sweet wine brought to England from Greece in the 16th century, Malmsey is now produced on the island of Madeira. Shakespeare writes about Malmsey in *Love's Labour's Lost* (5.2.240) and *2 Henry IV* (2.1.36), but the most famous reference to Malmsey in all of literature can be found in *Richard III*, when Richard orders the execution of his brother, the Duke of Clarence. Richard's hired assassins decide to drown Clarence in a large cask (butt) of the brew. When they arrive at the Tower of London to carry out the task, the unsuspecting Clarence asks for a cup of wine. The Second Murderer offers this ghastly retort: "You shall have wine enough, my lord, anon" (1.4.153).

Некоторые статьи дают подробную толково-энциклопедическую справку, как, например, в статье на слово **wine**:

Wine Wine was available in Shakespeare's England, but it was very expensive (about twelve times more costly than ale (Singman, 137)) and so only the upper classes could enjoy it regularly. English grapes were not adequate for winemaking so they imported their wines from France, Spain, and Greece. Sack, a sweet wine fortified with brandy (known today as sherry), was most popular with the Elizabethans.

Shakespeare's Prince Hal and the lads at the Boar's-head Tavern in Eastcheap were impassioned by the scrumptious drink. The following is a passage from *2 Henry IV* in which Falstaff praises the benefits of sack to one's constitution:

A good sherris sack hath a two-fold operation in it. It ascends me into the brain; dries me there all the foolish and dull and curdy vapours which environ it; makes it apprehensive, quick, forgetive, full of nimble fiery and delectable shapes, which, delivered o'er to the voice, the tongue, which is the birth, becomes excellent wit. The second property of your excellent sherris is, the warming of the blood; which, before cold and settled, left the liver white and pale, which is the badge of pusillanimity and cowardice; but the sherris warms it and makes it course from the inwards to the parts extreme: it illumineth the face, which as a beacon gives warning to all the rest of this little kingdom, man, to arm; and then the vital commoners and inland petty spirits muster me all to their captain, the heart, who, great and puffed up with this retinue, doth any deed of courage; and this valour comes of sherris. So that skill in the weapon is nothing without sack, for that sets it a-work; and learning a mere hoard of gold kept by a devil, till sack commences it

and sets it in act and use. Hereof comes it that Prince Harry is valiant; for the cold blood he did naturally inherit of his father, he hath, like lean, sterile and bare land, manured, husbanded and tilled with excellent endeavour of drinking good and good store of fertile sherris, that he is become very hot and valiant.. . (4.3.100)...

По отзывам многих пользователей, настоящий справочный ресурс является эффективным инструментом для театральных и кино-режиссеров при постановке пьес великого драматурга.

Таким образом, терминологические словари Шекспира нового поколения, охватывающие значительное количество предметных областей, созданы по запросам и требованиям определенной группы пользователей – театральных режиссеров. Эти справочники являются качественно новым типом терминологических словарей, способствующих раскрытию эстетического значения писательского словоупотребления, которые существенно обогащают современную лексикографическую картину мира.

ЛИТЕРАТУРА

1. *Karpova, O. English Author Dictionaries (the XVIth – the XXIst cc.)* / O. Karpova. – Newcastle upon Tyne : Cambridge Scholars Publishing, 2011. – 270 p.
2. *Карпова, О. М. Шекспир – профессиональная языковая личность?* / О. М. Карпова // Язык. Общество. Сознание. – Казань : Отечество, 2013. – С. 37–47.
3. *Карпова, О. М. Очерки по шекспировской лексикографии* / О. М. Карпова. – Saarbrücken : Lambert Academic Publishing, 2014. – 236 с.
4. *Karpova, O. Shakespeare Dictionaries of a New Generation* / O. Karpova // *New Trends in Lexicography : Ways of Registrating and Describing Lexis*; ed. by O. Karpova, F. Kartashkova. – Newcastle upon Tyne : Cambridge Scholars Publishing, 2010. – P. 158–178.
5. *Edelman, Ch. Shakespeare's Military Language. A Dictionary* / Ch. Edelman. – L., N.Y. : Athlone Press, 2000. – 448 p.
6. *Hassel, R. Series Editor: Sandra Clark Shakespeare's Religious Language. A Dictionary* / R. Hassel. – Bloomsbury Arden Shakespeare. – Bloomsbury Arden Shakespeare, 2015. – 480 p.
7. *Wilson, Ch. Music in Shakespare: A Dictionary* / Ch. Wilson, M. Calore. – L. : Athlone press, 2005. – 508 p.
8. *Sokol, B., Sokol, M. Shakespeare's Legal Language: A Dictionary* / B. Sokol, M. Sokol. – L. : Continuum, 2005. – 510 p.
9. *Innes, P. Class and Society in Shakespeare* / P. Innes. – L., N.Y. : Continuum, 2007. – 596 p.
10. *Thomas, V. Shakespeare's Political and Economic Language* / V. Thomas. – Bloomsbury Arden Shakespeare, 2015. – 416 p.
11. *Foster, J. A Doctor's Shakespeare: A Collection of Medical Quotations from Shakespeare* / J. Foster. – N.Y. : Sparks, 2011. – 489 p.
12. *Iyengar, S. Shakespeare's Medical Language: A Dictionary* / S. Iyengar. – Bloomsbury Arden Shakespeare, 2014. – 432 p.
13. *Thomas, V. Faircloth N. Shakespeare's Plants and Gardens: A Dictionary. Arden Shakespeare Dictionaries* / V. Thomas. – Bloomsbury Arden Shakespeare, 2016. – 432 p.
14. *Findlay, A. Women in Shakespeare: A Dictionary (Continuum Shakespeare Dictionary Series)* / A. Findlay. – L., N.Y. : Continuum, 2010. – 546 p.

15. *Fitzpatrick, J.* Shakespeare and the Language of Food: A Dictionary (Continuum Shakespeare Dictionaries) / J. Fitzpatrick. – L. : Continuum, 2010. – 480 p.
16. *Gibson, M.* Shakespeare's Demonology. A Dictionary / M. Gibson, J. A. Esra. – Bloomsbury Arden Shakespeare, 2014. – 248 p.
17. *Richmond, H.* Shakespeare's Theatre. A Dictionary of his Stage Context / H. Richmond. – L., N.Y. : Continuum, 2003. – 584 p.
18. *Beker, H.* Scientific and Technical Dictionaries; Coverage of Scientific and Technical Terms / H. Beker // The Oxford Handbook of Lexicography; ed. by Ph. Durkin. – Oxford : Oxford University Press, 2016. – P. 393–407.
19. *Зубов, А. В.* Информационные технологии в лингвистике / А. В. Зубов. // Учебник для студентов учреждений высшего профессионального образования. – Минск : Изд-во «Академия» (Academia), 2012. – 208 с.
20. *Mabillard, Amanda.* Shakespeare's Drinking. Shakespeare Online [Electronic resource] / Amanda Mabillard. – Mode of access : <http://www.shakespeare-online.com/faq/shakespearedrinking>. – Data of access : 20.08.2000.

И. А. Меньшенина

О ПОДХОДЕ К ОТБОРУ ЛЕКСИЧЕСКОГО МИНИМУМА ДЛЯ ОБУЧЕНИЯ СПЕЦИАЛЬНОСТИ

Под лексическим минимумом обычно понимают некоторое множество слов, подлежащих обязательному усвоению на определенном этапе изучения неродного языка. Обычно выделяют два этапа развития теории лексических минимумов, в каждом из которых преобладал один из следующих подходов к изучению лексики: элементарный или структурный [1, с. 46–47]. В основе первоначального подхода к отбору лексического минимума лежали отдельные свойства лексической системы языка, наиболее релевантные с точки зрения лингвистов, либо отдельные свойства слова как элемента этой системы. В рамках такого элементарного подхода отдельные лингвисты выбирали одно, наиболее значимое, наиболее характерное на их взгляд свойство лексической системы или лексической группы этой системы, которое служило основой для отбора лексического минимума. Среди таких свойств выделяли: 1) принадлежность слова к основному словарному фонду языка, выделяемого с помощью этимологического, или исторического критерия (К. Магер); 2) способность слова объединяться в тематические группы (тематический критерий) (Я. А. Коменский); 3) частотность слова, раскрываемая в частотных словарях различных типов (Э. Торндайк, Э. Хорн и др.); 4) способность слова замещать другие слова и выступать средством семантизации слов более низкого иерархического ряда (описательная способность слова) (М. Уэст, Ч. Огден). Со временем стало очевидно, что отбор лексики на основе какого-либо одного критерия не позволяет создать качественный учебный словарь.

В начале 30х гг. XX века в основу отбора лексического минимума были положены критерии структурного характера. При данном подходе лингвисты использовали разнообразные структурные и статистические свойства слова как единицы лексической системы языка. Поиск свойств, наиболее полно

раскрывающих пользу или ценность слова для обучения, велся в разных направлениях: рассматривались текстообразующие свойства лексики, т.е. частотность слова, необходимость слова для целей обучения, употребительность, активность слова в речи. Таким образом, были выделены критерии (принципы) для отбора лексики.

В это же время лексическую систему стали рассматривать не только как лингвистическое, но и как дидактическое и психологическое явление. Поэтому в оценке значимости слова ученые вышли за пределы лингвистики, привлекая такие науки как психология, педагогика и др.

Основоположник структурного подхода Г. Палмер дополнил принцип частотности еще 12-ю критериями отбора лексики, такими как принципы эргонической комбинированности (т.е. языковой сочетаемости), стилистической нейтральности, конкретности, общей целесообразности, пропорции и др. С течением времени количество критериев отбора лексического минимума значительно возросло и продолжает расти.

Сегодняшнее понимание сущности критериев сводится к следующему. Все критерии делят на три большие группы: а) лингвистические критерии, отражающие релевантные свойства лексики того или иного иностранного языка (критерий частотности, употребительности, сочетаемости, многозначности, словообразовательной ценности, семантической ценности и др.); б) лингводидактические критерии, обладающие языковой природой, но направленные на обучение (критерии целевого назначения, тематический критерий, координационные критерии, сопоставительные критерии, основанные на учете специфики родного языка учащихся); в) дидактические критерии, отображающие наиболее существенные закономерности учебного процесса и обусловленные потребностями обучения (критерии практической необходимости слова, соответствия учебным потребностям, учета возрастных особенностей учащихся и т.д.).

Свойства лексической системы языка и дидактические и психологические закономерности обучения крайне многообразны, количество критериев можно увеличивать до бесконечности.

И все же для того чтобы служить целям отбора учебной лексики, критерии должны обладать рядом особенностей: 1) способностью транспонироваться в конкретный лексический материал (селективная функция); 2) способностью разграничивать одну группу лексики от другой, определять ее состав (сигнификативная функция); 3) способностью ограничивать тот или иной объем лексики (рестриктивная функция); 4) способностью качественной и количественной характеристики состава лексики (кваликативная функция); 5) способностью служить целям отбора только одной определенной группы слов, не включать в себя и не соподчинять в себе другие критерии (целостность и неразложимость).

Рассматривая проблему создания лексических минимумов по определенным конкретным специальностям, выделяют три следующие важные методические задачи [2, с. 65]:

1. Установление объема учебного словаря-минимума.
2. Определение принципов отбора лексического материала.

3. Адекватное отражение в создаваемом словаре-минимуме качественных и количественных характеристик единиц этого словаря-минимума.

При решении данных задач наиболее сложным вопросом является определение принципов отбора лексического материала. Учебные словари составляются на основе статистического анализа специально отобранных текстов определенного подъязыка научно-технической литературы с опорой на имеющиеся частотные словари. В основу отбора текстового материала чаще всего берут тематический принцип. Таким образом, в сферу нашего исследования вошли тексты наиболее актуальные с точки зрения данной науки, а также тексты, относящиеся к ее основным направлениям.

Говоря об объеме словаря-минимума, следует отметить ранее упомянутого английского лингвиста и педагога Палмера, который отобрал 3 тысячи слов, разделил их на 3 части по тысяче слов и установил, что этого количества слов достаточно для понимания 95 % текста. Его ученик, голландец Богус, установил список из 3 тысяч слов, покрывающих, по его мнению, 97,5 % текста. На сегодняшний день установлено, что 1500–2000 слов, наиболее часто встречающихся в оригинальных источниках определенного подъязыка, являются оптимальным объемом учебного словаря-минимума, учитывая ограниченное количество учебных часов, отводимых на изучение иностранного языка в неязыковом вузе, покрывают 80–85 % любого текста в данной области научно-технической литературы. При увеличении словаря до 4000 слов, обладающих наиболее высокой частотой, наблюдается увеличение охвата текста всего на 3–4 %. И принимая во внимание такие немаловажные факторы, как возможности учащихся в усвоении языкового материала, ограниченное количество учебного времени, отводимого в неязыковых вузах на изучение иностранного языка, а также цели его изучения в таких вузах, можно сделать вывод, что оптимальный объем учебного лексического минимума по определенной специальности составляет 2000 слов.

Третья из упомянутых выше задач связана с количеством разделов той научной области, по которой составляется словарь-минимум. В таком словаре должны быть использованы частоты употребления конкретных слов словаря-минимума в различных разделах научной области.

Эти принципы и были использованы в нашем исследовании для построения лексического минимума для освоения англоязычных учебных медицинских текстов.

ЛИТЕРАТУРА

1. Томилов, В. А. Общие проблемы теории отбора лексических минимумов по иностранным языкам / В. А. Томилов // Иностранные языки в школе. – 1983. – № 5. – С. 45–50.
2. Вороненкова, Л. А. О методике составления учебного словаря-минимума / Л. А. Вороненкова // Квантитативные методы отбора учебного материала по иностранному языку для неязыкового вуза : науч. тр.; Яросл. политехн. ин-т : УРГУ. – Свердловск, 1986. – С. 65–66.

ЭТИМОЛОГИЧЕСКИЙ СЛОВАРЬ В ЦИФРОВОЙ СРЕДЕ

Как и все фундаментальные лексикографические работы, этимологические словари постепенно переводятся в цифровой формат. В интернете наиболее доступен «Этимологический словарь русского языка» М. Фасмера, первая его цифровая версия была издана на CD-ROM еще в 2004 году. Однако все цифровые репрезентации подобных лексикографических систем являются не более чем разновидностями машиночитаемых текстов, отсутствуют возможности масштабирования и динамического индексирования.

В данной работе рассматриваются метод и технология представления в цифровой среде Этимологического словаря украинского языка (ЭСУЯ, в печатной версии – 6 томов), которые предлагаются как универсальные для любой этимологической лексикографической системы [1]. Разработана формальная модель словарной статьи, в которой структурные элементы разделяются на два класса: лингвистические (описательные) и структурообразующие; последние вычленяются по формальной процедуре из лингвистических. На основе этой модели построена схема компьютерной базы данных (в рассматриваемой реализации – реляционной) и технология парсинга текста словаря, что позволило в автоматическом режиме конвертировать текст. Для поддержки цифровой версии ЭСУЯ разработан инструментальный комплекс, в который входит подсистема индексирования словаря по любому сформированному пользователем языковому регистру.

При выполнении работы по созданию цифровой версии ЭСУЯ использовались методы, которые уже были успешно опробованы для решения подобных задач, в частности, для создания компьютерной лексикографической базы данных нового Словаря украинского языка [2; 3].

Мы рассматриваем словарь как информационную систему особого типа – лексикографическую. Согласно теории лексикографических систем это абстрактный языково-информационный объект, ориентированный на реализацию комплексного информационного описания лексико-грамматических структур определенного языка или совокупности языков [2].

Архитектура системы отвечает стандартной трехуровневой архитектуре информационных систем ANSI/SPARK, согласно которой в информационной системе выделяются концептуальный, внутренний и внешний уровни данных [4].

В качестве концептуальной модели мы используем лексикографическую модель данных [3]. Ниже мы приводим ее в несколько упрощенном виде:

$$\{I_0(D), V(I_0(D)), \beta, \delta[\beta], Red[V(I^Q(D))]\},$$

где D – объект моделирования – Этимологический словарь украинского языка; $I_0(D) = \{x_i\}$ множество реестровых единиц словаря, в теории лексикографических систем его принято называть множеством *элементарных информационных единиц*; $V(I_0(D))$ – множество описаний (интерпретаций) элементарных информационных единиц, то есть текстов словарных статей:

$V(I_0(D)) = \{V(x_i)\}$ – словарная статья с заголовковым словом (реестровой единицей) x_i ; β – множество структурных элементов, которые были абстрагированы в результате анализа текста словаря; $\delta[\beta]$ – структура, которая порождается на β оператором δ ; ограничения $\delta[\beta]$ на $V(x)$ порождает микроструктуру словарной статьи $\delta(x)$; $Red[V(I_0(D))]$ – механизм *рекурсивной редукции лексикографической системы*. Он дает возможность последовательно выявлять все более тонкие детали структуры лексикографической системы, в частности – осуществлять распределение структурных элементов словарной статьи на реестровую и интерпретационную части.

Концептуальная модель словаря строится на основе анализа полиграфической версии ЭСУЯ, то есть анализируется типографское оформление, организация и структура печатных текстов словарных статей, которые интерпретируются как идентификаторы соответствующих элементов лексикографических структур β и $\delta(x)$.

В качестве базового структурного элемента словарной статьи определен *этимологический класс* (обозначается *ECL*), представляющий собой блок линейного текста, в котором описываются определённые генетические связи заголовочного слова. Для ЭСУЯ этимологические классы подразделяются следующим образом: *класс реестрового слова (HEAD)*, *класс дериватов (DERIVAT)*, *класс славянских соответствий (SLAVIA)*, *языковый класс (LANG)*, *библиографический класс (BIBL)* и *классы ссылок (REF и COMP в зависимости от типа ссылки)*.

Проиллюстрируем сказанное на примере двух небольших, но достаточно репрезентативных с точки зрения структуры словарных статей. Тексты приводим в форме, максимально приближенной к печатной версии.

*Пример 1 (словарная статья с заголовковым словом **абетка**):*

абетка, [абетло] Пі, абетний (заст.) «элементарний»;— власне українська назва азбуки, утворена за вимовою перших двох букв алфавіту (*а, бе*), очевидно, під впливом назв *азбука, альфабет* і п. *abecadło* «тс.» (від вимови перших трьох букв *а, be, се*).— Sadn. — Aitz. VWb. I 42.— Пор. **азбука, алфавіт**.

HEAD ≡ <**абетка**>

DER ≡ <[абетло] Пі, абетний (заст.) «элементарний»>

LANG ≡ <власне українська назва азбуки, утворена за вимовою перших двох букв алфавіту (*а, бе*), очевидно, під впливом назв *азбука, альфабет* і п. *abecadło* «тс.» (від вимови перших трьох букв *а, be, се*)>

BIBL ≡ <Sadn. — Aitz. VWb. I 42>

LINK ≡ <Пор. **азбука, алфавіт**>

*Пример 2 (словарная статья с заголовковым словом **абзац**):*

абзац;— р. бр. *абзац*, болг. *абзац*, схв. *абзац*;— запозичення з німецької мови; нім. *Absatz* «перерва, пауза, уступ, абзац» є похідним від дієслова *absetzen* «відсувати, відставляти», утвореного з префікса *ab-* «від-, з-», спорідненого з гот. *af* «від», лат. *ab* «тс.», і дієслова *setzen* «садити», пов'язаного з днп. *sezzen*, дангл. *settan*, англ. *set* і спорідненого з псл. *saditi*, укр. *садити*.— СІС 7; Фасмер I 56; Paul DWb 8, 10; Kluge — Mitzka 705.— Див. ще **абажур, садити**.— Пор. **обцас**.

HEAD ≡ <**абзац**>

SLAVIA ≡ <р. бр. *абзац*, болг. *абзац*, схв. *абзац*>

LANG $\equiv$ <запозичення з німецької мови; нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sezzen, дангл. settan, англ. set і спорідненого з псл. saditi, укр. *cadumu*>

BIBL $\equiv$ <CIC 7; Фасмер I 56; Paul DWb 8, 10; Kluge — Mitzka 705>

*LINK*₁ $\equiv$ <Див. ще **абажур, садити**>

*LINK*₂ $\equiv$ <Пор. **обцас**>

В тексті кожного етимологічного класу встановлюються зв'язи реєстрового слова з визначеними словами інших мов. Всі ці слова, включаючи реєстрові, ми будемо називати *етимонами*. При аналізі текстів етимологічних класів було виявлено вісім параметрів, за допомогою яких описуються етимони. Два параметри є обов'язковими: це *P_L* (*маркер мовної належності*) і *P_A* (*знакове представлення етимона*). Ці два параметри забезпечують унікальність кожного етимона словникової статті. Решта параметрів – факультативні. Для кожного параметра визначена формальна процедура, яка дозволяє виділити відповідний параметр з тексту для кожного етимологічного класу.

Набір параметрів ми будемо називати *етимон-структурою* і будемо позначати символом *ETYM*(*e_i*), де *e_i* – відповідний етимон; індекс – порядковий номер даного етимона в тексті.

Приклад 9 (етимон-структури для мовного класу):

LANG (абзац) $\equiv$ <запозичення з німецької мови; нім. Absatz «перерва, пауза, уступ, абзац» є похідним від дієслова absetzen «відсувати, відставляти», утвореного з префікса ab- «від-, з-», спорідненого з гот. af «від», лат. ab «тс.», і дієслова setzen «садити», пов'язаного з двн. sezzen, дангл. settan, англ. set і спорідненого з псл. saditi, укр. *cadumu*>

ETYM (*e*₁) $\equiv$ {<*P_L*= <нім.>, *P_A*= <Absatz>} }

ETYM (*e*₂) $\equiv$ {<*P_L*= <нім.>, *P_A*= <absetzen>} }

ETYM (*e*₃) $\equiv$ {<*P_L*= <нім.>, *P_A*= <ab->} }

ETYM (*e*₄) $\equiv$ {<*P_L*= <гот.>, *P_A*= <af>} }

ETYM (*e*₅) $\equiv$ {<*P_L*= <лат.>, *P_A*= <ab>} }

ETYM (*e*₆) $\equiv$ {<*P_L*= <нім.>, *P_A*= <setzen>} }

ETYM (*e*₇) $\equiv$ {<*P_L*= <двн.>, *P_A*= <sezzen>} }

ETYM (*e*₈) $\equiv$ {<*P_L*= <дангл.>, *P_A*= <settan>} }

ETYM (*e*₉) $\equiv$ {<*P_L*= <англ.>, *P_A*= <set>} }

ETYM (*e*₁₀) $\equiv$ {<*P_L*= <псл.>, *P_A*= <saditi>} }

ETYM (*e*₁₁) $\equiv$ {<*P_L*= <укр.>, *P_A*= <saditi>} }

Основна проблема створення комп'ютерних словників, виходячи з їх друкованих версій, – це формування відповідної бази даних в автоматичному режимі безпосередньо з тексту словника (парсинг). Досвід переконує, що формування лексикографічних баз даних «вручну» з великих і складних словникових текстів практично неможливо. Основна задача парсинга – автоматичне виділення визначених нами структурних елементів безпосередньо з тексту словника, оскільки саме вони виконують роль елементів лексикографічної бази даних.

Перед конверсией тексты всех томов были переведены в формат HTML и унифицированы как относительно структуры файлов, так и относительно знаковой системы. Это позволило выполнить инвентаризацию символов алфавита для представления этимонов каждого языка.

Для поддержки цифровой версии словаря построен инструментальный комплекс, который обеспечивает такие основные функции [3]:

- 1) автоматическую конверсию текстов этимологического словаря в компьютерную базу данных;
- 2) традиционный вход в систему по реестровому слову и отображение текста словарной статьи;
- 3) редактирование любого структурного элемента словарной статьи;
- 4) построение этимон-структуры для словарной статьи в ручном режиме;
- 5) автоматическое построение этимон-структуры для словарной статьи;
- 6) создание словарной статьи с определенной структурой;
- 7) индексирование словаря по заданным параметрам.

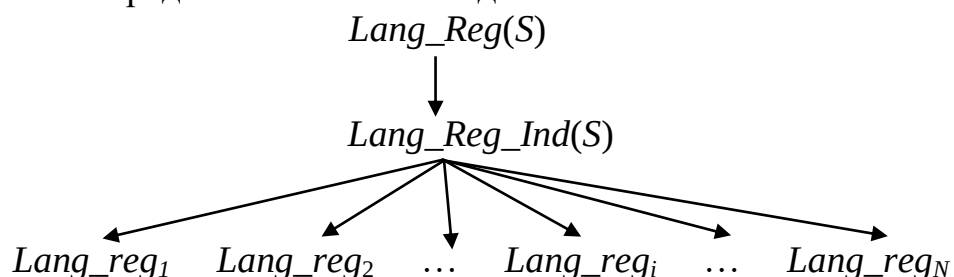
Одним из основных инструментов для эффективной работы со словарем является индекс. Индекс в его латентной форме присутствует в любом словаре: это заголовочные слова словарных статей, графически выделенные из массива текста и организованные в алфавитном порядке. При возрастании сложности структуры словарной статьи создание эффективной системы индексирования становится методологической и технологической проблемой, причем равно актуальной как для печатных словарей, так и для цифровых.

Исторически сложилось, что индексы разрабатывались, как правило, для двух типов словарей: фразеологических и этимологических. В первом случае необходимость индекса обусловлена структурной сложностью реестровой единицы; с помощью индекса организуется доступ к словарной статье по любой компоненте фразеологического словосочетания. Для этимологических словарей индекс является инструментом для установления генетической (этимологической) связи по параметру языковой принадлежности слова.

В идеале индекс должен быть построен по любому лексикографическому параметру словаря. Для реализации этого требования элементы индекса должны быть структурообразующими параметрами лексикографической системы, которая индексируется. Очевидно, развитые технологии индексирования могут быть построены только для цифровых лексикографических систем.

В предлагаемой модели структурообразующими являются обязательные параметры этимон-структуры: вход в словарь возможен по любому языку и по любому слову в алфавите этого языка.

Обобщенную схему многоязыкового индексирования этимологического словаря можно представить в таком виде:



$Lang\_Reg(S)$ – регистр всех языков (назовём его системным), наявных в этимологических описаниях словаря (к языкам также относим наречия и диалекты).

$Lang\_Reg\_Ind(S)$ – языковой регистр, выбранный для индексирования словаря. Она может совпадать с $Lang\_Reg\_Ind(S)$, или может быть его подмножеством: $Lang\_Reg\_Ind(S) \subseteq Lang\_Reg(S)$.

$Lang\_reg_i$ – некоторое подмножество индексного регистра. Как правило, в состав субрегистров входят родственные языки. Каждый субрегистр, в свою очередь, может подразделяться на более специализированные регистры ($Lang\_reg_{ij}$).

При разработке инструментального комплекса языкового индексирования мы стремились минимально ограничивать пользователя системы. Разработанный инструментарий позволяет формировать языковой регистр с любым составом языков, зафиксированных в словаре, то есть объединять в одном регистре произвольно выбранные языки, не учитывая их родственность. Можно создать любое количество таких регистров. Регистры должны различаться только именем (названием), которое присваивается пользователем. Можно задать перечень структурных элементов (этимологических классов), которые будут индексироваться.

Регистр «СЛОВ'ЯНСЬКІ МОВИ» («славянские языки») на рис. 1 сформирован следующим образом: на основе системного регистра последовательно были сформированы «східно-слов'янські мови» («восточно-славянские языки»), «західно-слов'янські мови» (западно-славянские языки), «південно-слов'янські мови» («южно-славянские языки»), а затем языковые списки были объединены.

Возможность использование разработанного инструментария для других этимологических лексикографических систем в первую очередь обусловлено возможностью использования формальной модели. Модель строится поэтапно: анализируется общая структура словаря, его знаковая система, метаязык, выделяются тексты словарных статей из общего корпуса словаря, анализируется способ распределения статьи на этимологические классы и возможность распределения класов по типам, последний шаг – способ представления этимонов и их параметров в тексте и построение этимон-структур. Наши исследования показывают, что для любой этимологической систем вполне релевантна модель, разработанная для ЭСУЯ.

На базе инструментального комплекса построена виртуальная лексикографическая лаборатория – ВЛЛ «ЕСУМ» [2], которая обеспечивает профессиональное взаимодействие лексикографов в Интернет-среде и предоставляет возможность использования всего функционала системы в режиме on-line. Эволюция системы определяется четырьмя взаимосвязанными факторами: выделением более тонких структурных элементов из базовых структур, введением новых параметров словарной статьи, расширением функционала системы и разработкой новых интерфейсных схем.

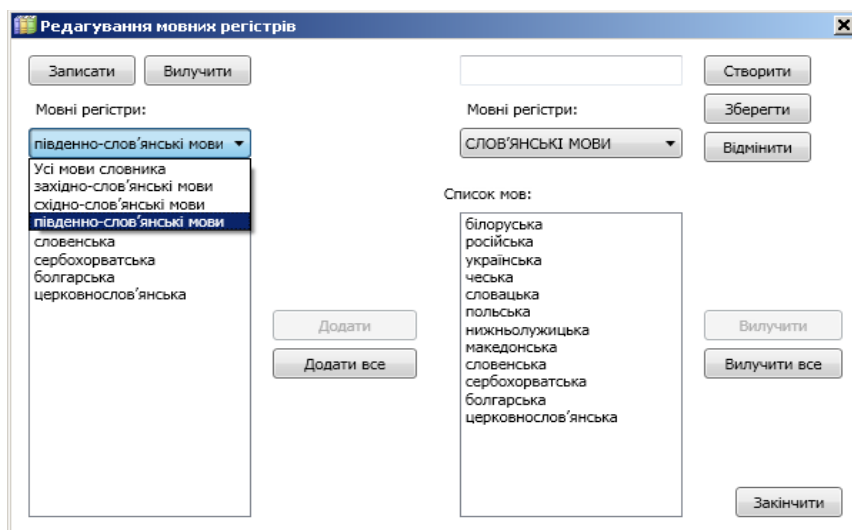


Рис. Формирование регистра «СЛОВ'ЯНСЬКІ МОВИ»

ЛИТЕРАТУРА

1. Етимологічний словник української мови. – Т. 1–6. – Київ : Наукова думка, 1982. – 2012.
2. Остапова, И. В. Виртуальная лексикографическая лаборатория для толковых словарей / И. В. Остапова, В. А. Широков // Компьютерная лингвистика и интеллектуальные технологии : матер. ежегод. Междунар. конф. «Диалог»; Бекасово, 26–30 мая 2010 г. – М. : РГГУ, 2010. – Вып. 9 (16). – С. 363–367.
3. СУМ: Словник української мови [Електронний ресурс]. – Режим доступа : <http://lcorp.ulif.org.ua/ExplS>. – Дата доступа : 12.03.2016.
4. Остапова, И. В. Лексикографическая структура этимологического словаря и его представление в цифровой среде / И. В. Остапова // Компьютерная лингвистика и интеллектуальные технологии : матер. ежегод. Междунар. конф. «Диалог 2009», Бекасово, 27–31 мая 2009 г. – М. : РГГУ, 2009. – Вып. 8 (15). – С. 359–365.

Л. В. Рычкова

ОТРАЖЕНИЕ СЕМАНТИЧЕСКОЙ ЭКВИВАЛЕНТНОСТИ В СЛОВАРЕ СОЦИАЛЬНО ЗНАЧИМОЙ СПЕЦИАЛЬНОЙ ЛЕКСИКИ БЕЛОРУССКОГО ЯЗЫКА

Язык как средство социального взаимодействия, выражения мыслей и консолидации общества, по праву считается базовым, основополагающим элементом культуры. Белорусский язык, исторически долгое время выполнявший вторичные социальные функции по отношению к польскому / русскому языкам, и в независимом Белорусском государстве по-прежнему остается языком лингвистического меньшинства: «Як і любая іншая мова, яка знаходзіцца ў суседстве з мовай самай вялікай дзяржавы свету, што мае значна большую вядомасць, значны міжнародны прэстыж, прытым, што рускамоўнае насельніцтва Беларусі вызначаецца значна большай сацыяльнай актыўнасцю, займае кіраўніцкія пасады ў дзяржаве, беларуская мова не

змога захаваць высокі сацыяльны статус. Адным з найважнейшых момантаў выступае ў гэтым напрамку пашырэнне функцый беларускай мовы ў дзяржаўнай сферы, у грамадскім жыцці, яе выкарыстанне ў органах дзяржаўнага кіравання, у дзяржаўных арганізацыях, установах, на прадпрыемствах» [1, с. 97–98].

Расширение функционала белорусского языка сдерживается недостаточным знанием большинством населения Беларуси специальной лексики, поскольку именно в профессиональной сфере, за очень немногими исключениями, практически повсеместно применяется именно русский язык: «Нават калі тэкст пішацца непасрэдна па-беларуску, і яго піша адукаваны носьбіт беларускай мовы, то з прычыны практычнага несутыкнення з беларускай мовай у прафесійных сферах адпаведны беларускі тэрмін проста невядомы, і тут у якасці арыенціра выступаюць рускія адпаведнікі» [2]. Такое положение дел приводит к тому, что в белорусскоязычных текстах зачастую вместо неизвестных авторам белорусских терминов появляются заимствования из других языков, прежде всего, путем перевода или калькирования соответствующих лексем русского языка [3], что вполне объяснимо в условиях коммуникативно несбалансированного белорусско-русского официального двуязычия в стране. Однако, как справедливо отмечает П. Н. Денисов, «необходимо различать ... заимствование, когда заимствованное слово приходит на пустое место и когда на занятое» [4, с. 74]. Как результат, «сучаснаму стану беларускай навуковай тэрміналогіі ўласціва высокая ступень сінаніміі і мнагазначнасці» [1, с. 148], которую связывают «з удасканаленнем моўнага афармлення і матываванасці тэрмінаў», одновременно отмечая «варыянтнасць беларускіх тэрмінаў розных галін ведаў» [Там же, с. 150], включая лексическую вариантность, «а таксама варыянты, якія ўзніклі на аснове выкарыстання тэрмінаў, утвораных рознымі спосабамі тэрмінаўтварэння: прамога запазычвання, калькавання, стварэння нацыянальнага эквівалента і г.д.» [Там же, с. 151].

Разнообразие терминологических вариантов повышает степень насыщенности классов семантической эквивалентности в различных терминосистемах, соответственно затрудняя понимание специальной лексики, находящей применение за пределами собственно профессиональной коммуникации, например, белорусскоязычных текстов СМИ, социальной рекламы и прочих коммуникативно валидных текстов, задачей которых является не столько информирование граждан, сколько трансляция системы ценностей белорусского государства, полезного социального знания. Очевидно, что необходима переориентация существующей лексикографической традиции с собственно лингвистических задач на задачи, обусловленные социальной прагматикой. Именно о таком типе общефилологического словаря, который «должен быть ориентирован не на систему языка, а на языковую индивидуальность, на человека – носителя языка, выполняя эстетически-воспитательную и духовно-преобразующую роль», писал в свое время Ю. Н. Караулов [5, с. 11–12]. Г. Н. Складарская назвала подобный принцип лексикографирования «антропоцентричным», переформулировав посыл Ю. Н. Караулова следующим образом: «Такой словарь должен быть антропоцентричным в двух

аспектах: 1) представлять собой идеологизированную модель, предназначенную для активного воздействия на общество; 2) ориентироваться не на систему языка, а на языковую индивидуальность, на человека...» [6, с. 18].

С учетом особенностей функционирования белорусского языка в Республике Беларусь, принцип «антропоцентризма» должен реализоваться, прежде всего, при разработке словаря социально значимой лексики терминологического и профессионального характера этого языка. Пилотная версия такого словаря (далее – Словарь) разрабатывается сотрудниками кафедры общего и славянского языкознания Гродненского государственного университета имени Янки Купалы при участии ученых Центра исследований белорусской культуры, языка и литературы НАН Беларуси в рамках совместного белорусско-российского проекта, финансируемого БРФФИ и РГНФ (соисполнителями с российской стороны являются ученые Института русского языка имени В. В. Виноградова РАН).

Особенностью микроструктуры Словаря является отражение в словарных статьях, которые строятся по принципу макростатьи, всех видов семантической эквивалентности, включая различные варианты специальных номинаций, встречающиеся в аутентичных белорусских текстах. Это соответствует принципам формирования «хорошего ТС» (терминологического словаря): «Хороший относительно полный толковый одноязычный ТС должен включать в себя максимальное число синонимов и вариантов как собственно научных, так и профессиональных, диалектных, разговорных...» [7, с. 110]. При отборе аутентичного языкового материала как «среды обитания» специальных номинаций использован принцип ориентации на современную терминологию [8, с. 137].

Вопрос семантической эквивалентности (синонимии) и вариативности специальной лексики по-разному решается в лингвистической литературе. Подробный обзор различных точек зрения, представленных в работах, посвященных этой проблеме, позволил С. Д. Шелову и А. Э. Цумареву обосновать «настоятельную необходимость уточнения соответствующих понятий и стоящей за этим уточнением совокупности теоретических положений...» [9, с. 54]. Анализ ряда научных работ, также посвященных вышеозначенной проблеме, но не вошедших в обзор [9], позволяет прийти к следующим выводам относительно синонимии, дублетности и вариативности (именно эти три понятия представляют наибольший интерес с учетом задач Словаря):

1. В зависимости от предметной области, типов текстов и анализируемых вариантов, степень вариативности терминологии в языковом материале варьируется в пределах от 15 до 30 % [10, с. 181].

2. Синонимы и гиперонимы рассматриваются как семантические варианты наряду с морфологическими и морфосинтаксическими вариантами; типология терминологических вариантов может варьироваться с учетом особенностей решаемых прикладных задач [10].

3. Отрицается терминологическая синонимия, а семантически эквивалентные термины рассматриваются как семантические варианты, или дублеты: «У адрозненне ад слоў агульналітаратурнай мовы тэрміны з розным моўным

афармленнем і аднолькавай семантыкай не маюць стылявых і сэнсавых адрозненняў, што дае падставу большасці даследчыкаў адносіць іх да ліку варыянтаў, або дублетаў» [1, с. 151].

4. Терминологические дублеты рассматриваются как синонимы. Так, под терминологическими дублетами понимают «слова или словосочетания, которые объединяются особой терминологической соотнесенностью с одним и тем же научным понятием и объектом действительности», при этом различают «три вида терминологической синонимии»: «термины-дублеты иноязычного происхождения», «термины – синтаксические синонимы», а также «дефиниционную синонимию» [11, с. 54, 57]; «Терминологические дублеты, обозначая одно и то же понятие, различаются чаще всего по своему происхождению и степени полноты раскрытия понятия» [12, с. 14].

5. Отражение вариантности в словаре позволяет «показать лексическую систему в ее функционировании», не нарушая принципа нормативности словаря: «нормативность поддерживается не включением/невключением той или иной лексической единицы в словник, а собственно лексикографическими приемами: широким набором функциональных и стилистических помет; специальной пометой запретительного характера при словах, отражающих частотные ошибки; демонстрацией лексических, морфологических и акцентологических вариантов» [13, с. 68].

Очевидно, что проблемы культуры речи важны и для составителей Словаря, который рассматривается в контексте проблемы языкового планирования в условиях интенсивного социального взаимодействия [14]. Поэтому для отражения семантической эквивалентности в Словаре используется система специальных помет, рекомендующих/нерекомендующих определенное словоупотребление. При этом семантическими дублетами считаются исключительно полные эквиваленты, которые в словарных статьях получают помету *экв.*; признается существование синонимии специальных номинаций, соответственно синонимы получают помету *сін.*; формальные варианты любого типа вводятся при помощи строчной буквы *i*, которая подается курсивом. Среди вариантов (при их наличии) один из вариантов в обязательном порядке избирается как предпочтительный и помечается пометой *рэк.*(рекомендуется), которая подается в круглых скобках. В случае наличия нежелательных либо ошибочных вариантов они помечаются пометой *(не рэк)*. Такая помета не является обязательной. Аналогичный принцип применения помет *(рэк.)/(не рэк)* используется для маркирования дублетов и синонимов, за тем только исключением, что обе эти пометы для дублетов либо синонимов не являются обязательными. Тем не менее, в случае дублетов такие пометы применяются значительно чаще, чем при синонимах, так как составители Словаря придерживаются той точки зрения, что «дублетность ... загромождает терминологию, и потому ее устранение – одна из существенных задач при составлении терминологических словарей» [12, с. 15].

Ниже в качестве примера отражения семантической дублетности, синонимии и вариативности, а также использования системы соответствующих помет приведен фрагмент словарной статьи, разработанной при участии автора О. М. Николаевой для лексемы *адзінаўладдзе*:

АДЗІНАЎЛА□ДДЗЕ (рэк.) і АДЗІНАЎЛАДСТВА

1. *гіст.* Форма праўлення, пры якой вышэйшая дзяржаўная ўлада непадзельна належыць адной асобе.

**Прыярытэтам дзейнасці Пятра I у першыя гады адзінаўладдзя быў працяг вайны з Крымам.* [Web (I): BWiki]

**Падчас далейшага ваенна-палітычнага канфлікту змог перамагчы Брута і Касія пад Філіпамі (42 да н.э.) і Марка Антонія і Клеопатру VII пад мысам Акцый (31 да н.э.), і дамогся, такім чынам, адзінаўладдзя.* [Web (I): BWiki]

экв.: абсалютызм

аўтакратыя

сін.: адзінадзяржаўе

самадзяржаўе

самаўладдзе (рэк.) і самаўладства

ТСБМ:

Энц.:

2. *гіст.* Форма праўлення, пры якой вышэйшая дзяржаўная ўлада належыць адной асобе, але пры гэтым улічваюцца асобныя палітычныя інтарэсы падданных.

**Адзінаўладдзе справядлівага манарха павінна ўлічваць інтарэсы розных груп насельніцтва і не выключае элементаў арыстакратыі і поліснай дэмакратыі.* [Web (I): BWiki]

**Марк Аўрэлій у сваіх “Развагах” піша пра прадстаўленне “аб дзяржаве, з законам, роўным для ўсіх, дзе прызнаюцца роўнасць і роўнае права на гаворку; таксама аб адзінаўладдзі, што ўсяго больш пачытае свабоду падданных”.* [Web (I): BWiki]

сін.: адзінадзяржаўе

самаўладдзе (рэк.) і самаўладства

ТСБМ:

Энц.:

3. *паліт.* Кіраванне, пры якім уся ўлада сканцэнтравана ў адных руках; паўната, неабмежаванасць чыёй-н. улады.

Іосіф Сталін, бязлітасна выразаўшы ўсіх супернікаў свайго адзінаўладдзя і падтрымліваючы сваю ўладу далейшым тэрорам, здолеў дасягнуць адзінства ўсіх трох фактараў, забяспечыўшы СССР дамінаванне і ў наступныя гады. [БН: Звязда]

сін.: самаўладдзе (рэк.) і самаўладства

ТСБМ:

Энц.:

4. *паліт.* Неабмежаванасць чыёй-н. палітычнай улады.

Каб у плане вытворчасці ўсе шасцярэнькі круціліся як у Еўропе – дакладна і без скрыту, а вось у палітыцы – каб заставалася анахранічная машына адзінаўладдзя і задушэння любой альтэрнатывы. [БН: БелаПАН]

сін.: самаўладдзе (рэк.) і самаўладства

ТСБМ:

Энц.:

Приведенный пример наглядно отражает динамику понятийного объема заголовочной единицы, актуализацию ее политической составляющей в противовес составляющей исторической. Одновременно видно, что при существовании ранее закрепившихся в белорусском языке терминов заголовочная номинация, представляющая собой кальку русского *единовластие*, несмотря на высокую степень полисемии, постепенно теснит закрепившиеся в истории и *абсалютызм*, и *аўтакратыю*.

Как видно из не получивших наполнения структурных частей словарной статьи, вводимых пометами *ТСБМ* и *Энци.*, лексикографируемая специальная номинация ни в одном из своих значений не была представлена ни в наиболее полном Толковом словаре белорусского языка (под редакцией К. Крапивы) [15], ни в Белорусской энциклопедии в 18 томах [16].

ЛИТЕРАТУРА

1. Антанюк, Л. А. Беларуская мова: прафесійная лексіка : Курс лекцый / Л. А. Антанюк, Б. А. Плотнікаў. – 3-е выд. – Мн. : Акадэмія кіравання пры Прэзідэнце Рэспублікі Беларусь, 2005. – 240 с.
2. Пошук беларускай тэрміналогіі [Электронны ресурс]. – Режим доступа : <http://movazbor.net/poshuk-terminalohii/>. – Дата доступа : 29.04.2016.
3. Рычкова, Л. В. Терминологические дублеты как результат перевода с близкородственных языков / Л. В. Рычкова // Индустрия перевода. – Пермь : Изд-во Пермского нац. исследов. политехнич. ун-та, 2016 (в печати).
4. Денисов, П. Н. Еще о некоторых аспектах изучения языков науки / П. Н. Денисов // Проблемы языка науки и техники. Логические, лингвистические и историко-научные аспекты терминологии. – М. : Наука, 1970. – С. 52–89.
5. Караулов, Ю. Н. Современное состояние и тенденции развития русской лексикографии / Ю. Н. Караулов // Советская лексикография. – М. : Русский язык, 1978. – С. 5–18.
6. Складаревская, Г. Н. «Нормативный толковый словарь живого русского языка»: основные постулаты и типологические признаки / Г. Н. Складаревская // Нормативный толковый словарь живого русского языка: теоретические проблемы и практические трудности. – СПб. : Филол. фак-т СПбГУ, 2007. – С. 17–41.
7. Герд, А. С. Терминологический словарь среди других типов словарей / А. С. Герд // Современная русская лексикография, 1980. – Л. : Наука, 1981. – С. 106–112.
8. Герд, А. С. Терминологические словари и их источники / А. С. Герд // Современная русская лексикография, 1981. – Л. : Наука, 1983. – С. 136–143.
9. Шелов, С. Д. О парадигматических отношениях в терминологии: вариативность и синонимия / С. Д. Шелов, А. Э. Цумарев // Язык. Культура. Речевое общение. – 2013 – № 2. – С. 42 – 59.
10. Daille, B. Variations and application-oriented terminology engineering / B. Daille // Terminology. – 2005. – № 11:1. – P. 191–197.
11. Головин, Б. Н. Лингвистические основы учения о терминах / Б. Н. Головин, Р. Ю. Кобрин. – М. : Высш. шк., 1987. – 104 с.
12. Гаранина, Н. С. Специальная лексика / Н.С. Гаранина // Учеб. пособие по лексике современного рус. языка для студентов-заочников фак-тов журналистики гос. ун-тов. – М. : МГУ, 1967. – 48 с.
13. Дурнева, А. А. Вариативность в современном русском языке (проблемы описания в «Нормативном толковом словаре живого русского языка») / А. А. Дурнева // Нормативный толковый словарь живого русского языка: теоретические проблемы и практические трудности. – СПб. : Филол. фак-т СПбГУ, 2007. – С. 64 –79.

14. Рычкова, Л. В. Языковое планирование в условиях интенсивного социального взаимодействия / Л. В. Рычкова // Язык и социум : матер. X Междунар. науч. конф., Минск, 15–17 окт. 2015 г., БГУ. – Минск : Колорград, 2016. – С. 8–12.
15. Тлумачальны слоўнік беларускай мовы; пад аг. рэд. К. К. Атраховіча : у 5 т. – Мінск : Бел. сав. энцыклапедыя, 1977–1984. – Т. 1–5.
16. Беларуская энцыклапедыя; гал. рэд. Г. Пашкоў, Б. Сачанка : у 18 т. – Мінск : Бел. Энцыклапедыя, 1996–2004.

Н. М. Сянкевіч

КАНКАРДАНС ЯК СРОДАК ВЫЯЎЛЕННЯ АСАБЛІВАСЦЕЙ МОВЫ МАСТАЦКАЙ ЛІТАРАТУРЫ

Канкарданс традыцыйна лічыцца спецыфічным даведнікам, якім карыстаюцца навукоўцы пры правядзенні даследаванняў у розных галінах мовазнаўства. Слова *канкарданс* доўгі час было зразумелым толькі для вузкіх спецыялістаў, якія мелі досвед у камп’ютарнай лексікаграфіі, аўтаматычнай апрацоўцы тэкстаў ці корпуснай лінгвістыцы. І гэта нягледзячы на тое, што гэты лексікаграфічны даведнік, які ствараецца паводле тэксту ці корпуса тэкстаў і ў самым простым выглядзе ўяўляе сабой спіс усіх без выключэння словаўжыванняў тэксту разам з іх тэкставымі ілюстрацыямі і адрасамі ў тэксце, мае даўнюю гісторыю. Лічыцца, што першы канкарданс быў створаны яшчэ ў XIII стагоддзі. Гэта быў канкарданс да Бібліі. У працы над ім удзельнічалі 500 манахаў. Менавіта з-за вялікага аб’ёму працы ў дакамп’ютарны час канкардансы складаліся толькі для твораў, якія мелі найвялікшую важнасць: Біблія, Веды, Каран, творы Шэкспіра.

Сярэдзіну мінулага стагоддзя можна назваць эрай росквіту гэтай справы. Развіццё вылічальнай тэхнікі дазволіла ў кароткі тэрмін апрацоўваць вялікія аб’ёмы тэкстаў, аўтаматызаваць нятворчыя віды работ, узяўшы такім чынам лексікаграфічную працу на якасна новы ўзровень. Канкардансы да твораў асобных аўтараў ці асобных твораў сталі карыстацца вялікай папулярнасцю, асабліва сярод замежных лексікографаў. Аб гэтым сведчыць той факт, што яны былі складзеныя да твораў амаль усіх вядомых англамоўных аўтараў (найбольш вядомым з іх з’яўляецца канкарданс да твораў У. Шэкспіра [1]), а таксама некаторых рускіх: А. Пушкіна [2], В. Мандэльштама [3], Ф. Цютчава [4], Я. Баратынскага [5] і інш. Канкардансы і цяпер застаюцца асноўным тыпам пісьменніцкіх слоўнікаў у замежнай лексікаграфіі. У апошнія дзесяцігоддзі назіраецца ўздым цікавасці расійскіх моваведаў да гэтага тыпу даведнікаў. Ствараюцца як кніжныя варыянты канкардансаў, напрыклад, канкардансы да твораў М. Кузміна [6], рамана А. Пушкіна “Евгений Онегин” [7], так і электронныя: паводле твораў Ф. Цютчава [8], Ф. Дастаеўскага [9], В. Даля [10], М. Ламаносава [11], С. Аксакава [12] і інш.

Першыя спробы стварыць канкарданс да беларускамоўных літаратурных твораў былі зробленыя напрыканцы 70-х гадоў мінулага стагоддзя, калі ў Інстытуце мовазнаўства Акадэміі навук Беларусі пачалася праца па складанні канкарданса беларускай мовы XIX ст. Асноўным крытэрыем выбару тэкстаў

для стварэння першага канкарданса стала значнасць гэтага перыяду ў развіцці беларускай мовы як этапа станаўлення яе літаратурнай формы і недаследаванасць тагачаснай мовы, выкліканая адсутнасцю дастатковага фактычнага матэрыялу. Няма сумнення ў тым, што даследаванні літаратурнай мовы павінны мець у сваёй аснове вывучэнне літаратурных крыніц. А для даследавання беларускай мовы XIX ст., якая складалася як мова мастацкай літаратуры, гэта асабліва актуальна. Вынікам збору тэкставага матэрыялу стаў корпус тэкстаў беларускамоўнай мастацкай літаратуры XIX ст., які складаўся з 283 твораў прозы і паэзіі. Быў распрацаваны камплект праграм для аўтаматычнай апрацоўкі тэкстаў, з дапамогай якога на аснове корпуса тэкстаў былі складзеныя ўласна канкарданс, некалькі частотных слоўнікаў, адваротны слоўнік і паказальнік слоў. Праца над першым беларускім канкардансам была завершаная ў 1985 годзе, але вялікі аб'ём матэрыялу (каля 5 тыс. старонак) не дазваляў выдаць канкарданс у кніжным варыянце, таму яго раздрукавалі ў адным экзэмпляры, і толькі ў такім выглядзе ён стаў даступны для карыстальнікаў.

Другім завершаным беларускім даведнікам такога тыпу з'явіўся канкарданс паводле твораў класіка беларускай літаратуры XX ст. К. Крапівы. Праца над ім скончылася ў 1996 г., аднак пытанне аб выданні даведніка на той момант таксама не ставілася. Канкардансы церпяліва чакалі таго часу, калі стане магчымым змяшчаць шматтомныя даведнікі на маленькім электронным дыску. Калі гэты час настаў, канкардансы былі ўзноўленыя, і ў 2011 г. пабачыў свет Канкарданс мовы Кандрата Крапівы [13].

Гісторыя ж стварэння канкарданса беларускай мовы XIX ст. атрымала свой працяг. Справа ў тым, што за той час, што канкарданс чакаў свайго выдання, даследчыкамі беларускай літаратуры былі знойдзеныя шматлікія літаратурныя помнікі, якія не ўвайшлі ў корпус тэкстаў, паводле якога ён быў складзены. Выданне няпоўнага канкарданса губляла сэнс, таму што была б страчаная асноўная вартасць даведніка такога тыпу: паўната і аб'ектыўнасць пададзенай інфармацыі, якая забяспечваецца апрацоўкай максімальна поўнага тэкставага матэрыялу. На працягу 5 гадоў быў складзены новы корпус тэкстаў, які павялічыўся амаль удвая: з 283 да 497 тэкстаў, з 145 тыс. да 257 тыс. словаўжыванняў. З дапамогай мадэрнізаванага камплекта праграм быў створаны новы канкарданс, тэкставым матэрыялам для якога сталі практычна ўсе вядомыя творы мастацкай літаратуры, напісаныя на беларускай мове ў XIX стагоддзі. Напрыканцы 2015 г. быў выдадзены электронны варыянт даведніка [14]. Агульны канкарданс дапоўнілі канкардансы мовы асобных пісьменнікаў, якія адыгралі вялікую ролю ў фарміраванні новай беларускай літаратурнай мовы ў XIX ст.: В. Дуніна-Марцінкевіча, А. Ельскага, К. Каганца і К. Каліноўскага.

Усе створаныя канкардансы пабудаваныя паводле адной схемы, таму складаюцца з аднолькавых кампанентаў: уласна канкарданс, частотныя і алфавітна-частотныя слоўнікі, паказальнік слоў, адваротны слоўнік, спісы тэкстаў і тэксты, на аснове якіх яны складзеныя. Слоўнікавы артыкул уласна канкарданса ў кожным з даведнікаў таксама мае аднолькавую структуру і ўключае загалоўнае слова, агульную частату, частату паводле асобных жанраў, адрасы слова ў тэкстах (нумары тэкстаў, старонак, радкоў) і тэкставую ілюстрацыю слова.

Фрагмент канкарданса беларускай мовы XIX ст.

Слова	Частага	Част. у прозе	Част. у паэзіі	№ тэксту	№ старонкі	№ радка	Кантэкст
ШЧАСНУ	3	1	2	123	057	11	!- КАБЫ МАЦІ , МАУЛЯУ , СВАЯМУ ДЗІЦЯЦІ ДОЛЮ ШЧАСНУ ДАРАВАЛА ;
				278	061	05	. ДЗЯКУЙ ТАБЕ ПАНЕ ЗА ШЧАСНУ ГАДЗІНУ .
				283	029	21	НА СТРОП ГЛЯНУЦЬ ,- ПЕСЦІУ ДУМУ ШЧАСНУ , ШТО ПЕУНЕ ТАМ УГЛЕДЗІЦЬ СУКЕНАЧКУ ЯСНУ,
ШЧАСНЫ	1		1	283	044	05	.- ПАГЛЯД НА ТАДАУША ШЧАСНЫ ВІДЫ ГЭТКІ , МАНІЛІ БЛУДНЫ МЫСЛІ ПРЫГОЖАЙ СУСЕДКІ,-
ШЧАСЦЕ	4		4	018	037	06	ГРОШАХ , ЕН ДУМАУ , УСЕ ЛЮДЗКОЯ ШЧАСЦЕ , А ЕН ЖЭ ПАВІНЕН ГДЗЕСЬЦІ СКАРБ
				022	068	06	; ІДУ І МЯРКУЮ : ЦІ ТО ЦЯПЕР ШЧАСЦЕ , ЯК ПАНОУ НЯ СТАЛА , ЦІ ТО БЫЛА ДОЛЯ ?
				136	007	13	; НЕХАЙ ЯГО РАДНІ БУОГ ШЧАСЦЕ ГАТУЯ ШТО ЮОН НАШЫХ БРАТОУ ВОЛЬНАСЦЮ ДАРУЯ .
				156	019	06	. ВЕРЦЕ ЛЮДЗІ : ШЧАСЦЕ БУДЗЕ , ТАК , ЯК ПЧОЛКАМ У УЛІ !..
ШЧАСЦІЕ	3	3		024	017	12	! БО ТАМ ТО ШЧАСЦІЕ І РАДОСЦЬ ВЕЧНАЯ ТАМ З БОГОМ І АНЕЛАМІ ЖЫЦЬ .
				024	023	17	БОГА І У ЛАСЦЭ ЯГО СВЯТОЙ , КАБ ПОСЛІ СМЕРЦІ ЗАРОБЦЬ НА ШЧАСЦІЕ ВЕЧНОЕ .
				024	030	10	; ЯКОЕ ГЭТО ШЧАСЦІЕ УМЕЦЬ УСЯЧЫНУ ПЕРЭЧЫТАЦЬ , ШТО ЗАХОЧЭШ НАПШЭШ , І ПІСАРА НІЯКОГО ПРАСЦЬ НЕ
ШЧАСЦІНЬКА	1		1	209	014	19	НАС ДОЛІ НЕТ НІ У НАС ШЧАСЦІНЬКА А ПАСТЫУ НАМ СВЕТА ЗЯМЛЯ МАТЫНЬКА
ШЧАСЦЮ	1	1		024	019	02	. П. А ЧЫ БУДЗЕЦЬ КАНЕЦ ШЧАСЦЮ НЯБЕСКОМУ ?

З выданнем канкардансаў лінгвісты атрымалі магутны сродак для вывучэння асаблівасцяў беларускай мовы на розных этапах яе развіцця. У XIX ст., як вядома, умовы функцыявання беларускай мовы былі вельмі неспрыяльнымі. Больш за стагоддзе яна не ўжывалася як літаратурная, таму сучасная пісьмовая мова не магла карыстацца пісьмовымі здабыткамі папярэдніх часоў. Яна пачала фарміравацца непасрэдна на аснове жывых беларускіх гавораў. Пісьменнікі ў сваёй творчасці адлюстроўвалі рысы сваіх родных дыялектаў, уключаючы ў творы словы, якія часта ўжываліся толькі ў той мясцовасці, дзе яны жылі. Некаторыя з дыялектызмаў у далейшым сталі літаратурнай нормай, іншыя не замацаваліся ў сучаснай літаратурнай мове. Канкарданс, які змяшчае поўны спіс усіх словаўжыванняў, зарэгістраваных у творах мастацкай літаратуры таго часу, і бібліяграфічныя звесткі аб крыніцах, дзе яны зафіксаваныя, дае магчымасць прасачыць гісторыю многіх слоў, якія з народнага слоўніка трапілі ў літаратурную мову. Ён забяспечвае гісторыка беларускай мовы вычарпальнай інфармацыяй аб тым, хто з пісьменнікаў першы ўжыў слова, якое яго цікавіць, у чыіх яшчэ творах яно сустракаецца. Напрыклад, з дапамогай гэтага даведніка можна лёгка выявіць, адкуль і калі такія цікавыя дыялектызмы, як *мэрам* 'быццам' і *суздром* 'зусім, поўнасьцю' прыйшлі ў літаратурную мову. Так, слова *суздром* даволі часта ўжывае В. Дунін-Марцінкевіч: зафіксавана 8 выпадкаў у 6 тэкстах. І калі яно сустракаецца толькі ў творах гэтага аўтара, то слова *мэрам* у розных графічных варыянтах часцей за ўсё (7 разоў) ужываецца ў творах А. Ельскага (*мэрам*, *мерам*), 3 разы ў творах В. Дуніна-Марцінкевіча (*мэрэм*) і 1 раз у творы Шахава М. "З жыцця" (*мерем*). Калі ўлічыць час напісання твораў, дзе сустрэлася гэтае слова, то можна прыйсці да высновы, што менавіта В. Дунін-Марцінкевіч першым увёў слова *мэрам* у літаратурную мову.

З дапамогай канкарданса можна прааналізаваць слоўнікавы склад літаратурнай мовы не толькі ў адносінах паходжання, але і з пункту гледжання семантыкі. Поўны пералік кантэкстаў для кожнага словаўжывання дазваляе аб'ектыўна семантызаваць тое ці іншае слова, выяўляючы такім чынам полісемантычныя словы, семантычныя лакалізмы. Некаторыя словы, якія ўжываюцца і ў сучаснай беларускай мове, у XIX ст. мелі адрознае ад традыцыйнага значэнне. Напрыклад, слова *злыдзень* зафіксавана ў тэкстах твораў трох аўтараў: Ф. Багушэвіча (5 разоў), В. Дуніна-Марцінкевіча (2 разы) і А. Ельскага (4 разы). Усе аўтары ўжываюць яго ў прамым значэнні 'шкодны чалавек, ліхадзей'. Аднак у вершах Ф. Багушэвіча ў 4 выпадках гэтае слова ў множным ліку абазначае дробныя гаспадарчыя рэчы: *Так я праляжаў тры тыдні, прадаў усе свае злыдні*. Незвычайнае значэнне 'труна' набывае слова *дошка* ў вершах Ф. Багушэвіча: *Якось яно будзе, дажывем да дошкі, хоць лепей, хоць горэй, абы не так трошкі!*, ці *Склаў я гора ды у дошкі, закопаў у велькі роў*. У перакладзе паэмы А. Міцкевіча "Пан Тадэвуш" В. Дунін-Марцінкевіч ужывае слова *агуркі* для абазначэння белых шнуркоў з вузлякамі, якімі падпярэзваліся бернардынцы.

У літаратурных творах XIX ст. сустракаецца шмат слоў, якія могуць быць незразумелымі для сучаснага чытача. Вырашыць гэтую праблему прызначаны кароткі тлумачальны слоўнік, які ствараецца на аснове канкарданса

ў Інстытуце мовазнаўства. З дапамогай комплексу даведнікаў, з якіх складаецца канкарданс, можна знайсці шматлікія графічныя варыянты слоў і роднасныя словы, каб скласці дакладны рэестр слоўніка, выявіць спецыфічныя значэнні таго ці іншага слова.

Такая лексікаграфічная форма як канкарданс вельмі карысная пры складанні анамастычнага слоўніка. Ён дазваляе правесці даследаванне асаблівасцей анамастычнай лексікі, апісаць яе ролю ў словатворчасці пісьменнікаў. Наяўнасць поўнага спісу словаўжыванняў дапаможа складальніку анамастычнага слоўніка ў сціслыя тэрміны сфарміраваць слоўнікавы рэестр, не згубіўшы ніводнага з уласных імёнаў, зарэгістраваных у тэксце. Часцей за ўсё канкардансы з'яўляюцца слоўнікамі форм, а не лем. А калі пры складанні даведніка лематызацыя ўсё ж праводзіцца, то граматычная інфармацыя падаецца асобна. Гэта забяспечвае зручнасць пры фарміраванні спісу склонавых форм слова, якія, як правіла, прыводзяцца ў слоўнікавым артыкуле анамастычнага слоўніка. Звесткі пра частотнасць слоўнікавых адзінак ці іх форм здабыць у канкардансе таксама вельмі лёгка. Несумненна, асноўнай вартасцю канкардана з'яўляецца наяўнасць кантэкстаў для кожнага словаўжывання, таму складальнік анамастычнага слоўніка можа з лёгкасцю выбраць той адрэзак тэксту, які, на яго думку, найлепш ілюструе ўжыванне пэўнага оніма ў тэксце. Усе перавагі, якія дае канкарданс пры стварэнні анамастычнага слоўніка, ужо паспелі ацаніць укладальнікі анамастычнага слоўніка беларускай мовы XIX ст., праца над якім пачалася ў Інстытуце мовазнаўства Акадэміі навук.

Канкардансы карыстаюцца вялікай папулярнасцю ў даследчыкаў мовы асобных аўтараў як самастойны інструмент даследавання, а таксама ў якасці асновы для складання слоўнікаў мовы пісьменніка. Такі даведнік дазваляе не толькі прааналізаваць індывідуальную лексіка-семантычную сістэму словаўжывання пісьменніка, правесці пошук і аналіз фразеалагізмаў, раскрыць асаблівасці аўтарскага стылю, вызначыць тэматычную накіраванасць іх твораў. Ён мае таксама і сацыяльна-гістарычнае значэнне: характарызуе літаратурную мову і адлюстроўвае рэаліі таго часу, калі жыў пісьменнік. У гэтым сэнсе вялікае значэнне мае канкарданс К. Крапівы. Калі прыняць да ўвагі той факт, што творчы шлях К. Крапівы доўжыўся амаль 60 гадоў, на працягу якіх была сфарміраваная багатая творчая спадчына пісьменніка (корпус тэкстаў уключае 277 твораў розных жанраў), і ўлічыць яго ролю ў выпрацоўцы нормаў і развіцці беларускай мовы, то становіцца відавочным, што з дапамогай канкардана яго мовы можна даследаваць асаблівасці беларускай мовы ў цэлым на працягу пэўнага перыяду.

Поўнавартаснае даследаванне асаблівасцей мовы таго ці іншага аўтара немагчыма без аналізу колькасных паказчыкаў. І з гэтага пункту погляду створаныя канкардансы мовы пісьменнікаў XIX ст. і К. Крапівы ўяўляюць сабой вельмі зручны інструмент для даследавання асаблівасцей, якія характэрныя як для мовы пісьменнікаў ў цэлым, так і для асобных жанраў. Адной з задач частотных слоўнікаў, як вядома, з'яўляецца дыферэнцыяцыя слоўнікавага запasu. Не выклікае сумнення той факт, што ў даследаванні мовы пісьменніка вялікае значэнне мае як высокачастотная лексіка, якая дазваляе

вызначыць тэматычную накіраванасць яго твораў, так і нізкачастотная, якая адлюстроўвае індывідуальнае словаўжыванне аўтара. З дапамогай частотнага слоўніка даследчык можа прасачыць, як змяняліся творчыя інтарэсы пісьменніка, параўнаць тэматыку яго твораў з тэматыкай твораў яго калег па прэу, прааналізаваць развіццё лексічнага складу мовы на працягу пэўнага гістарычнага перыяду.

Шматгадовая гісторыя існавання канкардансаў пацвердзіла значнасць гэтай лексікаграфічнай формы як эфектыўнага інструмента для рознабаковых мовазнаўчых даследаванняў у галіне лексікі і фразеалогіі, семантыкі і стылістыкі, марфалогіі і сінтаксісу. Слушную ацэнку ролі канкардансаў даў расійскі акадэмік М. Л. Гаспараў у прадмове да расійскага выдання “Канкарданса да вершаў А. С. Пушкіна”: “Конкорданс – это расширенный словарь языка писателя. ... Все мельчайшие оттенки значения каждого слова выступают при этом так исчерпывающе ясно, как ни в каком словаре. Изучать по-настоящему язык и литературу, не имея таких конкордансов, невозможно” [15, с. 111].

ЛІТАРАТУРА

1. *Spevack, M. A Complete and Systematic Concordance to the Works of Shakespeare / M. Spevack. – Hildesheim, 1968.*
2. *Shaw, J. Thomas. Pushkin: A Concordance To The Poetry / J. Thomas Shaw. – Slavica Publishers, 1985.*
3. *A Concordance to the Poems of Osip Mandelstam (The Cornell concordances) / Demetrius J. Koubourlis (editor). – Cornell University Press, 1974.*
4. *Bilokur, Borys. Concordance To The Russian Poetry Of Fedor I. Tiutchev / Borys Bilokur. – 1975.*
5. *Shaw, Joseph Thomas. Baratynski: A dictionary of the rhymes & a concordance to the poetry (Wisconsin slavic publications) / Joseph Thomas Shaw. – University of Wisconsin Press, 1975.*
6. Конкорданс к стихотворениям М. Кузмина; сост. А. В. Гик. – М. : Языки славянских культур, 2005–2011. – Т. 1–3.
7. *Гайдуков, Д. А. Опыт конкорданса к роману в стихах А. С. Пушкина “Евгений Онегин” с приложением текста романа / Д. А. Гайдуков; науч. ред. С. И. Гиндин. – М. : ЛексЭст, 2003. – 592 с.*
8. Частотный словарь-конкорданс словоформ языка Ф. И. Тютчева [Электронный ресурс]. – Режим доступа : <http://www.ruthenia.ru/tiutcheviana/-stihi/freq/k.html>. – Дата доступа : 22.02.2016.
9. Конкордансы всех произведений Ф. М. Достоевского (1997–2007) [Электронный ресурс]. – Режим доступа : <http://dostoevskij.karelia.ru>. – Дата доступа : 22.02.2016.
10. Конкорданс произведений В. Даля (2001) [Электронный ресурс]. – Режим доступа : http://elibrary.karelia.ru/dahl/user_new/index.php. – Дата доступа : 22.02.2016.
11. Конкорданс к текстам Ломоносова [Электронный ресурс]. – Режим доступа : <http://feb-web.ru/feb/lomocopc/abc>. – Дата доступа : 22.02.2016.
12. Конкорданс произведений С. Т. Аксакова [Электронный ресурс]. – Режим доступа : <http://aksakov.do.am/index/konkordans/0-22>. – Дата доступа : 22.02.2016.
13. Канкарданс Кандрата Крапівы [Электронны рэсурс] / Аўтар-складальнік Н. М. Сянкевіч; навуковае суправаджэнне В. У. Мартынава. – Электронныя тэкставыя даныя і праграма (133 Мб). – Мінск : Нацыянальная бібліятэка Беларусі, 2011.

14. Канкарданс беларускай мовы XIX стагоддзя [Электронны рэсурс] / Аўтар-складальнік Н. М. Сянкевіч. – Электронныя, тэкставыя даныя і праграма (350 Мб). – Мінск : Беларуская навука, 2015.
15. *Гаспаров, М. Л.* Предисловие к российскому изданию / М. Л. Гаспаров // Конкорданс к стихам А. С. Пушкина. – М. : Языки русской культуры, 2000. – Т. 1.

В. К. Щербин

СРАВНИТЕЛЬНАЯ ЛЕКСИКОГРАФИЯ И ПАРАЛЛЕЛЬНАЯ ТЕКСТОГРАФИЯ: АНАЛИЗ ВЗАИМОСВЯЗЕЙ

Как свидетельствуют историки лексикографии [1, с. 189], первоначальной формой словарной деятельности была текстовая лексикография (или глоссография), которая развивалась в виде разнообразных глосс (толкований непонятных слов) на полях или между строками в текстах древних рукописей XI–XVI столетий. Со временем практика глоссирования непонятных слов в текстах настолько усовершенствовалась и специализировалась, что стали различаться следующие виды глосс: толковые, текстологические, этимологические, ссылочные [2, с. 216]; эквивалентные [3, с. 222], грамматические [4, с. 62] и др. Именно эти отличающиеся виды глосс послужили прообразами тех лексикографических параметров (толкового, иллюстративного, этимологического, отсылочного, переводного, грамматического и др.), доминированием которых в том или ином словаре сегодня определяется тип такого словаря [5, с. 29]. Для данной статьи наибольший интерес представляют иллюстративный и переводной параметры, в основе которых, по мнению Е. М. Верещагина и В. Г. Костомарова, лежат приемы контекстного изъяснения и эквиваленции [3, с. 222].

На наш взгляд, «прием контекстного изъяснения слова (путем включения его в контекст литературной цитаты, пословицы, поговорки, специально созданного лексикографом типового предложения) выступает одним из ключевых лексикографических параметров в справочных изданиях таких типов, как толковые, исторические словари, конкордансы, словари языка писателя и др.» [5, с. 29]. Причем наиболее широкое и разнообразное использование иллюстративный (цитатный, контекстологический) параметр или «контекст-глосса» (по выражению Ю. Н. Трубачева) получил не в общеязыковых справочниках, а в двуязычных терминологических словарях, в которых, как правило, ставится задача дать максимально точное и по возможности однозначное толкование понятийного содержания описываемого термина и его иноязычного эквивалента. С этой целью в двуязычных терминологических словарях используются следующие типы научных цитат: 1) фрагменты научных текстов (терминированные словосочетания и сложные номинативные единства); 2) цитаты-дефиниции из других словарей; 3) авторские глоссы из научных текстов; 4) текстовые единицы понимания; 5) контекстные изъяснения [6].

К примеру, научные цитаты первого типа (фрагменты научных текстов) активно использовались в двух выпусках «Тематического русско-белорус-

ского словаря языка научных отчетов» (далее ТРБСЯНО) [7; 8]. Так, в первом выпуске указанного словаря (в первой части его макроструктуры) насчитывается около 4 тысяч словарных статей со стандартной микроструктурой (русская реестровая единица – белорусский эквивалент – порядковый номер текстового фрагмента из второй части словаря, в котором данная реестровая единица употребляется). Во второй части макроструктуры указанного справочника «подаются «оправдательные» цитаты для всех без исключения реестровых русских единиц и их белорусских эквивалентов из первой части словаря. В качестве таких цитат используются полностью эквивалентные по содержанию текстовые фрагменты научных отчетов на русском и белорусском языках, которые приводятся в виде параллельных колонок на одной и той же странице словаря. Каждая пара таких колонок имеет свой порядковый номер (от 1 до 510). При этом, как легко заметить, общее количество «оправдательных» цитат во второй части словаря уступает общему количеству реестровых единиц из первой части словаря, поскольку каждый текстовый фрагмент используется в качестве иллюстрации в среднем для 10 реестровых единиц» [7, с. 11]. Кроме того, объем отдельных текстовых фрагментов в ТРБСЯНО колебался от 2–3 до 60-ти слов и более в зависимости от степени известности реестрового термина. Как правило, менее известные термины иллюстрировались и семантизировались посредством использования более объемных текстовых фрагментов.

Научная общественность Беларуси положительно восприняла новшества, представленные в указанном словаре: 1) иллюстрацию и семантизацию русских реестровых единиц и их белорусских эквивалентов при помощи текстовых фрагментов научных отчетов, совокупность которых представлена в макроструктуре словаря отдельной, второй частью (ее содержание упорядочено при помощи цифрового указателя); 2) возможность использования второй части словаря в качестве русско-белорусской сопоставительной грамматики. Свидетельством позитивной оценки ТРБСЯНО стало значительное количество заявок на его получение, поступивших от академических институтов физико-математического и физико-технического профиля. Более того, один из читателей словаря, доктор технических наук А. И. Добролюбов, используя макроструктурную схему ТРБСЯНО, составил близкий по типу словарь английского языка [9] и даже запатентовал метод его создания в России [10]. И хотя уже в самом названии словаря А. И. Добролюбова декларируется его принципиальная новизна, в структурном отношении данный словарь как две капли воды похож на ТРБСЯНО, изданный на несколько лет раньше.

Определяя перспективы дальнейшего использования иллюстративного параметра в словарях разных типов, можно сделать следующий вывод: в современных условиях, 1) когда наличие Интернета и электронных корпусов национальных языков позволяет практически мгновенно находить необходимый фрагмент текста, в рамках которого используются интересные исследователя слова, термины и другие языковые единицы; 2) когда требования международного авторского права и словарной этики становятся все более жесткими в отношении случаев плагиата и компиляции чужих высказываний

в справочных изданиях; 3) когда юридически корректное использование высказываний известных ученых придает справочному изданию ореол научности; 4) когда источниками цитат все чаще становятся авторские тексты, представленные в электронной форме, что фактически исключает ошибки в цитировании этих текстов и позволяет представить их в оригинале, не вызывая сомнений обоснованность прогноза о том, что частота и объемы использования цитат в словарях разных типов со временем будут только увеличиваться [6, с. 168].

Более того, по мере развития инструментальных возможностей современной корпусной лингвистики и расширения практики использования цитат в словарях разных типов, можно смело прогнозировать развитие ряда новых лексикографических направлений: словарно-ориентированной текстографии, конкордансной и гипертекстовой лексикографии. К примеру, в области авторской лексикографии сегодня уже нередким является включение в состав словаря языка писателя (в виде приложения) полнообъемных художественных текстов самого писателя (см., например, [11]). В свою очередь, в рамках диалектной лексикографии организован выпуск региональных текстографических сборников, неотъемлемыми компонентами которых являются а) подобранные по географическому и тематическому принципам образцы диалектных и этнолингвистических текстов; б) лексические и фразеологические словари говоров (см., например, [12]).

Успешно развивается в Беларуси и конкордансная лексикография, саму основу которой составляет широкое использование иллюстративного параметра, поскольку конкорданс к тому или иному тексту (совокупности текстов) – это не что иное, как полный перечень выявленных в данном тексте словоупотреблений с контекстами их использования и адресными характеристиками. К настоящему времени в виде CD-дисков выпущены: 1) электронный «Канкарданс Кандрата Крапівы» (2011); 2) электронный «Канкарданс беларускай мовы XIX стагоддзя» (2016), являющийся комплексом словарей, которые были составлены на основе корпуса текстов, включающего все произведения белорусской художественной литературы XIX века (всего 497 текстов).

Наконец, импульс развитию гипертекстовой лексикографии дало появление Интернета и лежащих в его основе сетевых технологий. Именно широкое использование современных компьютерных гипертекстовых технологий позволило издательской компании «Encyclopaedia Britannica» создать в электронном и печатном вариантах такой объемный и комплексный (линейно-нелинейный) гипертекст, как «Великие книги западной цивилизации». 60 томов печатного варианта данного гипертекста содержат 517 линейных прецедентных текстов 130 авторов, к которым прилагается уникальный двухтомный индекс «Синтопикон». При помощи индекса содержание указанных линейных текстов группируется по 3020 «революционным идеям», наиболее повлиявшим на развитие западного мира [13, с. 32]. Создание данного гипертекста наглядно показало, что «объем памяти и быстродействие современных ЭВМ уже таковы, что позволяют, в отличие от описанных выше печатных словарей цитат, хранить на машинных носителях огромное

количество полнообъемных линейных текстов, осуществляя при этом мгновенный вход в эти тексты практически через любое слово, встречающееся в них» [Там же, с. 32].

В свою очередь, «прием эквиваленции является, пожалуй, наиболее распространенным средством семантизации реестровых слов в таких типах словарей, как переводные (двуязычные и многоязычные), синонимические и др.» [5, с. 29]. Причем, если раньше в переводных словарях использование приема эквиваленции ограничивалось рамками отдельных слов и словосочетаний, то в микроструктуре словарной статьи сравнительных (двуязычных и многоязычных) словарях концептов нередко представлены полностью эквивалентные (параллельные) тексты на разных языках. Иногда объем этих текстов составляет несколько страниц (см., например, русскоязычные, польскоязычные и англоязычные части биографических словарных статей в польском пятитомном справочнике «*Idee w Rosji. Идеи в России. Ideas in Russia: Leksikon rosyjsko-polsko-angielski*» (1999–2003)).

В Беларуси, где еще в советский период широко практиковалось одновременное использование как белорусского, так и русского языков, имеющих сегодня статус государственных языков, нередким является издание полностью эквивалентных по содержанию белорусскоязычного и русскоязычного справочников. Например: пятитомные «Беларуская ССР. Кароткая энцыклапедыя» (1978–1981) и «Белорусская ССР. Краткая энциклопедия» (1979–1982); «Слоўнік фізічных і астранамічных тэрмінаў» А. И. Болсуна и Е. Н. Рапановича (1979) и их же «Словарь физических и астрономических терминов» (1986) и т.д. Аналогичным образом дело обстоит и с учебной литературой Беларуси: «учебники представляют собой параллельные тексты, т.е. учебник по какому-либо предмету, написанный на белорусском языке, является переводом того же самого учебника, написанного на русском языке, и наоборот» [14, с. 44].

Сравнительная лексикография Беларуси может получить дополнительный импульс в своем развитии, если в каждой области профессиональных знаний будет создана необходимая текстовая база в виде электронных корпусов параллельных текстов на государственных языках Беларуси. Наличие такой текстовой базы содействовало бы также качественному выполнению всех требований Закона «О языках в Республике Беларусь». Кроме того, всегда следует помнить о том, что «полноценность и адекватность спектра лексикографических систем в значительной степени определяют уровень и достоверность извлечения информации и знаний из текстов различного состава, структуры и назначения» [15, с. 38].

Масштабная работа по формированию таких электронных корпусов параллельных текстов в нашей республике ведется на протяжении многих лет. Особенно весомый вклад в теорию и практику создания указанных корпусов параллельных текстов сделан профессором А. В. Зубовым, который, начиная с защиты в 1969 г. кандидатской диссертации на тему «Переработка текста естественного языка в системе «человек–машина», уже много лет занимается компьютерной обработкой текстов. В частности, для рассматриваемой нами темы А. В. Зубов не только дал точные и исчерпывающие

определения многим ключевым понятиям в этой области (см., например, следующее определение: «Корпус параллельных текстов состоит из множества текстов, написанных на каком-либо исходном языке, и текстов-переводов этих исходных текстов на один или несколько других языков» [16, с. 132]), но и раскрыл взаимосвязи, существующие между параллельной текстографией и сравнительной лексикографией: «...корпуса параллельных текстов ... позволяют: 1) автоматически строить двуязычные и многоязычные словари; 2) автоматически создавать и пополнять словари для систем машинного перевода...» [Там же, с. 136].

Таким образом, рассмотрение взаимосвязанного развития двух направлений прикладной лингвистики (сравнительной лексикографии и параллельной текстографии) показало следующее:

- в рамках обоих направлений рассматриваются совокупности текстов: объект сравнительной лексикографии – это тексты дву- и многоязычных словарей разных типов, т.е. нелинейные тексты; объект параллельной текстографии – это параллельные тексты на разных языках, обработанные специальным образом (тэггирование, лемматизированная, морфологическая, синтаксическая и иная разметка [17]), т.е. линейные тексты;

- главной функцией сравнительной лексикографии является обслуживание тех совокупностей линейных разноязычных текстов, которые репрезентируют соответствующие национальные культуры. На наш взгляд, «удельный вес нелинейных текстов (словарей, энциклопедий, справочников и т.п. гипертекстов) в общей совокупности разножанровых текстов, репрезентирующих ту или иную национальную культуру, будет непрерывно увеличиваться, обеспечивая этим самым целостность и информационную доступность данной культуры» [18, с. 19–20];

- перечень основных функций параллельной текстографии намного шире. В их числе: познавательная, трансляционная, учебная, информационная и иные, более специализированные функции. К примеру, представителям сравнительной лексикографии электронные корпуса параллельных текстов помогают: а) находить переводные эквиваленты многоязычных слов [19]; б) вырабатывать алгоритм перевода терминологических словосочетаний [20]; в) автоматически устранять полисемию лексических единиц [16, с. 136]; г) переводить несвободные словосочетания [21], а также решать множество иных словарных задач.

ЛИТЕРАТУРА

1. Ковтун, Л. Языкознание у восточных славян / Л. Ковтун // История лингвистических учений. Позднее Средневековье. – СПб. : Наука, 1991. – С. 182–207.
2. Жураўскі, А. І. Тыпы і асаблівасці глос у “Евангеллі” В. Цяпінскага / А. І. Жураўскі // Пр. Інстытута мовазнаўства АН БССР. – Мінск : Выд-ва АН БССР, 1960. – Вып. VII. – С. 213–226.
3. Верещагин, Е. М. Лингвострановедческая теория слова / Е. М. Верещагин, В. Г. Костомаров. – М. : Русский язык, 1980. – 320 с.
4. Анічэнка, У. В. Лексікаграфічная дзейнасць Скарыны / У. В. Анічэнка // Питання східнослов'янської лексикографії XI–XVII ст. : матер. симп. – Київ : 1979. – С. 58–64.

5. *Щербин, В. К.* Белорусская лексикографическая традиция: история и современность / В. К. Щербин // Славянская лексикография; отв. ред. М. И. Чернышева. – М. : Изд. центр «Азбуковник», 2013. – С. 21–61.
6. *Щербин, В. К.* Функции научных цитат в терминологии и специальной концептографии / В. К. Щербин // Термінологічний вісник. – 2013. – № 2. – С. 162–170.
7. Тематический русско-белорусский словарь языка научных отчетов / Физ.-мат. науки; сост. В. К. Щербин. – Минск : Ротапринт УД АН Беларуси, 1992. – Вып. 1. – 248 с.
8. Тематический русско-белорусский словарь языка научных отчетов / Физ.-техн. науки; сост. В. К. Щербин. – Минск : Ротапринт УД АН Беларуси, 1993. – Вып. 2. – 293 с.
9. *Добролюбов, А. И.* Словарь типа INSITU английского языка / А. И. Добролюбов // Словарь, грамматический справочник и сборник текстов для изучающих английский язык. – Минск : Амалфея, 1999. – 367 с.
10. Патент Российской Федерации № 2054200 на изобретение «Метод «INSITU» для создания словарей с целью перевода на иностранные языки». – М., 1996.
11. *Кимягарова, Р. С.* Словарь языка басен И. А. Крылова / Р. С. Кимягарова // Крылов И. А. Избранные басни. – М. : Оникс, 2006. – 928 с.
12. Дыялектная мова Гродзеншчыны: хрэстаматыя; аўт.-склад.: М. А. Даніловіч, Н. К. Памецька. – Гродна : ГрДУ, 2007. – 260 с.
13. *Щербин, В. К.* Словарные истоки практики нелинейного письма / В. К. Щербин // Весн. Бел. дзярж. ун-та. Сер. 4. – 2003. – № 2. – С. 28–33.
14. *Зубов, А. В.* О возможностях использования параллельных учебных текстов в лексикографии / А. В. Зубов // Новое в теории и практике лексикографии: синхронный и диахронный подходы : матер. VIII Междунар. shk.-семина., Иваново, 10–12 сент. 2009 г. – Иваново : Иван. гос. ун-т, 2009. – С. 44–45.
15. *Беляева, Л. Н.* Корпус научных текстов как источник извлечения лексикографической информации / Л. Н. Беляева // Новое в теории и практике лексикографии: синхронный и диахронный подходы : матер. VIII Междунар. shk.-семина., Иваново, 10–12 сент. 2009 г. – Иваново : Иван. гос. ун-т, 2009. – С. 38–41.
16. *Зубов, А. В.* Корпусная лингвистика и ее возможности для проведения лингвистических исследований / А. В. Зубов // Сучасныя праблемы беларускай лексікалогіі і лексікаграфіі : матэр. Міжнар. навук. канф., Мінск, 19–20 ліст. 2005 г. – Мінск : Права і эканоміка, 2006. – С. 129–140.
17. *Зубов, А. В.* О целях издания и принципах организации параллельного тэгірованнаго русско-белорусского корпуса учебных текстов / А. В. Зубов // Язык и дискурс в статике и динамике. – Минск : МГЛУ, 2008. – С. 37–38.
18. *Щербин, В. К.* Семиотические особенности словарно-справочных текстов / В. К. Щербин // Актуальные проблемы теоретической и прикладной лингвистики : матер. Междунар. науч. конф., посвящ. памяти проф. Р. Г. Пиотровского, Минск, 15–16 июня 2010 г. : в 2 ч.; отв. ред. А. В. Зубов. – Минск : МГЛУ, 2010. – Ч. 2. – С. 16–21.
19. *Зубов, А. В.* Использование параллельных текстов для определения переводных эквивалентов многозначных слов / А. В. Зубов // Матер. Междунар. конф. «Корпусная лингвистика и лингвистические базы данных». – СПб., 2002.
20. *Зубов, А. В.* Алгоритм перевода терминологических словосочетаний с использованием параллельных текстов / А. В. Зубов // Теория и практика перевода. – 2005. – № 1. – С. 64–67.
21. *Зубов, А. В.* Использование параллельного корпуса текстов для перевода несвободных словосочетаний / А. В. Зубов // Вестн. Моск. гос. обл. ун-та. Сер. Лингвистика. – 2011. – № 6. – С. 79–83.

КОРПУСНЫЕ ИССЛЕДОВАНИЯ

А. К. Жубанов

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ СЕМАНТИЧЕСКОЙ РАЗМЕТКИ ЗНАМЕНАТЕЛЬНЫХ КЛАССОВ СЛОВ В НАЦИОНАЛЬНОМ КОРПУСЕ КАЗАХСКОГО ЯЗЫКА

При семантической классификации лексики казахского языка имеются свои трудности и своя специфика. Это одна из труднейших задач семантики, которая не находит единого решения и в других языках мира. Количество таких классов и подклассов, их состав зависят от целей исследования, лингвистической школы и целого ряда других факторов.

Работ, посвященных исследованию казахского языка в области семантики, относительно мало. Здесь стоит назвать имя М. Оразова, его монографию «Қазақ, тілінің семантикасы» и его исследования по семантике казахского глагола [1; 2], которые затрагивают многие вопросы семантики казахского языка. Имеется и ряд других работ по некоторым проблемным вопросам семантики казахского языка [3]. Хотелось только отметить, что результаты этих работ во многом помогли нам при решении поставленных перед нами задач по вопросам формализации содержания казахского текста [4; 5]. Здесь необходимо отметить, что данная тема – «Основные принципы формализации содержания казахского текста» мною была защищена 2002 году в виде докторской диссертации, научным консультантом которой является Александр Васильевич Зубов, а текст диссертации опубликован в виде монографии» [5].

В самом общем виде можно выделить три принципа семантической классификации лексики казахского языка: 1) денотативно-десигнативный; 2) парадигматический; 3) синтагматический [1; 6; 7].

Денотативно-десигнативный принцип учитывает онтологическое расчленение действительности. При этом слова объединяются в группы в соответствии с тем, что они обозначают в реальной действительности (названия животных, названия растений, наименования определенных классов предметов и т.п.). При парадигматической классификации слова объединяются в группы на основе тождества определенных семантических компонентов, выделенных в значении слов (модальные и каузативные глаголы и т.п.). По синтагматическому принципу слова объединяются в лексико-семантические группы с учетом количества и качества валентностей классифицируемых слов [1, с. 189–192; 6, с. 39–42].

Исходя из целей исследования, за основу нашей семантической классификации принят денотативно-десигнативный принцип, хотя в некоторых случаях он будет дополняться двумя другими перечисленными выше принципами.

Далее, по семантической классификации знаменательных классов слов казахского языка, мы решили остановиться только на следующих двух:

1) семантическая классификация имен существительных (зат есім) казахского языка; 2) семантическая классификация глаголов (етістік) казахского языка.

А семантические классификации остальных видов знаменательных слов казахского языка можно будет прочесть в монографии [5].

Семантическая классификация имен существительных (зат есім) казахского языка. Как известно, основным критерием разграничения имен существительных (наименований явлений, предметов, процессов, понятий) казахского языка, как и в других языках, на отдельные типы является, прежде всего, анализ их прямого номинативного значения, детерминированного типом денотата и сигнификата. При этом «слово» можно рассматривать двояко, т.е. как единицу первичного означивания в системе номинативных средств языка и как единицу, семантически реализованную в речи. Нами различаются две модификации «денотатов»: денотат как объект обозначения единичного предмета или класса предметов, а также денотат как предмет референции знака в речи.

А. А. Уфимцева в своей книге «Типы словесных знаков» [8, с. 102] подразделяет сигнификат как компонент знакового значения слов на шесть понятий, которые, как нам кажется, справедливы и для казахского языка. Имена с денотативным и денотативно-сигнификативным значением относятся к разряду имен, называемых конкретной лексикой, а слова с сигнификативной и сигнификативно-денотативной основой значения составляют абстрактную лексику любого языка, в том числе и казахского. Необходимо заметить, что определяющим для семантики именных лексем является обязательное наличие в их знаковом значении денотата и сигнификата. Если словесный знак служит специализированным названием (термином) в какой-либо узкой предметной области, то доминирует денотат над сигнификатом. В случае, когда словесный знак служит названием абстрактного понятия, то, наоборот, или сигнификат доминирует над денотатом, или словесный знак имеет чисто сигнификативное значение.

Именами существительными в казахском языке, так же как и во многих тюркских языках, являются различные наименования предметов и понятий в широком смысле слова. Это различные имена: собственные, например имена людей, географические названия и т.д.; нарицательные – названия человека: *кісі* (человек), *шеше* (мать), *әйел* (женщина); названия животных; названия растений; названия различных веществ и предметов быта; научные, технические и другие термины; наименования разных обобщенных отвлеченных понятий: *ел* (народ), *ру* (род), *тағдыр* (судьба) и т.д.; наименования действий, явлений, состояний: *жүріс* (ходьба), *хабар* (известие), *ұйқы* (сон) и т.д.; названия результатов действия: *жиналыс* (собрание; ср. *жинау* – собираться), *кеспе* (латиша; ср. *кесу* – резать); названия орудий (инструментов) действия: *бұрғы* (сверло; ср. *бұрау* – винтить, завинчивать), *сытырғыш* (метла, веник), *ара* (пила) и т.п.

Как пишет М. Оразов, среди слов казахских существительных самыми частыми по сложности их значений являются нарицательные имена. Эти имена существительные по отношению к значению имеют в языке измен-

чивый характер, т.е. иногда у них превалирует денотативное значение, иногда сигнификативное или же приобретает как бы суммарное значение – и денотативное, и сигнификативное. Примером может служить использование слов в метафорическом и метонимическом значениях. Безусловно, когда слова употребляются в переносном значении, то изменяется и их денотативное и сигнификативное значение. Для убедительности сказанного можно сравнить слово «нар» в двух следующих предложениях: *Қырық балам бар еді, қырқы да бірдей нар еді.* (Батырлар жыры). *Айырдан туган жампоз бар – Нарға жүгін салғызбас* (Махамбет) [1, 98]. В первом предложении слово «нар» применено в переносном значении как сильный, «боевой», а во втором – в прямом значении как одnogорбный верблюд. Их денотативные и сигнификативные значения несравнимы.

В количественном отношении тематические группы имен существительных в любом языке, в том числе и в казахском языке довольно велики. М. Оразов для примера приводит 8 разновидностей таких групп: 1) родственные термины (туысқандық, атаулар); 2) человек и названия частей его тела (дене мүше атаулары); 3) названия домашних животных (төрт түлік мал атаулары); 4) названия одежды, обуви человека (киім-кешек атаулары); 5) название птиц (құс атаулары); 6) название диких животных (жабайы аң атаулары); 7) название обиходных предметов-сосудов (посуды) (ыдыс-аяқ атаулары); 8) названия болезни человека (адамға тән ауру атаулары).

В нашем исследовании в основу семантической классификации казахских существительных положено деление реальной действительности на определенные зоны, сегменты [7]. С учетом целей нашего исследования, анализируемого материала, а также отдельных идей, изложенных в работах [1; 3; 9] нами выделены семантические классы существительных, используя, денотативно-десигнативный принцип при выделении наиболее крупных семантических классов существительных. Более мелкие лексико-семантические группы, например «части тела человека», «воинские звания», «дни недели» и др., связаны парадигматическими отношениями.

Семантическая классификация глаголов казахского языка. Знаковое значение глагольных лексем. Имена существительные отражают и обозначают предметы или явления, а глаголы – отношения последних. Значение имен существительных – более автономно и абсолютно. Значение глагольных лексем – относительно и детерминировано многими факторами. В глагольных лексемах, выражающих понятие отношения, их взаимодействие как бы выносится за рамки глагольной лексемы и переносится в область синтагматических отношений действия к его объекту или субъекту (или одновременно). Поэтому соотносительность глагольных лексем с предметным рядом предполагает наличие смысловых связей глагола с предметными именами.

Знаковое значение глагольных лексем представляет собой сложный феномен, детерминируемый тремя факторами: 1) соотносительностью с реальной действительностью, которая представляет собой в данном случае не мир предметов, а мир их отношений, действий и состояний; 2) категориальной семантикой сочетающихся с глаголом предметных имен; 3) типом смысло-

вых отношений между действием, его субъектом и объектом, а соответственно типом моделей «объект – действие», «действие – объект», «субъект – действие – объект» [8, с. 113].

По сравнению с именами существительными класс глагольных лексем представляет большие трудности при выделении типов по характеру их знакового значения.

Во-первых, собственно лексическое (индивидуальное) значение глагольных слов осложнено целым рядом семантических признаков, выражающих способ протекания действия, обстоятельства его свершения, видовременные характеристики глагольного действия, модальные и экспрессивные наслоения.

Во-вторых, определение глагольной семантики затруднено уточнением, какой из признаков присущ только данной лексеме, какой – целой группировке, а какой обобщен до уровня категории глагольных лексем и т.п.

Традиционно глаголы по их семантике делятся на три класса: 1) глаголы действия; 2) глаголы состояния; 3) глаголы процессов.

У каждой лексико-семантической группы – свои парадигматические параметры, семантические признаки.

Для семантики эмотивных глаголов существует иная шкала – глаголы воздействия, глаголы восприятия, глаголы оценки или субъективного отношения и т.п. [8, с. 148].

М. Оразов в своей книге «Қазақ тілінің семантикасы» пишет, что неконкретность лексического значения глагола в основном зависит от его денотативного и сигнификативного значения. Когда речь идет о денотативном значении глагола, мы говорим о движении предмета, физическом действии (амал-эрекет) и о состоянии (қалпы). Если учесть, что движение происходит в какое-то определенное время и в определенном месте, то имеется в виду присутствие субъекта, который осуществляет физическое действие, известна также цель передвижения. При этом лексическое значение глагола дополняют множество других элементов. В качестве примера он приводит следующее предложение: *Шаңғышылар селоза кіргенде көз байланып та қалған еді. (М. Сундетов). /Когда лыжники вошли в село, уже наступило сумерки/*. В этом предложении рассматривается структура значения глагола *кір* (входи): 1) наличие движения; 2) движется некоторый предмет или человек (субъект); 3) движение происходит на земле; 4) переход через некоторую границу (вошли на территорию села); 5) имеется элемент значения войти «внутрь». Перечисленные значения определяются благодаря отношениям и связям слов в данном предложении. Таким образом, действие *кіру* (входить) происходит на лыжне (на снегу) движением лыжников, *кіру* означает вхождение на территорию, где проживают люди (в село). Благодаря сочетанию слова *кір* с другими словами (*село, шаңғышылар*), мы узнаем направление движения, что оно происходит «извне в село».

Глагол *кір*, сочетаясь с другими словами, может принять абсолютно другое значение. Например, *Мен 1962 жылы партия ұйымына кірдім*. В этом предложении слово *кір* уже не означает движения, а дает понять, что он стал членом партийной организации. Последний пример позволяет говорить о том, что значение слова непосредственно не зависит от значения слово-

сочетаний (от контекста, в широком смысле) в предложении. В первом примере глагол *kір* сочетается с объектом, означающим местность, поэтому его граница является основой понимания лексического значения из сем движения. Во втором примере глагол *kір* сочетается с понятием, т.е. названием организации, здесь сема глагола *kір* единственная – стать членом данной организации. Изменение значения некоторых глаголов требует очень тонкого подхода при анализе их семантики. Например, глаголы движения в одних сочетаниях входят в структуру глагола общения, а в других – в структуру глагола мышления, Сравните: *үйге келу* (прийти домой), *қарсы келу* (возразить), *ойға келу* (вспомнить), *тілге келу* (заговорить) и др. В структуре предложения глаголы образуют отношения с определенными словами. Такие отношения происходят между субъектом или объектом и движением, между физическим действием (*амал-әрекет*) и объектом или субъектом. Когда в предложении глагол требует связи с другим словом, то он с этим словом имеет связь в отношении и значения, и формы (валентность глагола).

Тематические и лексико-семантические классы казахских глаголов.

Основанием для отнесения слов в различные тематические и лексико-семантические классы являются обобщающие (жалпылаушы) семы. Необходимо отметить, что в индо-европейских языках имеется определенный опыт в разделении на лексико-семантические классы отдельных частей речи, но в отношении тюркских языков, в частности казахского, подобные исследования начали появляться только в последние годы (большинство из них имеют сопоставительный с русским языком характер). В настоящее время описаны следующие лексико-семантические группы глаголов казахского языка: глаголов физиологического воздействия, глаголов речи, глаголов ориентированного положения в пространстве, глаголов движения, эмотивные глаголы, глаголы физического восприятия, глаголы созидания и др.

Таким образом, мы своей семантической классификации знаменательных классов слов казахского языка старались придерживаться результатов исследований М. Оразова, а именно основных условий объединения слов в одну лексико-семантическую группу, а также и других работ вышеупомянутых ученых.

В центре лингвистической проблематики находится изучение фундаментальных семантических категорий со сложной формально-содержательной структурой. К числу таких категории, отражающих одну из форм всеобщей связи явлений и фактов объективной действительности, следует отнести категорию каузальности (причинности).

Необходимо отметить, что при семантической классификации каузативных глаголов казахского языка мы в своей работе в основном придерживались положений работы Г. М. Алиной «Каузальные отношения в системе русского и казахского языков» (рукопись канд. дис., руков. З. К. Ахметжанова).

Необходимо отметить, что бытийные глаголы казахского языка еще не освоены до конца. В существующих классификациях (И. Маманова, А. Ыскакова, Н. Оралбаевой, М. Оразова, Н. Душимовой и др.) лексико-семантических групп казахского языка бытийные глаголы не выделяются как отдельная группа. Глагол «быть» является основным средством выражения понятия бытия в русском языке. В казахском языке, глаголы с чисто бытий-

ным значением, такие, как *еді (екен, емес), бол*, характеризуются в основном как вспомогательные, которые самостоятельно не употребляются, но образуют описательные формы времен и наклонений основных глаголов.

А. Ыскаков, в свою очередь, определяет глаголы, образованные от корня *-е (еді, екен, емес)*, как неполные слова, которые утратили свое лексическое значение либо полностью, либо наполовину и играют в предложении вспомогательную роль, придавая грамматическое значение отдельным словам (*бала еді, қырық-отыздар шамасында екен, қара-торы ала емес*).

По результатам нашей классификации семантический класс казахских глаголов может рассматриваться по 10 разновидностям: 1. Глаголы существования; 2. Глаголы физических действий; 3. Глаголы движения; 4. Глаголы, означающие отношения между людьми, животными и неживыми объектами; 5. Глаголы психической деятельности; 6. Глаголы физического состояния и нормы; 7. Глаголы психического состояния; 8. Фазисные глаголы; 9. Модальные глаголы; 10. Каузативные глаголы. Каждый класс глагола может быть разделен еще и на семантические подклассы.

В заключении хотелось бы отметить, что материалы исследования по семантической классификации знаменательных слов казахского языка в данное время используются для осуществления семантической разметки текстов Национального корпуса казахского языка.

ЛИТЕРАТУРА

1. Оразов, М. Қазақ тілінің семантикасы / М. Оразов. – Алматы, 1991. – 213 б.
2. Оразов, М. Семантика казахского глагола: дис. докт. филол. наук / М. Оразов. – Алма-Ата, 1983.
3. Ахметжанова, З. К. Функционально-семантические поля русского и казахского языков / З. К. Ахметжанова. – Алма-Ата, 1989. – 106 с.
4. Джубанов, А. Х. Формализация процедуры построения текстов естественного языка / А. Х. Джубанов, А. В. Зубов // Научные приборы и автоматизация научных исследований. – Алма-Ата : Ғылым, 1992. – С. 190–196.
5. Жубанов, А. К. Основные принципы формализации содержания казахского текста / А. К. Жубанов. – Алматы, 2002. – 250 с.
6. Васильев, Л. М. Семантика русского глагола / Л. М. Васильев. – М., 1981. – 184 с.
7. Зубов, А. В. Вероятностно-алгоритмическая модель порождения текста : автореф. дис. ... д-ра филол. наук / А. В. Зубов. – Минск, 1985. – 39с.
8. Уфимцева, А. Типы словесных знаков / А. Уфимцева. – М., 1974.
9. Апресян, Ю. Д. Экспериментальное исследование семантики русского глагола / Ю. Д. Апресян. – М., 1967. – 252 с.

О. М. Зубань

ЗАДАЧИ И МЕТОДЫ АВТОМАТИЧЕСКОГО МОРФЕМНОГО АНАЛИЗА В КОРПУСЕ УКРАИНСКОГО ЯЗЫКА

Корпус украинского языка [1] (в оригинале Корпус української мови – КУМ) имеет исследовательский характер и по своему назначению ориентирован на решение широкого круга лингвистических задач, в частности

в области морфемики и словообразования, хотя создатели большинства корпусов славянских языков не ставят задачу параметризации текстов на морфемном уровне: по нашим сведениям, кроме Корпуса украинского языка, только Национальный корпус русского языка [2] имеет поиск по словообразовательным параметрам, но без аннотирования посткорневой зоны. Возникает ряд вопросов: нужна ли в корпусе морфемная (словообразовательная) аннотация текста? Какие исследовательские перспективы открывает перед лингвистом корпусноориентированный автоматический морфемный и словообразовательный анализ? Как методически правильно создать такой анализ?

Работая над проблемами конструирования параметризованных лингвистических баз, коллектив лаборатории компьютерной лингвистики Киевского национального университета имени Тараса Шевченко разработал методику компьютерного моделирования морфемной структуры украинского слова и в 2001 г. создал автоматизированную систему морфемно-словообразовательного анализа (АСМСА) украинского языка [3; 4]. АСМСА – лексикографическая система (база данных), выполняющая различные задачи автоматической систематизации лексики (исходных форм) в области морфемики и словообразования: группирует лексику в общекорневые и общеаффиксальные классы; классифицирует слова общекорневой выборки по словообразовательным тактам; классифицирует лексику по структурно-морфным моделям слов; создает алфавитно-частотные корневые и аффиксальные словари.

Система АСМСА создавалась на базе ≈ 170 тыс. слов, разбитых на морфемы по принципам морфемного членения словаря И.Т. Яценко [5]. Морфемный анализ проводился автоматизировано с помощью удобного компьютерного инструмента, в котором каждое слово анализировалось в отдельной электронной карточке (рис. 1). По методике моделирования структурно-функциональных связей морфов в слове каждому слову приписывалась модель, определяющая границы и функциональный тип морфа в морфемной структуре слова.

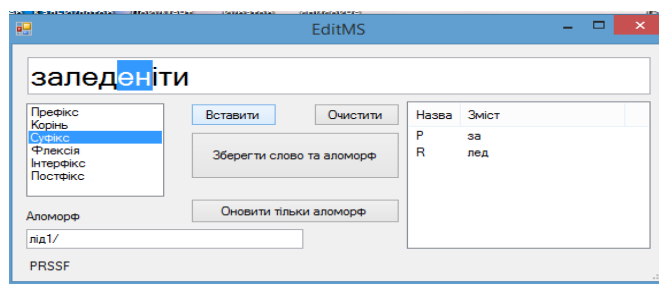


Рис. 1. Электронная карточка морфемного анализа слова *заледеніти*

Маркировка морфем в электронной карточке слова автоматически приписывает каждому слову формулу программной процедуры морфемной сегментации: *заледеніти* P2R5S7S8F10, где буквы обозначают функциональный тип морфа (P – приставка, R – корень, S – суффикс, F – флексия, I – интерфикс, X – постфикс), а цифры – границы морфа по порядковому номеру (с начала слова) конечной графемы каждого морфа (за – 2; лед – 5; ен – 7; і – 8; ти – 10). Таким образом, формула программной процедуры морфемной сегментации отображает функциональную структуру слова на морфемном

уровне – PRSSF, а субстанциональное выражение этой структуры представляется графемно-цифровой моделью, автоматически переводимой в базе данных на латиницу по номеру буквы в алфавите. Информация о слове с программной процедурой раньше записывалась в res.ini-файлы (рис. 2), объединяемые в единую базу данных. Сегодня работа с базой данных происходит в on-line режиме.

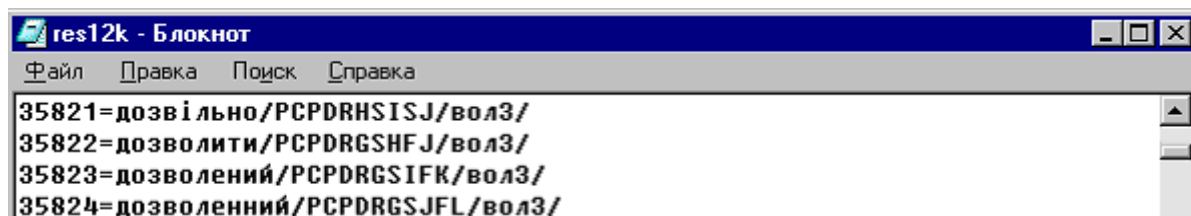


Рис. 2. Фрагмент res.ini-файла

С целью формирования общекорневых выборок в процессе создания морфемной базы данных (МБД) учитывалась омонимия и алломорфия корневых морфов: 1) омонимия снималась приписыванием каждому корню-омониму цифрового кода: в существительном *лід-ер-Ø* приписывается корень *лід2*, а в глаголе *за-лед-ен-і-ти* – корень *лід1*; 2) алломорфным корням приписывался условно первичный алломорф – маркер инварианта корневой морфемы: *за-лед-еніти* – *лід1*, *льод-овий* – *лід1*.

Сегодня система АСМСА объединяет три базы данных: 1) морфемных структур слов (исходных форм) ≈ 200 тыс.; 2) базу корней-алломорфов ≈ 2500 ; 3) базу корней-омонимов ≈ 3100 корней.

Система АСМСА с начала своего создания ориентировалась на проведение автоматической морфемной сегментации текстовых словоупотреблений. Использование АСМСА в процедуре автоматической морфемной сегментации в Параметризированной базе данных поэтической речи, созданной коллективом лаборатории компьютерной лингвистики в 2004 г., демонстрирует опыт автоматической морфемной аннотации текста.

id	cls	morfem	morfema	comm
378422	ПТ	R	за	за RC RC
378423	ЙИ	R	щит	щитом RDFE RDFF
378424	ЙИ	F	ом	щитом RDFE RDFF
378425	АИ	R	смарагд	смарагдових RHSJFL RHSJFL
378426	АИ	S	ов	смарагдових RHSJFL RHSJFL
378427	АИ	F	их	смарагдових RHSJFL RHSJFL
378428	ЙИ	R	ліс	лісів RDFE RDFF
378429	ЙИ	F	ів	лісів RDFE RDFF
378430	X	N	.	

Рис. 3. Фрагмент морфемной сегментации словоупотреблений поэтических текстов Л. Костенко

На рис. 3 показано морфемную сегментацию словоупотреблений текстового фрагмента *...за щитом смарагдових лісів*. Каждое словоупотребление сегментируется на морфы с автоматическим приписыванием двух программных процедур: первая – морфемная модель начальной словоформы лексемы; вторая – морфемная модель словоформы текстового фрагмента.

В 2010 г. началось создание Корпуса украинского языка, составляющего на сегодня корпус текстов, параметризированных на уровне морфологической аннотации, объемом ~ 44 млн. словоупотреблений. В процессе создания автоматического морфемного анализа в Корпусе с целью оптимизации поиска на больших текстовых массивах мы отказались от метода морфемной аннотации текста. Морфемная разметка текстовых словоупотреблений не производится, тексты Корпуса используются только как источник для составления разного рода алфавитно-частотных морфемных словарей по текстовым выборкам с помощью системы АСМСА, исполняющей функцию морфемного модуля-анализатора в Корпусе украинского языка.

На этапе работы морфологического модуля все словоупотребления текста с морфологической разметкой лемматизируются в исходные формы и попадают в морфемный модуль в виде алфавитно-частотного словаря исходных форм с информацией о части речи, частотных характеристиках, контекстах употребления и морфемной модели слова, автоматически приписываемой исходной форме способом сопоставления с этим же словом в АСМСА (рис. 4).

Код	wrд	cls	xabs	xsred	sigma	v	R	context	d	morfem
396	вгасати	Г	2	0,0327868852	0,1780783687	5,4313902456	2	Тліє, не вгаса	2,988105	PRSF
397	вголос	Н	1	0,0163934426	0,1269830605	7,7459666924	1	А то думають	1,110223Е-15	PR
398	вгородити	Г	1	0,0163934426	0,1269830605	7,7459666924	1	Вихватила ту с	1,110223Е-15	PRSF
399	вгорю	Н	4	0,0655737705	0,5079322421	7,7459666924	1	Все вгорю та ви	1,110223Е-15	PRS
400	вдавити	Г	1	0,0163934426	0,1269830605	7,7459666924	1	Чому мене не	1,110223Е-15	PRSF
401	вдарити	Г	2	0,0327868852	0,1780783687	5,4313902456	2	Налигачем скр	2,988105	RSF

Рис. 4. Фрагмент морфемной базы данных исходных словоформ
вгасати, вголос, вгородити, вдавити, вдарити

На выходе морфемного модуля – те же исходные формы слов, сегментированные на морфы, с информацией о функциональном типе морфемы (рис. 5).

морфемид	wrдFK	mname	mseg
799	396	P	в
800	396	R	гас
801	396	S	а
802	396	F	ти
803	397	P	в
804	397	R	голос
805	398	P	в
806	398	R	город
807	398	S	и
808	398	F	ти
809	399	P	в
810	399	R	гор
811	399	S	у
812	400	P	в
813	400	R	дав
814	400	S	и
815	400	F	ти
816	401	R	вдар
817	401	S	и
818	401	F	ти

Рис. 5. Фрагмент базы данных морфемной сегментации
исходных словоформ *вгасати, вголос, вгородити, вдавити, вдарити*

На основании этих баз по заказу пользователя создаются разного рода алфавитно-частотные словари морфем и морфемных структур слов по текстовым выборкам отдельных авторов или подкорпусов стилей. Эти словари выложены в рубрике "Частотні словники" Корпуса как автономные электронные лексикографические системы [6]. Рассмотрим автоматическое составление морфемных словарей на базе ≈ 80 тыс. словоупотреблений поэтических текстов Т. Шевченко [7].

Инвентарь единиц частотного словаря морфем составляется по двум параметрам – тип морфемы и часть речи, выбираемые пользователем перед построением словаря, в данном случае – корни имен существительных. Частотный словарь корней существительных по текстам Т. Шевченко (рис. 6) состоит из 1509 морфем, каждой из которых приписана информация об абсолютной и средней частоте. Словник строится по алфавиту или частотности: спаду / росту абсолютной частоты.

Морфемно-частотний словник Коренів всього записів: 1509					Частотний словник по морфемі: R:люд, частина мови: Іменник Всього записів: 7		
Морфема	Структури	Слова	Абсолютна частота	Середня частота	Слово	Частина мови	Абсолютна частота
бу	Структури	Слова	531	8,70	люд	ім. ч. р.	12
пі	Структури	Слова	380	6,23	люди	ім. множ.	270
ста	Структури	Слова	340	5,57	людина	ім. ж. р.	1
він	Структури	Слова	335	5,49	людоїд	ім. ч. р.	3
бог	Структури	Слова	306	5,02	людомор	ім. ч. р.	1
люд	Структури	Слова	306	5,02	недолюд	ім. ч. р.	3
світ	Структури	Слова	287	4,70	недолюдок	ім. ч. р.	2
да	Структури	Слова	280	4,59	1		
каз	Структури	Слова	280	4,59			
див	Структури	Слова	273	4,48			
1 2 3 4 5 6 7 8 9 10 ... >>							
Всього записів: 5					Частотний словник по морфструктурі: RIRSF з морфемою -люд- Всього записів: 2		
Структура	Абсолютна частота				Слово	Частина мови	Абсолютна частота
PRF	3				людоїд	ім. ч. р.	3
PRSF	2				людомор	ім. ч. р.	1
RF	282				1		
RIRSF	4						
RSF	1						

Рис. 6. Фрагмент частотного словаря корней (инвентарь и лексическая реализация)

Активация опции «Слова» конкретного корня открывает окно лексической реализации этой морфемы (Частотный словник по морфемі): фрагмент иллюстрирует реализацию частотного корня *-люд-* (абсолютная частота – 306 словоупотреблений) в семи словах разной продуктивности, самым частотным из которых является слово *люди* (270 текстовых словоупотреблений). Активация опции "Структура" этого корня открывает список моделей морфемных структур (нижняя первая таблица), в которых реализуется корень *-люд-*, а активация конкретной модели морфструктуры открывает окно лексической реализации выбранной структуры (Частотный словник по морфструктурі): фрагмент демонстрирует реализацию модели RIRSF с корнем *-люд-* в двух словах: *люд/о/їд/Ø/Ø*, *люд/о/мор/Ø/Ø*.

Частотний словник збірки "Твори в п'яти томах". Тарас Шевченко	
Конкорданс до слова: люди (ім. множ.)	
Морфемна структура: RF /люд/и/	
Контекст	Джерело
Пошли ж ти їй долю — вона молоденька, Бо люде чужії її засміють.	>>
На чужині не ті люде — Тяжко з ними жити !	>>
Кохайтесь, чорнобриві, Та не з москалями, Бо москалі — чужі люде, Роблять лихо з вами.	>>
Серце в'яне співаючи, Коли знає, за що ; Люде серця не побачать, А скажуть — ледащо !	>>
Кохайтесь ж, чорнобриві, Та не з москалями, Бо москалі — чужі люде, Згнущаються вами.	>>
Нехай собі ті люде, Що хотять, говорять : Вона любить, то й не чує, Що вкралося горе.	>>
Тойді Катерина Буде собі московкою, Забудеться горе ; А поки що, нехай люде, Що хотять, говорять.	>>
Батько . мати — чужі люде . Тяжко з ними жити !	>>
:ПРИЧИННА ("Рече та стогне Дніпр широкий..."):ТВОРИ:ТАРАС ШЕВЧЕНКО:ПОЕТИЧНА	
МОВА	
Стиль:поетичні тексти	
Приблизна кількість слововживань: 849	
Видано:Київ: Дніпро:	
Жанр: вірш	
Місце: Україна / Росія: XIX ст.	

Рис. 7. Фрагмент конкорданса лексеми *люди*

Активация выбранного слова в "Частотном словнике по морфемі" (верхняя правая таблица) открывает окно контекстов употребления слова (рис. 7). Каждый текстовый фрагмент опцией "Джерело" связан с экстралингвистической информацией о произведении в Корпусе: на рис. 7 демонстрируется информация о произведении по первой цитате конкорданса.

Автоматический частотный словарь морфемных структур слов имеет такую же организацию: частотный словарь морфструктур, реализация конкретной морфструктуры в лексиконе выборки, реализация слова выбранной морфструктуры в текстах. Рис. 8. иллюстрирует реализацию высокочастотной морфструктуры PRSF (3804 текстовых словоупотреблений) в 95 существительных по текстам Т. Шевченко.

Всього записів: 39			Частотний словник по морфструктурі: PRSF, частина мови: Іменник Всього записів: 95 Всього записів: 95								
Структура	Абсолютна частота	Середня частота	Слово	Частина мови	Абсолютна частота	Слово	Середня частота	Середньоквадратичне відхилення	Коефіцієнт стабільності		
RF	16323	267,59	пожар	ім. ч. р.	20	11	0,327868852459016	0,740432117418768	2,25831795812724		
RSF	7482	122,66		невольник	ім. ч. р.	15	11	0,245901639344262	0,644160088668475	2,61958436058513	
R	6605	108,28			пророк	ім. ч. р.	12	7	0,19672131147541	0,697058741818665	3,54338193757822
PRSF	3804	62,36	пророк	ім. ч. р.							
PRF	1631	26,74		пророк		ім. ч. р.					
RSSF	898	14,72				пророк	ім. ч. р.				
RS	604	9,90	пророк	ім. ч. р.							

Рис. 8. Инвентарь морфструктур и лексическая реализация морфструктуры PRSF

Использование автоматического морфемного анализа в Корпусе украинского языка раскрывает новые возможности исследования морфемной организации украинского слова в словаре и тексте. Полученные статистические данные могут использоваться в системных стилистических исследованиях [8]: сравнение частотных характеристик морфемных единиц в разных текстовых выборках Корпуса и в системе языка прогнозирует вероятность реализации морфемных единиц в разных текстах и формирует понятие статистической структуры текста на морфемном уровне как его стилеметрической модели. Систематизация морфструктур позволяет анализировать включение морфемных моделей в образование новых слов Корпуса, исследовать морфемную длину и глубину слов в украинских текстах разных стилей, а также морфотактику разных типов морфем. Связь частотных морфемных словарей с конкордансом позволяет проанализировать разноаспектное функционирование морфем и морфструктур в предложениях.

Опыт создания морфемного автоматического анализа показывает, что с целью извлечения из текста реляционно-функциональных характеристик морфемных единиц не обязательно проводить морфемную или словообразовательную разметку текстов. Методика морфемной аннотации словаря лексем (исходных форм) текстов не уменьшает эффективности и оперативности организации лингвистического исследования, а наоборот, за счет систематизации и разноаспектной классификации морфемной информации повышает разрешающую возможность исследования. Изучение организации текста на морфемном уровне посредством морфемного анализа исходных

форм слов, а не словоупотреблений, оправдан практикой создания частотных морфемных словарей в Корпусе украинского языка. Предложенный подход базируется на онтологическом устройстве морфемной структуры слов флективных языков: морфемная структура основы слова, отображающая эксплицитированную семантику слова, при словоизменении остается относительно стабильной: количество морфем не меняется, может возникать только словоизменительная алломорфия, но она учитывается при лемматизации и может автоматически приписываться морфемам как потенциальный признак. Использование методики морфемной аннотации словаря исходных форм в Корпусе украинского языка демонстрирует эффективность и оптимальность этой методики: ≈ 200 тыс. единиц морфемной базы данных АСМСА позволяют получить информацию о морфемной организации ≈ 44 млн. текстовых словоупотреблений с иллюстрацией контекстов их употреблений. Недостаток этой методики – невозможность автоматически снять омонимию слов одной части речи в случае разной морфемной сегментации омографов: *вида-ти (напечатать)*, *ви-да-ти (дать)*, *вид-а-ти (видеть)*. В таких случаях морфемная сегментация редактируется вручную.

Перед разработчиками АСМСА на данном этапе стоит задача автоматического построения на базе реестра АСМСА и лексики текстов Корпуса общекорневых выборок и словообразовательных гнезд с учетом омонимии и алломорфии корней; общеаффиксальных выборок с учетом омонимии и алломорфии аффиксов. В методологии компьютерного моделирования, использованной в построении АСМСА, обобщены теоретические и прикладные идеи современного языкознания, что делает эту систему эффективным и рациональным инструментом лингвистических исследований.

ЛИТЕРАТУРА

1. Корпус украинского языка [Электронный ресурс]. – Режим доступа : <http://www.mova.info/corpus.aspx>. – Дата доступа : 18.04.2016.
2. Национальный корпус русского языка [Электронный ресурс]. – Режим доступа : <http://www.ruscorpora.ru>. – Дата доступа : 18.04.2016.
3. Дарчук, Н. П. Комп'ютерне анутовання тексту: результати і перспективи : монографія / Н. П. Дарчук. – Київ, 2013.
4. Алексієнко, Л. А. Методика створення автоматизованої системи морфемно-слово-твірної аналізу (АСМСА) слів української мови / Л. А. Алексієнко, Н. П. Дарчук, О. М. Зубань // Наукова спадщина професора С. В. Семчинського : зб. наук. пр. – Київ, 2001, – Ч. 1. – С. 38–49.
5. Яценко, І. Т. Морфемний аналіз: словник-довідник / І. Т. Яценко. – Київ, 1980–1981. – Т. 1–2.
6. Частотні словники Корпусу [Электронный ресурс]. – Режим доступа : <http://www.mova.info/article.aspx?l1=210&DID=5215>. – Дата доступа : 18.04.2016.
7. Електронний словник мови Т. Шевченка [Электронный ресурс]. – Режим доступа : <http://www.mova.info/cfqsh.aspx>. – Дата доступа : 18.04.2016.
8. Зубань, О. Особливості морфемної будови слів у поетичних текстах Т. Шевченка (на матеріалі Корпусу української мови) / О. М. Зубань // Українське мовознавство. – № 44/1. – Київ, 2014. – С. 123–133.

КОРПУСНЫЙ АНАЛИЗ ГРАММАТИЧЕСКОГО СОСТАВА УЧЕБНЫХ ДИАЛОГОВ

Проекту SPbEFL LC¹, который создавался как «учебный корпус» (Learner Corpus), точнее, корпус текстов, порождаемых обучающимися иностранному языку, уже десять лет. Цель подобных проектов заключается в выявлении наиболее значимых ошибок, которые допускают обучающиеся в чужом языке, в особенности таких, которые не учитываются в дидактических материалах, в том числе ошибок национально-специфических, то есть связанных с влиянием родного языка обучающихся. На материале этого корпуса проведен целый ряд исследований, наиболее значимые результаты которых отражены в [1]. Однако до сих пор оставался мало исследованным подкорпус скриптов диалогов на английском языке, порожденных информантами (русскоязычными школьниками) в ответ на задание-стимул.

Предварительный анализ этих диалогов показал, что кроме очевидных ошибок (грамматика глагольных форм, порядок слов, артикли, предлоги, лексические ошибки и пр.) тексты диалогов обучающихся производят впечатление искусственности, «простоты без элегантности» [2]. Очевидно, что часть этого впечатления создается за счет многочисленных повторов слов и реплик партнера, заполненных пауз хезитации (*So..., Can you tell me..., And..., Well, yes..., Maybe...*), а также повторов формулировок задания, которые делают старательные попытки информантов варьировать микротемы внутри диалога нарочитыми и поверхностными. И все-таки за этими видимыми причинами неудач, на наш взгляд, кроются более веские основания, которые отличают диалогические тексты учащихся от обучающих образцов, тем более, от естественной диалогической речи носителей языка.

Более глубокое проникновение в причины неестественности учебных диалогов дает возможность корпусного анализа текстов сопоставимых корпусов. С этой целью был создан корпус диалогов-образцов из учебных пособий по английскому языку уровня Pre-Intermediate и Intermediate издательств Oxford University Press, Cambridge University Press, Express Publishing, Macmillan и Longman – Textbook Dialogue Corpus (TDC). Этот корпус сопоставим с корпусом диалогов SPbEFL LC, далее – Learner Dialogue Corpus (LDC) по объему (10 821 с/у и 11 561 с/у соответственно), по уровню владения языком информантов корпуса SPbEFL LC, учебным программам, темам диалогов и времени сбора материалов для корпуса.

Процедура анализа была подсказана работой Х. Ли и А.К. Фанга [3], в которой анализировался грамматический состав (grammatical composition) текстов диалогов матерей и детей – фрагмент известного корпуса CHILDES (*Child Language Data Exchange System*)² – в терминах частотных распре-

¹ SPbEFL LC – «Учебный корпус текстов петербургских школьников, изучающих английский язык» – URL: www.spbeflcorp.ru

² CHILDES – URL: <http://childes.psy.cmu.edu>

делений основных грамматических классов слов. Количественный анализ в корпусе основывался на сравнении числа словоупотреблений (токенов) и типов словоформ (типов токенов) основных грамматических классов слов в материнских и детских репликах, которые стали основой двух сопоставляемых корпусов текстов. Исследование показало, что согласно коэффициенту Пирсона, который показывает степень связи, корреляции сопоставляемых массивов текстов, материнские и детские тексты чрезвычайно близки, как по числу токенов разных грамматических классов слов ($r=0,936$), так и по типам токенов в этих грамматических классах ($r=0,991$). Однако распределение грамматических классов слов по сопоставляемым корпусам оказалось различным. Например, в текстах детей значительно преобладают существительные и местоимения, как наиболее референтно-прозрачные номинации, а употребление прилагательных, наречий, глаголов значительно уступает показателям материнских текстов. Напротив, в речи матерей наиболее представленным классом являются глаголы. Эти и другие наблюдения по корреляции инпута (речи матерей) и аутпута (речи детей) дали авторам возможность предположить, что проведенное исследование может оказаться полезным не только при исследовании процесса овладения ребенком родным языком, но и при обращении к изучению проблем овладения вторым или иностранным языком [3, с. 95].

На основе данной методики был проведен сравнительный анализ корпусов текстов диалогов-образцов («инпут») и текстов ученических диалогов («аутпут»). Анализ опирается на сравнение распределений существительных, глаголов, прилагательных, наречий, местоимений, союзов, предлогов и междометий (см. табл. 1 и 2) в текстах корпусов.

Т а б л и ц а 1

Распределение грамматических классов слов
в корпусе диалогов-образцов (TDC)

Часть речи	Кол-во типов токенов (типов словоформ)	Кол-во токенов (словоупотреблений)	Количественное соотношение токенов (токен / всего токенов %)	Количественное соотношение типов (тип / всего типов%)
Существительные	538	1353	12,5%	40%
Глаголы	302	2625	24,26%	22,47%
Прилагательные	133	425	3,93%	9,9%
Наречия	88	809	7,48%	6,55%
Местоимения	44	2179	20,14%	3,27%
Предлоги	26	851	7,86%	1,86%
Союзы	15	353	3,26%	1,12%
Междометия	32	531	4,59%	2,38%
Другое	166	1695	15,66%	12,35%
Всего	1 344	10 821	100.00%	100.00%

Распределение грамматических классов слов
в корпусе диалогов учащихся (LDC)

Часть речи	Кол-во типов токенов (типов словоформ)	Кол-во токенов (словоупотреблений)	Количественное соотношение токенов (токен / всего токенов %)	Количественное соотношение типов (тип / всего типов%)
Существительные	474	1823	15,77%	44,76%
Глаголы	201	2630	22,75%	18,98%
Прилагательные	145	572	4,95%	13,7%
Наречия	73	740	6,41%	6,89%
Местоимения	45	2227	19,26%	4,25%
Предлоги	23	852	7,37%	2,17%
Союзы	15	786	6,8%	1,42%
Междометия	35	644	5,57%	3,31%
<i>Другое</i>	48	1287	11,13%	4,53%
Всего	1 059	11 561	100.00%	100.00%

Значительная степень корреляции текстов двух корпусов подтверждается коэффициентом корреляции Пирсона¹. Коэффициент корреляции токенов (словоупотреблений) основных грамматических классов слов в сопоставляемых корпусах оказался очень высоким: $r=0,9462$. При вычислении коэффициента корреляции типов токенов (типов словоформ) этих грамматических классов в обоих корпусах был также получен высокий результат: $r=0,9639$.

Однако, как и в сравнении материнских и детских текстов, основные различия обнаружились в распределении грамматических классов слов и их реальном лексическом наполнении.

В исследовании [3] существенным различием материнских и детских текстов (инпута и аутпута) оказалась асимметрия в распределении основных грамматических классов – существительных и глаголов: в детских текстах количество существительных значительно превышало количество глаголов, в то время как в материнских текстах большая доля принадлежала глаголам. Это объясняется тем, что дети быстрее осваивают существительные (и местоимения), чем глаголы, прилагательные, наречия, семантика которых более абстрактна, в отличие от имен существительных и местоимений, которые в речи ребенка конкретно-референтны [3, с. 100].

Рисунок 1 в более явной форме демонстрирует, какие грамматические классы слов преобладают в диалогической речи информантов корпуса и в диалогах-образцах. В обоих корпусах количественно преобладают глаголы (как в речи матерей, т.е. взрослой речи), местоимения и существительные. Обращает на себя внимание количество союзных употреблений в текстах школьников.

¹ Вычисление коэффициента было осуществлено автоматически с помощью программного обеспечения IBM SPSS Statistics 22 IBM SPSS Statistics 22 – URL: www-01.ibm.com/support/

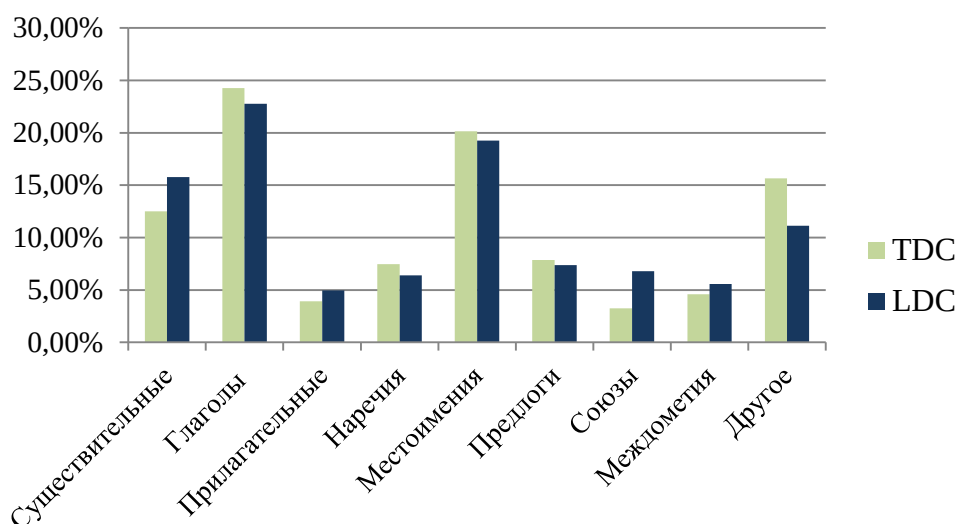


Рис. 1. Количественное соотношение грамматических классов слов (количество токенов) в обоих корпусах

Рисунок 2 дает представление о разнообразии репертуара глаголов, существительных и других грамматических классов:

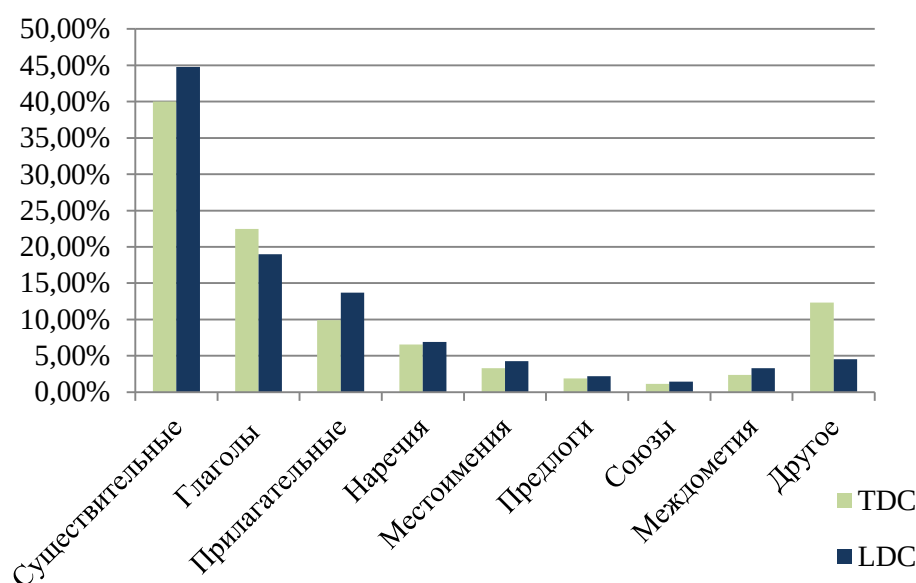


Рис. 2. Количественное соотношение грамматических классов слов (количество типов токенов) в обоих корпусах

Мы видим, что в обоих корпусах наиболее разнообразно представлены существительные, причем в ученических диалогах количество типов именных словоформ больше, чем в диалогах-образцах (как в речи детей), но их глагольный репертуар уже – свидетельство того, что обучающиеся избегают глагольных синонимов и сложных глагольных форм. Общее количество союзов в диалогах школьников почти в 2 раза (рис. 1) больше, чем в диалогах-образцах, но по репертуару они близки. Такой показатель – результат перепроизводства (overuse) в диалогической речи школьников союза *and*. Примечательно, что он

встречается в начале вопросительных предложений (*And what about/And do you/And what do you..?*), выполняя функцию слова-связки (linking word) и используется как некая «универсальная» форма вовлечения собеседника в разговор. Учащиеся заполняют им паузы хезита-ции, когда вспоминают или подбирают нужное слово (*So... and... I like very much that you are listening to rock music/ comedies and ...or horror.../Oh, it is a pity and... Oh, maybe we should go to my home. / But my friends go and...and I can't go with ...without them*).

Союз *and* эквивалентен русским союзам *и, а*, которые имеют высокую частоту в разговорной речи¹ и выполняют функцию связи предложений в высказывании. Поэтому частое употребление информантами союза *and* может быть вызвано интерференцией родного языка.

Эти и другие наблюдения, сделанные при сопоставлении распределений грамматических классов слов в диалогических корпусах текстов, позволяют заключить, что при очень высокой степени корреляции их грамматического состава тексты диалогов учащихся значительно отличаются от текстов диалогов-образцов из учебников репертуаром лексических средств, составляющих эти классы. Особо скудным представляется глагольный репертуар из-за повтора наиболее простых глаголов, которые усваиваются учащимися еще на самых ранних этапах обучения и часто выполняют в их речи роль неравнозначной замены, становясь основой конструкций внутренней, переходной грамматики и являясь тем самым интерференцией родного языка учащихся. Кроме того, глагольный репертуар школьников обеднен использованием простых глагольных форм и избеганием видо-временных форм. Наблюдается явный дефицит во владении так называемыми комментаторными глаголами – модальными, фазовыми и грамматическими, которые составляют вершину глагольного списка текстов диалогов-образцов. Более разнообразен репертуар существительных, что, весьма вероятно, как и в речи детей, осваивающих язык, может быть следствием референтной прозрачности именных номинаций.

Данные проведенного корпусного исследования могут внести поправки в методики и дидактические материалы обучения диалогической речи.

ЛИТЕРАТУРА

1. Камшилова, О. Н. Учебный корпус: потенциал, состав, структура / О. Н. Камшилова – СПб. : ООО «Книжный дом», 2012. – 56 с.
2. Hinkel, E. Simplicity without elegance: Features of sentences in L1 and L2 Academic texts / E. Hinkel // TESOL Quarterly – Vol.37. – № 2. – 2003. – P. 275–301.
3. Li. Word frequency of the CHILDES corpus: Another perspective of child language features / Li, Hanhong, Fang, C. Alex // ICAME Journal. – № 35 – 2010. – P. 95–112.

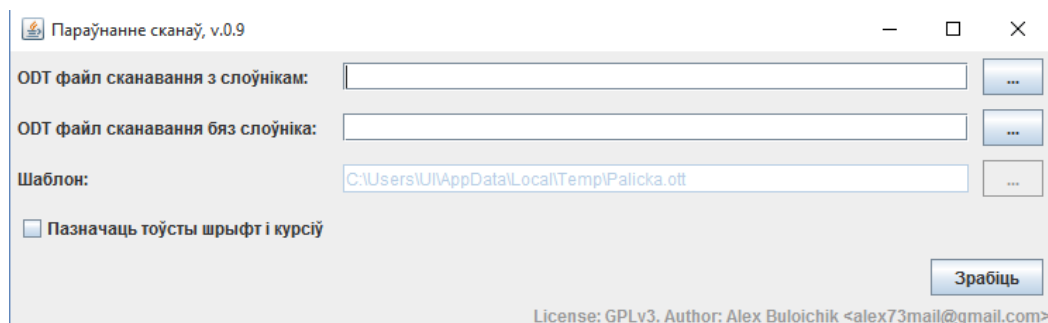
¹ Устный подкорпус Национального корпуса русского языка состоит из 189 481 предложения, в которых союз *и* встречается в начале предложения 5 511 раз, союз *а* – 16 236 раз.

ПРАГРАМНАЕ ЗАБЕСПЯЧЭННЕ І ПРАЦЭС ПАДРЫХТОЎКІ ТЭКСТАЎ ДЛЯ НАЦЫЯНАЛЬНАГА КОРПУСУ БЕЛАРУСКАЙ МОВЫ

Падчас працы над Нацыянальным корпусам беларускай (НКБ) мовы быў створаны шэраг праграм для працы з лексіка-граматычнай базай, якая ляжыць у аснове граматычнай разметкі корпусу, і ўласна Нацыянальным корпусам беларускай мовы: рэдактар метатэкставай разметкі (пашпарту) корпусу, рэдактар корпусу (дазваляе захоўваць тэкст у фармаце корпусу, здымаць рознаўроўневую аманімію ў паўаўтаматычным рэжыме, папаўняць лексіка-граматычную базу і г.д.), праграма правэркі лексіка-граматычнай базы на паслядоўнасць, праграма параўнання сканаваных тэкстаў, праграма канвертацыі тэкстаў у папярэдні фармат корпусу.

Папаўненне корпусу. Адсканаваны тэкст распазнаецца ў праграме для аптычнага распазнавання тэкстаў *ABBYY FineReader*. Для аптымізацыі працы на аснове лексіка-граматычнай базы быў створаны слоўнік для FineReader і шэраг праграмных сродкаў, якія дазвалялі аўтаматызаваць працэс.

Тэкст праходзіць распазнаванне два разы: першы раз распазнаецца аптычна без слоўніка, а другі раз – са слоўнікам. Атрыманыя два файлы параўноўваюцца пры дапамозе ўласнапрацаванай праграмы “*Параўнанне сканаў*”:



Праграма візуальна параўноўвае файл, адсканаваны без слоўніка, і файл, адсканаваны са слоўнікам, сінхранізуе гэтыя файлы, прымяняе шаблон і вынікі захоўвае ў асобны файл з пашырэннем .odt.

Выніковы тэкст у файле .odt мае наступны выгляд:

I
Восень заўсёды падкрадваецца неўзаметку. Янічэ {Яшчэ} прыграве сонца, усё вакол зялёнае, яшчэ цвітуць гуркі, на доўгай гарбузовай касе смела ўсміхаецца вялізная пяцікутная жоўта-залатая кветка. Лета не хоча здавацца. I {I} раптам прачнешся раніцаю {раніцаю} — на вуліцы туман, густы, белы, як малако. А калі ён развеецца, на бярозе ярчай засвецяцца жоўтыя лісцікі, бульбоўнік пачарнее, бы абвараны, і раса на траве халодная, аж коле ў босыя пяткі.

Звычайна на трэцім перапынку хлапчукі выбягалі на ўзгор’е і скакалі з абрыву ў пясчаную ямку. На дне яе жвірысты пясок быў халаднаваты, хоць зверху трохі праграваўся, ды неўзабаве парэпаны ад цыпак ногі перамешвалі яго, ператаўкалі, як проса ў ступе.

Скакалі хлапчукі не проста так сабе. Тут ішло зацятае спаборніцтва — хто далей скокне. Не раз быў чэмпіёнам Андрэйка Сахута. Але нечакана для ўсіх уперад вырываўся Парасчын Данілка. Андрэйка гэтага не мог перажыць. Калі б Паўлік, ягоны даўні сябрук, было б не так крыўдна, а то смаркаты Данька раптам сігануў далей за ўсіх. Мабыць, і Паўліка гэта раздражніла, ён шмаргануў носам і са здзіўленнем у голасе сказаў:

— Смелы {еты} байструк. У матку ўдаўся. Яна ваўка забіла...

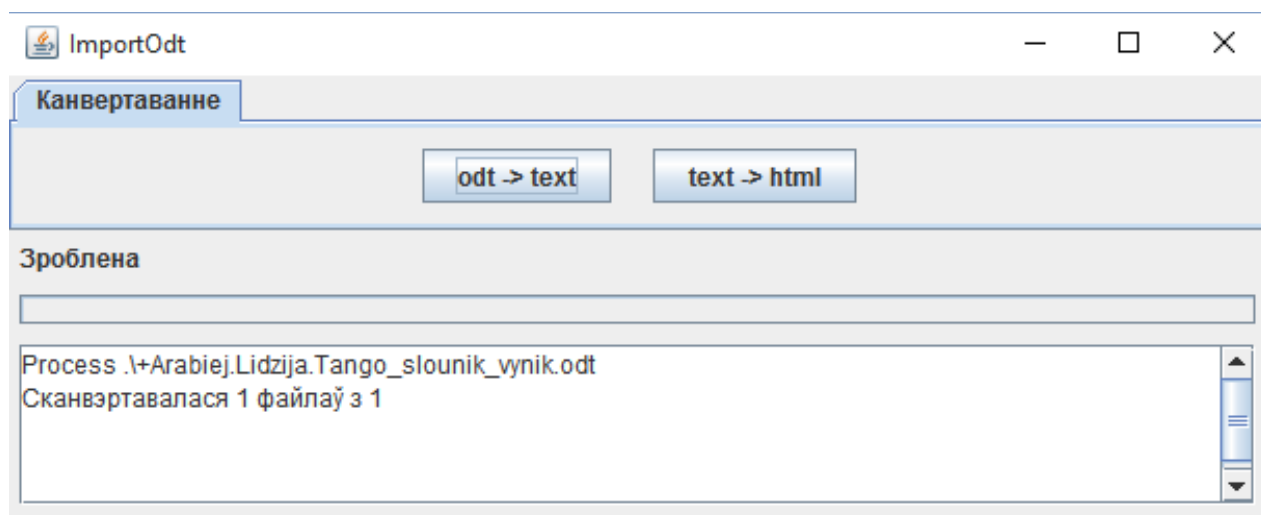
— Хто байструк? А то ў нос як дам зараз! — замахнуўся брудным кулачком Данька.

— Ну, папрабуй! — натапырыўся Паўлік. — Як дам выпятка, дык і пакоцішся прама ў Бесадзь.

Колерам у тэксце вылучаны спрэчныя месцы, на якія трэба асабліва звяртаць увагу. У аснову кладзецца тэкст, расчытаны пры дапамозе слоўніка, але распазнаванне са слоўнікам мае адзін мінус: праграма падбірае найбліжэйшае слова, якое, паводле алгарытму, найбольш пасуе ў гэтым месцы і здараюцца выпадкі, калі слова падабрана, але яно памылковае, таму побач для праверкі ў фігурных дужках пакідаецца слова, распазнанае без слоўніка. У нашым выпадку відаць, што формы “Янічэ” (Яшчэ) і “ёты” (еты) былі падабраны няўдала, на што ўказвае варыянт у фігурных дужках, а слова “раніцаю” – удала. Такі падыход дазволіў істотна сэканоміць час пры вычытцы тэкстаў.

Адначасова з вычыткай тэксту правяралася і пазначалася яго структура, закладзеная ў шаблоне: іерархія загаловаў, эпіграфы, подпісы, паэтычныя ўстаўкі ў прэзінтэчных тэкстах.

Такім чынам, вычытаны і структураваны тэкст далей апрацоўваецца пры дапамозе праграмы ImportODT (уласная распрацоўка):



Праграма правярае правільнасць разметкі структуры тэксту і канвертуе файл *odt* у фармат *txt* і потым у фармат *html* з структурнай разметкай:

```
<!-- BOOK_BEGIN -->
<p>&nbsp;</p>
<div class="nootnote-block">
</div>
<h3><a id="chapter1"></a>«ТАНЦУЙ·ТАНГО...»</h3>
<p>Гарадок·быў·нібыта·незнаёмы·чалавек,·пра·якога·думаеш·—·вельмі·ён·на·
некага·падобны,·толькі·не·ўспомніш·—·на·каго,·а·потым·пачынаеш·разумець,·што·
гэта·проста·такое·аблічча,·блізкае·табе·аблічча,·можа,·нават·земляка,·
чалавека·з·тваёй·вёскі.</p>
<p>Так·і·гэты·гарадок.·Колькі·іх,·вось·гэтакіх·—·з·зьялёным·скверыкам·у·
цэнтры,·з·Домам·культуры,·у·архітэктур·якога·можна·пазнаць·былую·царкву·
```

Калі ў структурнай разметцы ёсць недакладнасці ці супярэчнасці, то праграма ўказвае на месца і тып памылкі. Такім чынам, тэкст падрыхтаваны для ўключэння ў корпус.

Наступны этап – пашпартызацыя тэксту. Для гэтых мэтай была распрацавана асобная праграма (гл. малюнак ніжэй), якая дазваляе ў паўаўтаматычным рэжыме ствараць апісанне тэксту (метатэкставую інфармацыю):

Выпраўленне загалоўкаў кнігі

Захаваць

Аўтары(праз,): Альгерд Абуховіч

Назва: Старшына

Крыніца: Беларуская літаратура : Хрэстаматыя. Частка другая. (Другая палова XIX ст.) Дапаможнік для польскіх студэнтаў па спецыяльнасці "Беларуская філалогія" / Уклад. і камент. М. Хаўстовіча. Мн., 2006. - 204 с.

Год напісання: 1914?

Год першага выдання: 1915

Выданне: Беларуская літаратура : Хрэстаматыя. Частка другая. (Другая палова XIX ст.) Дапаможнік для польскіх студэнтаў па спецыяльнасці "Беларуская філалогія" / Уклад. і камент. М. Хаўстовіча. Мн., 2006. - 204 с.

Год выдання: 2006

Схаваны каментар:

Правапіс: ☒ ☐ A1933 ☐ A1957 ☐ A2008 ☐ T1929 ☐ K2005 ☐ NIES

Стыль/жанр:

☐ афіцыйна-дзелавы/аб'ява ☐ іншыя ☐ мастацкі/аповяданне ☐ навуковы/анатацыя

☐ афіцыйна-дзелавы/акт ☐ мастацкі/аповесць ☐ навуковы/даклад

☐ афіцыйна-дзелавы/аўтабіяграфія ☐ мастацкі/баллада ☐ навуковы/дысертацыя

☐ афіцыйна-дзелавы/афіцыйнае паведамленне ☒ мастацкі/басня ☐ навуковы/лекцыя

☐ афіцыйна-дзелавы/даведка ☐ мастацкі/верш ☐ навуковы/манаграфія

☐ афіцыйна-дзелавы/даверанасць ☐ мастацкі/казка ☐ навуковы/навуковы артыкул

☐ афіцыйна-дзелавы/дамова ☐ мастацкі/ода ☐ навуковы/навуковы выступ

☐ афіцыйна-дзелавы/загад ☐ мастацкі/паэма ☐ навуковы/навуковае паведамленне

☐ афіцыйна-дзелавы/закон ☐ мастацкі/п'еса ☐ навуковы/падручнік

☐ афіцыйна-дзелавы/заява ☐ мастацкі/раман ☐ навуковы/папулярны артыкул

Раздзел:

☐ Проза ☐ Гісторыя, палітыка ☐ Замежная паэзія ☐ Відэа

☒ Паэзія ☐ Рэлігія ☐ Літаратуразнаўства ☐ Рознае

☐ Драматургія ☐ Філасофія ☐ Іншамовныя творы і пераклады ☐ Архівы

☐ Дзіцячая літаратура ☐ Міталогія ☐ Кулінарыя ☐ Газеты, часопісы

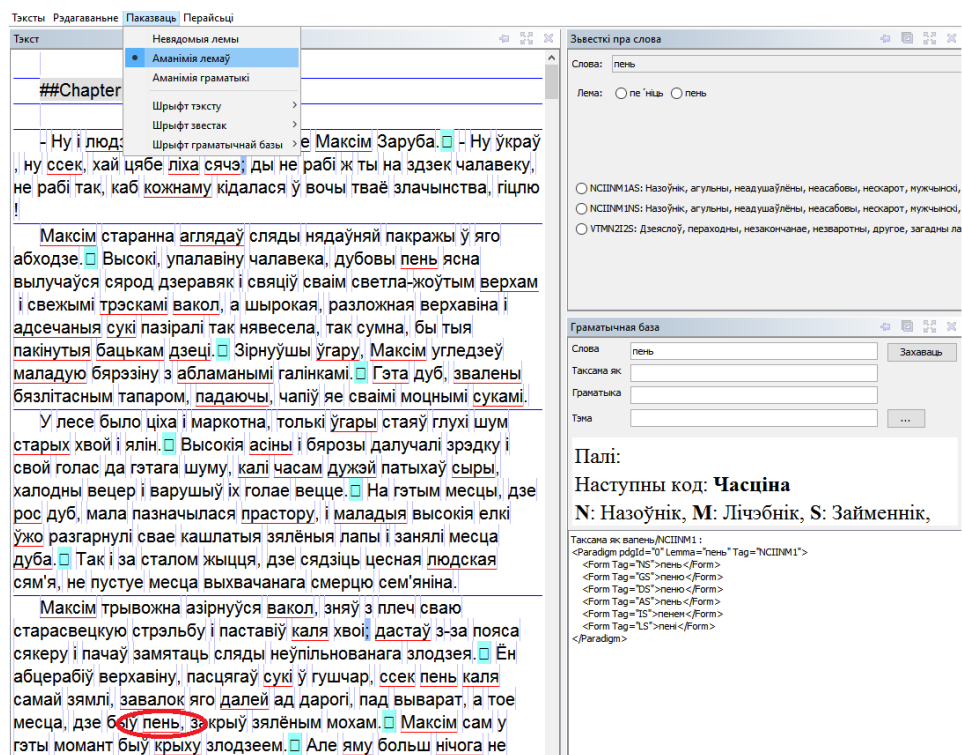
☐ Мова ☐ Аўдыёкнігі ☐ Мэдыцына

☐ Падручнікі ☐ Замежная проза ☐ Мастацтва, архітэктура

Пры дапамозе праграмы можна пазначыць аўтара (-аў) тэксту, яго назву, крыніцу атрыманага тэксту, год яго стварэння, першага выдання, актуальнага выдання, зрабіць дадатковыя заўвагі, пазначыць правапіс, стыль і жанр. Адкрытая архітэктара праграмы дазваляе пры патрэбе дадаваць іншыя параметры тэксту. У выніку ў файл дадаецца інфармацыя наступнага кшталту:

```
##Authors: Альгерд Абуховіч
##Title: Старшына
##SectionsTheme: Паэзія
##Source: Беларуская літаратура : Хрэстаматыя. Частка другая. (Другая палова XIX ст.) Дапаможнік для польскіх студэнтаў па спецыяльнасці "Беларуская філалогія" / Уклад. і камент. М. Хаўстовіча. Мн., 2006. - 204 с.
##Description: упершыню апублікаваны ў Беларускі календар «Нашай нівы» на 1915 г. – Вільня, 1915
##Edition: Беларуская літаратура : Хрэстаматыя. Частка другая. (Другая палова XIX ст.) Дапаможнік для польскіх студэнтаў па спецыяльнасці "Беларуская філалогія" / Уклад. і камент. М. Хаўстовіча. Мн., 2006. - 204 с
##StyleGenre: мастацкі/басня
##CreationYear: 1914?
##PublicationYear: 2006
##FirstPublicationYear: 1915
##HiddenSource:
```


Пасля дадавання метатэкставай інфармацыі, тэкст апрацоўваецца праграмай рэдагавання корпусу:



Праграма працуе на камп'ютары карыстальніка, выкарыстоўвае толькі чытанне з лексіка-граматычнай базы. З праграмай могуць працаваць розныя людзі адначасова (кожны на сваім камп'ютары), але кожны будзе апрацоўваць свой тэкст. Працэс нагадвае рэдагаванне звычайнага тэксту ў тэкставым рэдактары:

- адчыняецца тэкставы файл;
- праграма робіць аўтаматычнае сегментаванне і такенізацыю;
- спецыяліст рэдагуе тэкст, змяняе сегментаванне, такенізацыю, здымае аманімію і г.д.;
- тэкст захоўваецца ў фармаце корпусу і перадаецца ў сховішча тэкстаў корпусу.

Праграма дазваляе працаваць у некалькіх рэжымах: 1) *Невядомыя лемы* – выдзяляюцца словы, якія не трапілі ў ЛГБ. Праграма дазваляе разгарнуць парадыгму гэтых слоў, прысвоіць адпаведную граматыку і дадаць слова ў ЛГБ; 2) *Аманімія лемаў* – выдзяляюцца словы з часцінамоўнай аманіміяй (гл. прыклад на малюнку вышэй). Гэты рэжым дазваляе здымаць часцінамоўную аманімію ў паўаўтаматычным рэжыме; 3) *Аманімія граматыкі* – выдзяляюцца словы з граматычнай аманіміяй (аманімія ўнутры адной парадыгмы), зняцце граматычнай аманіміі таксама ажыццяўляецца ў паўаўтаматычным рэжыме.

Дадаткова праграма паказвае межы абзацаў, сказаў і слоў, што дазваляе кантраляваць правільнасць пазначэння структурных элементаў пры папярэдняй аўтаматычнай разметцы.

Пры захаванні, праграма канвертуе адпрацаваны тэкст ў фармат корпусу:

```
<XMLText>
  <Header>
    <Tag name="Authors">Якуб Колас</Tag>
    <Tag name="CreationYear">1913</Tag>
    <Tag name="FirstPublicationYear">1914</Tag>
    <Tag name="SectionAuthor">Апавяданьні</Tag>
    <Tag name="SectionsTheme">Проза</Tag>
    <Tag name="StyleGenre">мастацкі/апавяданне</Tag>
    <Tag name="Title">Малады дубок</Tag>
  </Header>
  <Content>
    <p/>
    <Tag name="Chapter">I</Tag>
    <p/>
    <p>
      <se>
        <z cat="KM1">-</z>
        <s char=" " />
        <w cat="E_У" lemma="ну">Ну</w>
        <s char=" " />
        <w cat="СКХ_Е" lemma="і">і</w>
        <s char=" " />
        <w cat="NCAPNM1NP" lemma="чалаве́к" manual="true">людзі</w>
        <z cat="KE">!</z>
      </se>
      <se>
        <s char=" " />
        <z cat="KM1">-</z>
        <s char=" " />
        <w cat="VDMN2PXSM" lemma="гавары́ць">гаварыў</w>
      </se>
    </p>
  </Content>
</XMLText>
```

На сённяшні дзень, створаны электронны корпус тэкстаў аб’ёмам 80 000 000 словаўжыванняў са структурнай і граматычнай разметкай і пашпартызацыяй і распрацаваны эксперыментальны рухавік корпусу.

Папярэдняя версія корпусу аб’ёмам каля 50 000 000 словаўжыванняў размешчана з мэтай тэсціравання ў інтэрнэце па адрасе <http://bnkorporus.info/>. Корпус дазваляе шукаць дакладную форму слова, слова з усімі словаформамі, паводле граматычных характарыстык. Пошук можна ажыццяўляць з улікам аўтара, году напісання, стылю, жанру і г.д. Створана магчымасць выбаркі па кластарых (спалучэннях суседніх слоў).

ПРОБЛЕМЫ ПЕРЕВОДА ТЕКСТОВ

О. Ю. Горбунова

ОБУЧАЮЩИЙ ФРАНЦУЗСКО-РУССКИЙ АВТОМОБИЛЬНЫЙ ТЕЗАУРУС В АСПЕКТЕ ДИДАКТИКИ ТЕХНИЧЕСКОГО ПЕРЕВОДА

В настоящее время практика составления тезаурусов в дидактических целях значительно расширяет свои границы. Успешно развивается так называемая учебная тезаурусная лексикография. Учебные или обучающие тезаурусы представляют собой эффективное средство овладения лексикой иностранного языка, в том числе и специальной терминологией конкретной предметной области знания в условиях технического перевода. Основная цель учебного (обучающего) тезауруса состоит в том, чтобы помочь студентам освоить семантические отношения, существующие между словами, терминами отдельного иностранного языка. Эта цель достигается путем объединения лексических единиц по тематическому признаку в отдельные группы или семейства, исходя из тезаурусного принципа.

Технический перевод представляет собой сложный процесс речемыслительной деятельности, который предполагает владение определенными компетенциями и требует особой подготовки. В настоящее время исследований по теории и практике такой подготовки явно недостаточно, что вызвано рядом причин. Эти причины обусловлены следующими недостатками:

- малая по времени переводческая практика в расчете на каждого обучающегося в группе;
- отсутствие необходимой индивидуализации и дифференциации обучения;
- недостаточная разработка учебно-методических материалов по переводу технической документации;
- отсутствие новаторских технологий и инноваций в подготовке лингвиста-переводчика.

Перечисленный ряд недостатков позволяет выявить противоречия между:

- насущной необходимостью переориентации процесса обучения переводческой деятельности, с одной стороны, и отсутствием эффективной модели обучения, призванной сформировать умения и навыки, лежащие в основе профессиональной переводческой компетентности, с другой;
- потребностью общества в подготовке квалифицированных лингвистов-переводчиков и ограниченным допуском к актуальной научно-технической информации на иностранном языке в сфере передовых отраслевых технологий, которая нередко имеет закрытый характер;
- переходом в сфере профессионального образования на двухступенчатую систему подготовки бакалавров и магистров и отсутствием инновационных педагогических технологий формирования их компетенций.

В Тольяттинском государственном университете накоплен значительный опыт по подготовке лингвистов-переводчиков, владеющих двумя иностранными языками: английским и французским. Подобное сочетание двух

иностранных языков в переводческой деятельности как никогда лучше отвечает сложившейся в настоящее время ситуации на АВТОВАЗе в связи с развивающимся техническим сотрудничеством с французским концерном RENAULT. В процессе формирования профессиональной компетентности у студентов – будущих переводчиков с французского на русский возникают значительные трудности при работе с текстами научно-технической направленности, в первую очередь – в области автомобильной индустрии. Проблемы вызывают не только грамматическое оформление текста технического перевода, но и, прежде всего, его лексическая составляющая, в частности, многочисленная техническая терминология. В этой связи актуальным представляется тезаурусный подход к формированию профессиональной компетентности переводчика в области специальной технической коммуникации. В основе этого процесса находится особый лингводидактический инструмент – обучающий словарь-тезаурус, охватывающий двуязычную техническую коммуникацию в предметной области (ПО) «Автомобилестроение».

В настоящей статье рассматривается проблема формирования профессиональной переводческой компетентности (ППК) у студентов направления подготовки бакалавров 45.03.02 «Лингвистика» (профиль «Перевод и переводоведение») на практических занятиях по дисциплине «Практический курс перевода второго иностранного языка – французского». Решение проблемы эффективного формирования ППК лингвиста находим в русле современной прикладной лингвистики и лингводидактики, в частности – в сфере учебной тезаурусной лексикографии. В основе настоящего исследования лежат лингвистические и лингводидактические традиции, сложившиеся в Международной научной группе «Статистика речи» [1; 2; 3].

Опираясь на опыт предшественников [4–10], нами был разработан обучающий французско-русский автомобильный тезаурус (ОФРАТ), в котором техническая терминология распределяется по семантическим полям с указанием синтагматических и парадигматических отношений. ОФРАТ был построен в результате предварительного анализа корпуса французских технических текстов, который включает учебники, научные статьи, монографии, технические описания, патенты изобретений и рекламу в ПО «Автомобилестроение». Выявленные в ходе анализа французских технических текстов термины и их определения в текстах были сопоставлены с дефинициями этих терминов, приводимых в словарях и энциклопедиях французских технических терминов. В результате компонентного анализа дефиниций французских технических терминов ПО «Автомобилестроение» была выявлена синоптическая схема, которая отражает систему ключевых понятий, вертикально и горизонтально связанных между собой сетью семантических отношений. Данная синоптическая сеть состоит из отдельных словарных статей тезауруса, каждая из которых отражает отдельное семантическое поле ключевого термина ПО «Автомобилестроение».

Словарная статья ОФРАТа содержит:

- 1) заглавный термин, отражающий одно из ключевых понятий «терминополья» и его русский эквивалент;
- 2) варианты термина во французском языке и их русские соответствия;
- 3) синонимы французского термина и их русские эквиваленты;

- 4) антонимы французского термина и их русские соответствия;
- 5) гипероним французского термина с переводом на русский язык;
- 6) гипонимы заглавного термина и их русские эквиваленты;
- 7) голоним французского термина и его русский эквивалент;
- 8) меронимы французского термина и их русские эквиваленты и т. д.

Термины обучающего тезауруса связаны семантическими отношениями, среди которых преобладают:

1) гиперонимия (*hyperonymie*): от *hyperonyme* – гипероним: понятие в отношении к другому понятию, выражающее более общую сущность. В отношении некоторого множества объектов гиперонимом является понятие, отражающее надмножество к исходному, т. е. это слово с более широким значением, выражающее общее, родовое понятие, название класса предметов, например: *automobile* (родовое понятие) и отдельные виды автомобиля: *cabriolet* – кабриолет, *limousine* – лимузин и т. д.;

2) гипонимия (*hyponymie*): от *hyponyme* – гипоним: понятие, выражающее частную сущность по отношению к другому, более общему понятию. Так, термины, выражающие видовые понятия автомобиля: *cabriolet* – кабриолет, *limousine* – лимузин – являются гипонимами по отношению к термину *automobile* – автомобиль;

3) меронимия (*méronymie*): это классификация явлений, основанная на отношениях меронимии и голонимии. Меронимия и голонимия, как семантические отношения являются взаимно обратными друг другу. Мероним выражает понятие, которое является составной частью другого. Голоним выражает понятие, которое является целым над другими понятиями, т.е. другое понятие является частью первого. Например, термины *moteur* – двигатель, *roue* – колесо являются меронимами по отношению к термину *automobile* – автомобиль. В свою очередь, термин *automobile* – автомобиль является голонимом по отношению к терминам *moteur* – двигатель и *roue* – колесо;

4) синонимия (*synonymie*): синонимы – слова одной части речи, различные по звучанию и написанию, но близкие по значению. Примеры синонимов во французском языке: *automobile* (автомобиль): *bagnole* – машина; *caisse* – колымага; *charrette* – телега;

5) корреляция (*corrélation*): от лат. *correlatio* – соотношение, взаимосвязь. В отношении корреляции находятся термины *automobile* – автомобиль и *route* – дорога, а также *route* – дорога и *police de route* – дорожная полиция.

Отличительная черта словарной статьи «Обучающего французско-русского автомобильного тезауруса» состоит в том, что она имеет открытый характер, то есть с учетом специальных целей словарная статья может быть расширена за счет новых терминов с помощью особых пометок, которые вводят в тезаурус новую терминологию, отражающую актуальные тенденции эволюции рассматриваемой предметной области.

ОФРАТ имеет явную дидактическую направленность, поскольку студент в процессе работы с этим словарем приобретает необходимые лингвистические знания (специальная техническая терминология), а также навыки и умения, которые в целом составляют профессиональную компетентность технического переводчика. Основное назначение настоящего тезауруса состоит

в его использовании в качестве лингводидактического инструмента формирования профессиональной компетентности у студентов – будущих переводчиков в сфере специальной технической коммуникации.

В процессе самостоятельной работы по техническому переводу с использованием настоящего словаря-тезауруса студент должен следовать алгоритму переводческих действий, который включает следующие операции.

1. Поиск французского термина в алфавитном указателе.
2. Определение адресов семантических полей, в которые входит французский термин.
3. Последовательное обращение к семантическим полям французского термина по указанным адресам.
4. Выбор русского эквивалента французского термина в результате анализа его семантических полей с учетом реального контекста.

В результате многократного обращения к различным семантическим полям того или иного французского технического термина в поисках соответствующего русского эквивалента у студента формируется профессиональный тезаурус, отражающий соответствующую предметную область. С течением времени в условиях практического обучения техническому переводу обучающий французо-русский автомобильный тезаурус закрепляется в профессиональном сознании переводчика и в конечном итоге становится показателем его профессиональной компетентности как специалиста в сфере специальной ПО «Автомобилестроение».

В конечном итоге можно констатировать, что:

- 1) обучение лингвистов-переводчиков должно строиться на основе тезаурусной подхода, так как данная технология способствует оптимизации учебного процесса и наиболее эффективному обучению специальному техническому переводу;
- 2) базовые и специфические знания, умения и навыки специального технического перевода формируются в процессе выполнения особых упражнений с учетом тезаурусного подхода;
- 3) использование в процессе обучения специальному техническому переводу тезаурусного подхода в профессиональной подготовке лингвистов-переводчиков оптимально способствует формированию знаний, умений, навыков перевода, составляющих ППК.

Предлагаемый тезаурусный подход не только способствует эффективному изучению содержания ПО «Автомобилестроение», но и содействует развитию у студентов специального технического мышления, которое необходимо при решении практических задач в будущей профессиональной деятельности лингвиста-переводчика с французского языка на русский в области специальной технической коммуникации.

Результаты внедрения «Обучающего французо-русского автомобильного тезауруса» в учебный процесс позволяет сделать вывод о его существенной значимости в русле развивающейся учебной тезаурусной лексикографии, поскольку его использование, как показали статистические данные его экспериментальной верификации, значительно содействует дальнейшему развитию компетентностного подхода в области теории и методики профессионального образования.

ЛИТЕРАТУРА

1. *Пиотровский, Р. Г.* Лингвистический автомат (в исследовании и непрерывном обучении) : учеб. пособие / Р. Г. Пиотровский. – СПб. : Изд-во РГПУ им. А. И. Герцена, 1999. – 256 с.
2. *Беляева, Л. Н.* Основы теории и практики перевода: материалы для учебно-методического сопровождения дисциплины / Л. Н. Беляева. – 2-е изд., испр. и доп. – СПб. : Изд-во «ЛЕМА», 2010. – 184 с.
3. *Зубов, А. В.* Информационные технологии в лингвистике : учеб. пособие для студ. лингв. фак-тов высш. учеб. заведений / А. В. Зубов, И. И. Зубова. – М. : Издательский центр «Академия», 2004. – 208 с.
4. *Тараносова, Г. Н.* Анализ художественного текста в профессиональной подготовке филолога: учебное пособие / Г. Н. Тараносова; под ред. действ. члена РАО, д-ра филол. наук, проф. Н. М. Шанского. – Тольятти : Изд-во ТГУ, 2004.
5. *Горбунов, Ю. И.* Французско-русский грамматический тезаурус : учеб. пособие / Ю. И. Горбунов. – Тольятти : Изд-во ТГУ, 2003. – 152 с.
6. *Горбунов, Ю. И.* Французская грамматическая терминология: Опыт тезаурусного исследования / Ю. И. Горбунов // Монография; науч. ред. Р. Г. Пиотровский. – СПб. : Изд-во РГПУ им. А. И. Герцена, 2004. – 187 с.
7. *Горбунов, Ю. И.* Тезаурусное моделирование французской грамматической терминологии : автореф. дис. ... д-ра филол. наук / Ю. И. Горбунов. – СПб. : Российск. гос. пед. ун-т им. А. И. Герцена, 2004. – 28 с.
8. *Горбунов, Ю. И.* Тезаурусное моделирование метаязыка французской фонетики и фонологии / Ю. И. Горбунов // Научное мнение. – СПб, Санкт-Петербургский университетский консорциум, 2012. – № 9. – С. 46–52.
9. *Матвеева, О. Н.* Обучающий тезаурус по английскому переводоведению: науч. издание / О. Н. Матвеева. – Тольятти : ТГУ, 2010. – 84 с.
10. *Матвеева, О. Н.* Формирование профессиональной компетенции у студентов – будущих переводчиков с использованием обучающего тезауруса : автореф. дис. ... канд. пед. наук / О. Н. Матвеева. – Самара : Поволжская гос. соц.-гум. академия, 2010. – 18 с.

Е. Б. Козеренко

ЛИНГВИСТИЧЕСКИЕ ПРЕДСТАВЛЕНИЯ В ИНТЕЛЛЕКТУАЛЬНЫХ МНОГОЯЗЫЧНЫХ СИСТЕМАХ

Текущий этап развития систем машинного перевода и систем извлечения знаний из текстов характеризуется интенсивным процессом «гибридизации» подходов и моделей, исследованиями в области когнитивной семантики, вероятностных языковых моделей и разработкой семантико-синтаксических представлений, в которых учитываются многозначность и неоднозначность синтаксических структур. Значительные вычислительные ресурсы современных систем позволяют накапливать и использовать ранее переведенные текстовые фрагменты, обеспечивать машинный перевод, основанный на прецедентах (“Example-Based Machine Translation”) [1; 2; 3], эффективно поддерживать компоненту «переводческой памяти» (“Translation

Memory”) [4; 5; 6]. Создатели систем, основанных на правилах, вводят в правила веса и вероятностные характеристики, которые позволяют отобразить динамику и разнообразие языковых форм и значений, порождаемых в процессе речевой деятельности, а сторонники статистических методов построения лингвистических моделей все чаще обращаются к подходам, основанным на лингвистических знаниях, рассматривая это как средства «интеллектуализации» систем [4; 5; 6; 7; 8].

Для машинного перевода наиболее сложной проблемой является реализация языковых трансформаций, которые необходимо производить при переводе с одного языка на другой. Новое содержание проблеме языковых трансформаций придают современные реалии: необходимость проектировать и развивать обучающие компоненты систем машинного перевода и обработки текстовых знаний на основе уже существующих и вновь создаваемых корпусов параллельных текстов. При разработке компьютерных моделей русской грамматики доминировали подходы, основанные на грамматиках зависимостей [9] или локально-синтаксических представлениях, при этом не разрабатывались грамматики составляющих. Предлагаемый нами подход, разработанный в лаборатории Компьютерной лингвистики и когнитивных технологий обработки текстов ФИЦ ИУ РАН – ИПИ дает возможность компактного представления структуры составляющих предложения (грамматика фразовых структур), с одной стороны, а, с другой стороны, учитывает механизмы зависимости между узлами дерева предложения.

Одно из направлений лингвистического моделирования связано с использованием функционального подхода. Разработка понятия функции, являющегося центральным в функциональной грамматике, связана с широкой проблематикой функций языка [10; 11; 12; 13; 14; 15; 16; 17; 18]. Функции связаны со значениями языковых единиц, но они не тождественны им. Исследование функции некоторой языковой формы включает анализ ее значения (или ряда значений в случае многозначности). Функциональный подход интегрирует разноуровневые языковые средства (синтаксические, лексические, словообразовательные и словоизменительные) на основе их функционально-семантических характеристик.

Отношение между понятиями категория и функция можно иначе выразить как исходное структурное значение и реализуемое структурное значение. Подобная концепция была высказана Леонардом Блумфилдом в его работе «Язык». Так, с точки зрения Л. Блумфилда, каждая лексическая форма связана с грамматическими формами в двух направлениях. С одной стороны, лексическая форма, даже когда она взята сама по себе, абстрагировано, обнаруживает значимую грамматическую структуру. С другой стороны, лексическая форма в любом конкретном высказывании, являясь особой языковой формой, всегда сопровождается той или иной грамматической формой. На основе различных функций могут возникать частично совпадающие формальные классы. Так, выполнение функции действующего лица характерно для субстантивных выражений и для типично инфинитивных словосочетаний (to scold the boys would be foolish «бранить мальчиков было бы глупо» [19].

Грамматика данного типа рассматривает в единой системе средства, относящиеся к разным языковым уровням, но объединенные на основе их семантических функций; при описании языкового материала используется подход «от семантики к ее формальному выражению» («от функций к средствам») как основной, определяющий построение грамматики, в сочетании с подходом «от формы к семантике» («от средств к функции»). Под единицами строя языка подразумеваются, прежде всего, грамматические формы слова и синтаксические конструкции, а также единицы «строевой лексики» [20]: модальные и фразовые глаголы, слова типа вчера, обычно, часто, прежде, долго и т.п.

Ю. С. Степанов [18] вводит понятие функтора как языкового средства транспозиции одного множества языковых единиц в другое множество языковых единиц того же языка. Функтор — это свойство или значение функтора

Понятие функции является одним из центральных в коммуникативной грамматике Г. А. Золотовой. Функция — это предназначенность элемента к определенному способу существования в системе, к определенному служению этой системе [21]. Если за целое принимаем предложение в его коммуникативном назначении, то функции его элементов, его составных частей определяются как их строительные, комбинаторные потенции, реализуемые в построении предложения.

Функции реализуются при взаимодействии языковых объектов и их контекстов.

Для рассмотрения семантики способов конфигурирования языковых структур мы пользуемся понятием структурного знака [22], предложенным в семиотической лингвистике С. К. Шаумяном. При этом слово также рассматривается нами не с точки зрения его лексического значения, а как функционально-категориальная единица, минимальный структурный знак.

В семиотической лингвистике вводится понятие суперпозиции функций, означающее, что каждый языковой объект обладает исходной первичной функцией, а происходящие в действующем языке сдвиги значений — это наложение вторичной и других функций на исходную. Таким образом, использование инструмента суперпозиции [2; 23] категорий дает возможность выразить функциональные свойства языковых объектов.

Функциональный подход, исследующий отношения «функциональной синонимии» разнородных и разноуровневых единиц языка, чрезвычайно актуален в настоящий момент, когда проводятся эксперименты по выявлению изофункциональных и изосемичных языковых структур из параллельных текстовых корпусов [24]. Именно этот подход позволяет найти соответствия в текстах на разных языках. В самом деле, заранее нельзя с полной достоверностью определить, каким именно образом была переведена та или иная языковая структура в текстовом корпусе. Поэтому необходимо строить и исследовать различные гипотезы при проектировании лингвистического процессора.

Отсутствие полного совпадения между английскими и русскими языковыми конструкциями в научно-технических текстах можно обнаружить при изучении сравнительной частоты употребления в них отдельных частей

речи, что важно для построения систем перевода, использующих машинное обучение. Для научного изложения в целом характерен признак номинативности, т.е. более широкое использование существительных, чем в других функциональных стилях. При этом сопоставительный анализ переводов показывает, что в русском языке эта тенденция выражена более четко, и при переводе английские глаголы нередко заменяются существительными. Проведенные нами статистические исследования параллельных текстов позволяют сделать вывод о том, что русский текст приблизительно на 35 % более номинативен, чем английский. Рассмотрим следующие примеры глагольно-именных трансформаций при англо-русском переводе.

1. The fuel system is designed to store liquid gasoline and to deliver it to the engine cylinders in the form of vapor mixed with air.

Система питания предназначена для заправки жидким топливом и подачи его в цилиндры в виде смеси паров бензина с воздухом.

to store and to deliver → для заправки и подачи.

2. A similar approach has marked the EU's efforts to expand the current club of 15 countries to embrace former communist countries further east.

Точно таким же подходом характеризуются усилия ЕС по расширению нынешнего клуба 15 стран дальше на восток путем присоединения к нему бывших коммунистических стран.

to embrace → по расширению.

Нами были проведены исследования на материале имеющихся в нашем распоряжении параллельных переводов научных статей и отдельно взятых примеров высказываний с исследуемыми конструкциями, а также мы обращались к опросу экспертов-переводчиков.

Как оказалось, в процессе грамматического разбора предложений важно учитывать вероятностные характеристики такого анализа. Рассмотрим подробнее, каким образом значения вероятности используются в процессе грамматического разбора. Например, вероятностная контекстно-свободная грамматика (PCFG – Probabilistic Context Free Grammar) и вероятностная грамматика замещения деревьев (PTSG – Probabilistic Tree Substitution Grammar) присваивают вероятность (P) каждому дереву разбора T (т.е. каждому деривату) предложения S. Эта информация является ключевой для решения неоднозначности синтаксических структур. Вероятность каждого возможного дерева разбора T определяется как произведение вероятностей всех правил r , используемых для развертывания каждого узла n в дереве разбора:

$$P(T, S) = \prod_{n \in T} p(r(n)) \quad (0.1)$$

Вероятность однозначного предложения (т.е. предложения, где нам не надо разрешать неоднозначность) равна вероятности единственного дерева разбора для этого предложения, т.е. $P(T, S) = P(T)$. Вероятность же неоднозначного предложения равна сумме вероятностей всех возможных деревьев разбора ($\tau(S)$) данного предложения:

$$P(S) = \sum_{T \in \tau(S)} P(T, S) = \sum_{T \in \tau(S)} P(T) \quad (0.2)$$

Вероятность полного разбора предложения вычисляется с учетом категориальной информации для каждой головной вершины каждого узла. Пусть n – синтаксическая категория некоторого узла n , а $h(n)$ – головная вершина узла n , $m(n)$ – материнский узел для узла n , таким образом, мы будем вычислять вероятность $p(r(n)|n, h(n))$, для этого мы преобразовываем выражение (0.1) таким образом, что каждое правило становится обусловленным своей головной вершиной:

$$P(T, S) = \prod_{n \in T} p(r(n) | n, h(n)) \times p(h(n) | n, h(m(n))) \quad (0.3)$$

В нашей системе грамматики функциональные значения языковых структур определяются категориальными значениями головных вершин. Вероятностные характеристики вводятся в правила унификационной грамматики в виде весов, присваиваемых деревьям разбора. Неоднозначные и многозначные синтаксические структуры учитываются в многовариантной грамматике когнитивного трансфера (переноса). Неоднозначность является коренным свойством естественного языка и вызывает основные затруднения при создании систем машинного перевода.

Если говорить в целом о разработке лингвистических процессов, то большую роль в результатах их функционирования имеет многоязычная лингвистическая база знаний.

Многоязычная лингвистическая база знаний создана и развивается в лаборатории Компьютерной лингвистики и когнитивных технологий обработки текстов ФИЦ ИУ РАН – ИПИ РАН на основе экспериментального лингвистического ресурса семантико-синтаксических представлений в лингвистических процессорах систем машинного перевода и обработки текстовых знаний. При разработке лингвистического процессора (на основе англо-русского и обратного трансфера) автором данной статьи [15] было предложено понятие полей функционального переноса (ПФП), явившихся базисом сегментации языковых структур для решения задач машинного перевода. Основная идея такого поля состоит в принятии гипотезы о том, что в основе грамматических структур лежат структуры когнитивные (ментальные фреймы); поле функционального переноса отражает взаимодействие элементов разных языковых уровней [14]. В основе многоязычной лингвистической базы знаний лежит разработка системы правил фразовых структур, отражающих также и отношения зависимости через механизм наследования атрибутов головной вершины. Как показано в [14; 15], этот подход более практичен с вычислительной точки зрения, и не применялся ранее для двуязычной ситуации. Функциональные значения языковых единиц закодированы как метки фразовых структур, и типы атрибутов-значений определяются функционально-категориальной семантикой. Множество языковых структур, представленных в виде синтактико-семантических комплексов, выстраиваются в иерархию правил. В данном случае это разновидность унификационно-порождающей грамматики, в которой структуры атрибутов и значений и их преобразования задаются в виде контекстно-свободных и мягко контекстно-зависимых продукционных правил. Отношения зависимости

реализуются через механизм головных вершин фразовых структур, а сами фразовые структуры задают линейные последовательности языковых объектов. В настоящий момент разрабатывается расширенная спецификация грамматики, в которой задаются категориально-функциональные признаки языковых объектов и структур и уточняются их вероятностные расширения.

ЛИТЕРАТУРА

1. Brown, R. D. Example-based machine translation in the Pangloss system / R. D. Brown // In COLING-96. – Copenhagen, 1996. – P. 169–174.
2. Козеренко, Е. Б. Логико-статистические методы представления языковых структур в машинном переводе / Е. Б. Козеренко // Тр. Междунар. конф. «Диалог'2005» «Умпыютерная лингвистика и интеллектуальные технологии». – М.: Наука, 2005.
3. Knight, K. Automating knowledge acquisition for machine translation / K. Knight. – AI Magazine 18, 1997. – P. 81–96.
4. Compendium of Translation Software / John Hutchins (editor), Seventh (edition). August, 2003.
5. Lagoudaki, E. Translation Memories Survey 2006: User's Perseptions Around TM Usage / E. Lagoudaki // In Proceedings of the Translating and the Computer 28 Conference, London, 16–17 Novem. 2006. – Aslib/IMI, London, 2006. – P. 1–29.
6. Wang, Ye-Yi. Modelling with structures in statistical machine translation / Ye-Yi Wang, Alex Waibel. – In ACL 36/COLING 17, 1998. – P. 1357–1363.
7. Alshawi, H. A comparison of head transducers and transfer for a limited domain translation application / H. A Alshawi, Adam L. Buchsbaum, Fei Xia. – In ACL 35/EACL 8, 1997. – P. 360–365.
8. Plana, E. SIMILIS Second-Generation Translation Memory Software / E. Plana // In Proceedings of the Translating and the Computer 28 Conference, London, 24–25 Novem. 2005. – Aslib/IMI, London, 2005.
9. Mel'cuk, I. A. Dependency Syntax: theory and practice / I. A. Mel'cuk. – Albany : State University of N.Y., 1988.
10. Якобсон, Р. О. Разработка целевой модели языка в европейской лингвистике в период между двумя войнами / Р. О. Якобсон // Новое в лингвистике. – М.. 1965. – Вып. 4. – С. 372–377.
11. Якобсон, Р. О. Шифтеры, глагольные категории и русский глагол / Р. О. Якобсон // Принципы типологического анализа языков различного строя. – М., 1972.
12. Степанов, Ю. С. имена, предикаты, предложения (семиологическая грамматика) / Ю. С. Степанов. – М. : Едиториал УРСС, 2004.
13. Halliday, M. A. K. An Introduction to Functional Grammar / M. A. K. Halliday; ed. by Edward Arnold. – London, 1985.
14. Halliday M. A. K. System and Gunction in Lanfuage / M. A. K. Halliday // Halliday M. A. K. Selected Papers. – London, 1976. – 250 p.
15. Слюсарева, Н. А. Проблемы функционального синтаксиса современного английского языка / Н. А. Слюсарева. – М., 1981. – 206 с.
16. Шведова, Н. Ю. Один из возможных путей построения функциональной грамматики русского языка / Н. Ю. Шведова // Проблемы функциональной грамматики. – М., 1985. – С. 30–37.
17. Звегинцев, В. А. Функция и цель в лингвистической теории / В. А. Звегинцев // Проблемы теоретической и экспериментальной лингвистики. – М., 1977. – С. 120–146.

18. Гак, В. Г. К типологии функциональных подходов к изучению языка / В. Г. Гак // Проблемы функциональной грамматики. – М., 1985. – С. 5–15.
19. Блумфилд, Л. Язык / Л. Блумфилд. – 2 изд. стереотип. – М. : Едиториал УРСС, 2002. – 608 с.
20. Щерба, Л. В. Языковая система и речевая деятельность / Л. В. Щерба. – Л., 1974. – 428 с.
21. Золотова, Г. А. Коммуникативная грамматика русского языка / Г. А. Золотова, Н. К. Онипенко, М. Ю. Сидорова. – М., 2004.
22. Шаумян, С. К. Семиотическая Лингвистика как Объяснительная Наука / С. К. Шаумян // Тр. Междунар. конф. «Диалог'2005» «Компьютерная лингвистика и интеллектуальные технологии». – М. : Наука, 2005. – С. 507–513.
23. Shaumyan, S. Sings, Mind, and Reality / S. Shaumyan. – John Benjamins Publishing Company, USA, 2006.
24. Зубов, А. В. Использование параллельного корпуса текстов для перевода несвободных словосочетаний / А. В. Зубов // Вестн. Московского гос. обл. ун-та. Серия Лингвистика. – 2011. – № 6. – С. 79–83.

М. В. Макарич

ПУТИ ДОСТИЖЕНИЯ АДЕКВАТНОСТИ ПЕРЕВОДА ТЕХНИЧЕСКИХ ТЕРМИНОВ

Рассмотрение перевода как процесса и как результата в разных аспектах, в том числе и в связи с разными, условно разграничиваемыми уровнями языка, предполагает, что объектом этого рассмотрения всегда является текст. Однако текст интересует переводоведение не только как некоеместилище тех или иных языковых явлений (фонем, слов, грамматических структур), но и как самостоятельный феномен, обладающий признаками, релевантными для перевода. Именно эти признаки позволяют переводчику выбрать общую стратегию переводческих действий для передачи смысловой организации текста. Смысл – «это не нечто аморфное, а более или менее строго организованная сущность. И адекватность перевода связана с раскрытием структуры и элементов этой сущности, иными словами, смысла как системы. Именно эта система определяет языковой стиль текста, т.е. выбор и комбинирование языковых единиц» [1, с. 5]. Исходя из этого, перевод любого текста, в особенности специального, содержащего большое количество терминов, должен иметь общий контекст, отходить от которого переводчик не имеет права во избежание смысловых потерь и искажения смыслов оригинала.

Ю. Н. Марчук в своей книге «Модели перевода» уделяет особое внимание проблеме адекватности перевода терминов. Выделяют два основных фактора [2, с. 83]:

- 1) экстралингвистические;
- 2) интралингвистические.

Экстралингвистические факторы являются следствием того, что каждый язык создает своеобразную «языковую картину мира». А это является одной из причин трудностей, возникающих при переводе. Надо отметить, что структура языка, действительно, способна определять возможные пути по-

строения сообщений, организуя определенным образом выражаемые мысли, порой навязывая говорящим обязательное употребление тех или иных форм. Но верно и то, что языковая форма высказывания не определяет однозначно содержание высказывания, выводимое на основе интерпретации значений составляющих его единиц, а служит лишь исходной базой для понимания глобального смысла. Один и тот же смысл может быть выведен из разных языковых структур, и, наоборот, одна и та же структура может служить основой для формирования и понимания сообщений. И зависимость выраженных мыслей от способа их языкового выражения оказывается относительной и ограниченной. Действующие коммуниканты могут сознавать различие между формой высказывания и сутью дела, преодолевать навязываемые языком стереотипы. Таким образом, оптимальность перевода обусловлена знанием алгоритмов «чужой» культуры, но также и столкновением ментальных пространств автора исходного текста и его переводчиков, то есть их индивидуально личностных особенностей.

Проблема текста – одна из центральных проблем переводоведения, напрямую связанная с интралингвистическими факторами: с языковой нормой, функциональными стилями, типами изложения. Именно текст является предметом анализа на первом этапе перевода, связанном с интерпретацией оригинала, и именно текст является предметом синтеза на его заключительном этапе. Анализируя исходный текст, переводчик ставит перед собой вопрос: какую цель преследует отправитель и использует для этого языковые средства? Понимание текста основывается на осознании его целостности. При этом важно не только сказанное, но и подразумеваемое: не только сказанное ранее, но и просто известное, «я» говорящего, его социальный статус, фоновые знания и т.д. В этой связи уместно вспомнить слова И. Р. Гальперина о роли подтекста, запланированного создателем текста. Выдвигаемое им положение о «содержательно-подтекстовой информации» как об органической части смыслового содержания текста, имеет самое непосредственное отношение к переводу [3, с. 42].

Основная особенность грамматического аспекта перевода состоит в жестко детерминированных отношениях между системами языка-источника и языка-приемника, которые в большинстве случаев предписывают вполне однозначное решение той или иной конкретной переводческой задачи. Возникающие в процессе перевода трансформации грамматических элементов текста являются, следовательно, системно обусловленными. Их общим свойством является то, что выбор соответствующей грамматической единицы в языке-приемнике совершается автоматически и какие-либо отклонения от предписываемых межсистемными отношениями соответствий, как правило, исключены. Так, совершенно не зависят от воли переводчика такие трансформации, как переход от единственного числа к множественному у некоторых существительных (*money* – деньги, *gate* – ворота, *watch* – часы, *contents* – содержание) и т.д. В то же время очевидно, что, несмотря на строгую детерминированность отношений между грамматическими системами двух языков, в них всегда есть элементы, допускающие известную вариативность. Эти отступления от принципа однозначных отношений

объясняются наличием синонимичных средств выражения, расхождением в структуре грамматических систем, наличием лакун в одной из них, предпочтительным использованием одного элемента в ущерб другому, взаимодействием элементов разных уровней. Так, английская личная конструкция *I'm cold* может быть переведена русской личной конструкцией «Я замерз», но если учесть предпочтение, которое русский язык отдает безличным оборотам, то более приемлемым будет другой вариант – «Мне холодно». Именно в такого рода случаях необходим выход за рамки чисто системных соответствий, ибо адекватность перевода требует здесь учета факторов, относящихся к норме.

Виды грамматических трансформаций чрезвычайно разнообразны и могут быть сведены к следующим четырем типам: добавления, опущения, перестановки, замены, что позволяет условно разделить их в соответствии с грамматическими уровнями на *синтаксические* и *морфологические*.

К *синтаксическим* можно отнести грамматические преобразования, ведущие к изменению синтаксических структур: добавления, опущения, перестановки, частично замены.

В английском языке преобладает глагольный способ выражения многих предикативных отношений, тогда как русский язык тяготеет к именному способу. Частым приемом *морфологического* преобразования является распространение исходной формы, которое заключается в расщеплении лексико-грамматической единицы на составляющие, каждая из которых несет часть исходной информации. Распространение применяется, например, для преобразования синтетических форм в аналитические.

В заключение мы можем привести основные требования, которым должен удовлетворять адекватный перевод [4, с. 113]:

- *Точность*. Переводчик обязан донести до читателя полностью все мысли, высказанные автором. При этом должны быть сохранены не только основные положения, но также нюансы и оттенки высказывания. Заботясь о полноте передачи высказывания, переводчик, вместе с тем, не должен ничего добавлять от себя, не должен дополнять и пояснять автора. Это также было бы искажением текста оригинала;

- *Сжатость*. Переводчик не должен быть многословным, мысли должны быть облечены в максимально сжатую и лаконичную форму;

- *Ясность*. Лаконичность и сжатость языка перевода, однако, нигде не должны идти в ущерб ясности изложения мысли, легкости ее понимания. Следует избегать сложных и двусмысленных оборотов, затрудняющих восприятие. Мысль должна быть изложена простым и ясным языком;

- *Литературность*. Перевод должен полностью удовлетворять общепринятым нормам русского литературного языка. Ввиду значительного расхождения в синтаксической структуре английского и русского языков, как отмечалось выше, редко оказывается возможным сохранить при переводе форму выражения подлинника. Больше того, в интересах точности передачи смысла зачастую бывает необходимым при переводе прибегнуть к изменению структуры переводимого предложения в соответствии с нормами

русского языка, т.е. переставить или даже полностью заменить отдельные слова и выражения, хотя замена даже одного слова другим весьма существенна. В переводе же не одно, а все слова заменяются другими, принадлежащими, помимо всего, к иной языковой системе, которая отличается своей особой структурой речи – порядок слов в предложении, слова, принадлежащие к одному синонимическому ряду, как правило, существенно отличаются в разных языках смысловыми оттенками.

Особенности терминов, а также их перевода с русского языка на английский и наоборот, были исследованы в работах многих авторов, в частности, Е. В. Бреуса, Л. Л. Нелюбина, В. Н. Комиссарова, [4; 5; 6]. Работы этих ученых послужили теоретической базой данного исследования. В их работах выявлены основные особенности русскоязычной и англоязычной терминологии и предложены определенные методы ее перевода. Однако техническая терминология представляет собой постоянно изменяющийся класс лексических единиц, который постоянно пополняется в связи с быстрыми темпами технического прогресса и разработкой новых технологий, и зачастую традиционные способы перевода терминов оказываются неприменимыми для передачи современных терминов. Ученые также рассматривали грамматические и стилистические особенности научных текстов, обуславливающие выбор подходов к их переводу. Тем не менее, приходится признать, что до настоящего времени особенности перевода научных текстов с английского языка на русский изучены недостаточно.

Цель данного исследования – выявить особенности при переводе англоязычной научно-технической терминологии и в связи с этим определить наиболее эффективные способы перевода английских технических терминов на русский язык.

Объектом данного исследования является научно-техническая лексика из следующих узких предметных областей: строительство, машиностроение, информатика, технология производства и связи. Тексты взяты из англоязычных периодических изданий *“Journal of Engineering & Technology”*, *“Journal of Information Technology”*, *“Architectural Design”*. При выделении 250 терминологических лексических единиц применялись два критерия – их наличие в составе узких технических предметных областей и степень их употребительности в качестве общеупотребительных слов.

Количественный анализ анализируемой лексики позволил сделать следующие выводы:

1. Наибольшее число общеупотребительных слов, имеющих терминологическое значение, являются общетехническими терминами (Polytechnic) – 93 лексические единицы, что составляет 37,2 % от общего количества исследуемой лексики.

2. Наименьшее количество общеупотребительных слов, имеющих терминологическое значение, относятся к компьютерной области (Computers) – 29 и к области телекоммуникаций (Telecoms) – 17 лексических единиц, что соответственно составляет 11,7 % и 6,9 % от общего количества лексических единиц.

3. Большое количество терминов относятся к двум и более узким техническим областям – 36 лексических единиц, что составляет 14,5 % от общего количества.

Качественный анализ исследуемой лексики показал, что:

1. В составе лексики есть слова, которые известны и понятны всем и могут быть употреблены как в устной, так и в письменной речи. Среди этих слов выделяются слова стилистически нейтральные, т.е. такие слова, которые можно встретить как в повседневном разговоре, так и прочесть в деловом техническом документе. Например: *available* – доступный, (общеупотр.), годный в соответствии со строительными нормами (строит.), *annex* – приложение (общеупотр.), пристройка (строит.).

2. Часть терминов, попадая в информационную среду, перестает функционировать только в узкоспециальном контексте. Это неизбежно оказывает влияние на терминологическую суть специальной лексической единицы и влечет за собой появление новых закономерностей функционирования терминов. Например, компьютерный термин *affirmation* – подтверждение (общеупотр.), квитиование (комп.). Первоначально данный термин обозначал подтверждение приёма-передачи структурной единицы информации. Современные пользователи интернета используют его для описания протоколов, работающих с установлением соединения между координаторами веб-сайтов для отслеживания доставки пакетов получателю.

3. В группе русских переводов англоязычной исследуемой лексики наблюдается явление транслитерации (1 случай из 250). Например, *able* – способный (общеупотр.), эйбл (комп.) – в среде системных инженеров данный термин обозначает спектр настройки интерфейса программного обеспечения.

ЛИТЕРАТУРА

1. Брандес, М. П. Стилистика текста. Теоретический курс / М. П. Брандес. – М : Прогресс-Традиция, 2004. – 416 с.
2. Марчук, Ю. Н. Модели перевода : учеб. пособие для студ. учреждений высш. проф. образования / Ю. Н. Марчук. – М. : Издательский центр «Академия», 2010. – 176 с.
3. Гальперин, И. О. Текст как объект лингвистического исследования / И. О. Гальперин. – М. : Наука, 1981. – 139 с.
4. Бреус, Е. В. Основы теории и практики перевода : учеб. пособие / Е. В. Бреус. – М. : УРАО, 2000. – 287 с.
5. Нелюбин, Л. Л. Введение в технику перевода (когнитивный теоретико-прогматический аспект) : учеб. пособие / Л. Л. Нелюбин. – 2 изд. – М. : Флинта : Наука, 2012. – 216 с.
6. Комиссаров, В. Н. Современное переводоведение : учеб. пособие / В. Н. Комиссаров. – М. : ЭТС, 2002. – 424 с.

КОМПЬЮТЕРНАЯ ЛИНГВОДИДАКТИКА

Ю. И. Горбунов

ТЕЗАУРУСНАЯ ЛЕКСИКОГРАФИЯ И КОМПЬЮТЕРНАЯ ЛИНГВОДИДАКТИКА

В настоящее время активно развивается взаимодействие тезаурусной лексикографии и компьютерной лингводидактики. Компьютерная лингводидактика представляет собой новую область знания, которую можно определить как научное направление, занимающееся вопросами компьютерной поддержки обучения языкам (Computer-Assisted Language Learning – CALL) [1; 2; 3].

Значительный опыт составления и применения тезаурусов накоплен в международной группе «Статистика речи» [4; 5; 6]. В данной группе разработана теория обучающего лингвистического автомата (ОЛА), который понимается как лингвистическая учебная среда, в которую входят лингвистические учебные курсы, универсальные системы автоматической переработки текста (АПТ), текстовые редакторы и другие программные средства, используемые при обучении иностранному языку на базе компьютера. ОЛА рассматривается как часть человеко-машинного дидактического комплекса, работающая в диалоговом режиме коллективного пользования и выполняющая следующие функции: 1) хранение и выдача учебной информации; 2) регистрация, обработка, хранение результатов учебной деятельности обучающихся; 3) анализ и оценка сообщений и ответов обучающихся; 4) управление учебной деятельностью; 5) моделирование учебных экспериментов и ситуаций [5, с. 216–217].

Теория ОЛА находится в основе научных проектов, разрабатываемых в Тольяттинском государственном университете. Так, на материале французской лингвистики конструируется ОЛА «Тезаурус лингвиста», который включает три лингвистических модуля: 1) грамматический модуль; 2) фонетико-фонологический модуль; 3) лексикологический модуль. В ходе исследования терминополья французской грамматики была выявлена семантическая область общеграмматических понятий, которые непосредственно связаны с отдельными разделами грамматики – морфологией и синтаксисом. В свою очередь, отдельные разделы грамматики представляют «терминополья», которые включают семантические области морфологических и синтаксических терминов. Перечисленные семантические области «стягиваются» в единую семантическую сеть, представляющую собой семантическое пространство или «терминополье» французской грамматики. Данное «терминополье» соотносится с соответствующими французскими терминами и их русскими эквивалентами. Каждый французский термин образует отдельное семантическое поле, которое соответствует отдельной словарной статье конструируемого тезауруса, а в совокупности словарные статьи образуют француско-русский грамматический тезаурус (ФРГТ) [7; 8; 9].

Ключевые термины сопровождаются дефинициями, которые приводятся со ссылками на научные источники (словари, монографии и т.п.). Дефиниции приводятся также и внутри словарной статьи при необходимости показать те или иные семантические отношения заглавного термина с другими терминами рассматриваемого метаязыка французской грамматики. Термины-гипонимы заглавного термина распределяются внутри словарной статьи по отдельным словарным зонам на основании семантического, формального, синтаксического и других критериев.

При составлении ФРГТ учитываются семантические отношения, которые объединяют совокупность терминов в единую систему. Данные семантические отношения образуют семантическую сеть, которая отражает структуру системы грамматических терминов. В тезаурусе данные семантические отношения представлены в фигурных скобках, например: помета {syn.} указывает на отношение синонимии, которое связывает заглавный термин с термином, перед которым эта помета стоит. ФРГТ снабжен входным индексом, который позволяет осуществлять эффективный поиск французских терминов и их русских эквивалентов в составе тезауруса.

Данная технология используется также на этапе лингвистического конструирования фонетико-фонологического и лексикологического модулей французско-русского лингвистического тезауруса, который отражает семантическое пространство французской фонетики, фонологии и лексикологии [10].

В исследовании Е. Ю. Сенницкой предлагается также тезаурусный подход, который используется с целью его последующего использования в составе комплексного ОЛА «Тезаурус лингвиста» [11; 12]. Сущность этого подхода состоит в том, что тезаурусная модель строится на принципах асимметричности, антирефлексивности и транзитивности. В данной работе французская социолингвистика моделируется с помощью особой семантической сети, структура которой позволяет учитывать не только традиционные – гиперогипонимические отношения, но и новые – специфические для терминополья французской социолингвистики семантические отношения, как-то: отношения корреляции, эквивалентности, объекта и его свойства, объекта и процесса и др.

Е. Ю. Сенницкая рассматривает французские социолингвистические термины в аспекте их семантической значимости, которая представляет собой «пучок» семантических отношений социолингвистического термина, поскольку они определяют место и статус термина в пределах социолингвистической терминосистемы. Что касается информационно-семиотической природы французской социолингвистической терминологии, то она представлена в виде тезаурусной модели метаязыка французской социолингвистики и составляет когнитивную основу информационной базы обучающего лингвистического автомата «Тезаурус лингвиста», которое рассматривается как средство оптимизации преподавания теории французского языка – французской лингвистики.

Из числа подобных работ на материале английского языка заслуживает внимания исследование Е. Ю. Горбунова. Прежде всего – опыт тезаурусного моделирования терминополья *Parts of speech* ‘части речи’ предметной области (ПО) «Английская морфология и морфосинтаксис», где подробно рассматривается специфика синергетических механизмов интенционалов терминов

[13]. В этой работе показаны два семиотических подхода к построению лингвистических классификационных моделей: формальный, ориентирующийся на означающее терминологического знака (ТЗ), и понятийно-функциональный, опирающийся на его означаемое. Первый, т.е. формальный подход, оправдывает себя при моделировании грамматических систем языков синтетического (флективно-агглютинативного) «покроя». Применительно к «мягкой» системе ПО «Английская морфология и морфосинтаксис» эффективнее использовать второй подход, опирающийся на компонентный анализ. Ведь только он позволяет выявить лексические примитивы в языке аналитического строя, определяющие положение рассматриваемой лексической единицы в терминологической системе. Компонентный анализ интенционалов терминов для обозначения частей речи схемы по П. Мэтьюзу, построенной по формальному критерию и классификационной модели, построенной, исходя из функционального и частично понятийного критериев, в версии схемы Р. Траска, позволил выявить синергетические расхождения между наборами примитивов. В качестве иллюстрации исследователь сопоставляет интенционалы трех терминов из сравниваемых таксономий, а именно *Pronoun* 'местоимение', *Adverb* 'наречие' и *Conjunction* 'союз'.

В конечном итоге исследователь приходит к выводу о том, что выявление, систематизация и унификация глубинных синергетических механизмов исследуемой ПО возможны только при компонентном анализе интенционалов моделируемых терминов с выделением набора примитивов, выступающего в качестве доказательно-экспериментального инструмента упорядочения слабо структурируемых грамматических объектов с нечетко-множественной природой. Такой семиотико-синергетический подход позволяет ослабить сложности, связанные с моделированием «мягких» грамматических систем, которые характерны для лингвистики, равно как и для других гуманитарных ПО, а также минимизировать информационные потери [14].

Таким образом, тезаурус является инструментом структурирования понятийного аппарата некой гуманитарно-социальной ПО на основании выделения понятийно-логических связей между объектами. Каждое понятие, включаемое в тезаурус, имеет строго закрепленное место в матрице терминополья, которое определяется путем анализа группы признаков. Каждый узел тезауруса заполняется понятием, закрепленным за соответствующим означающим, которым является термин или терминологическое словосочетание (ТСС). Эти термины или ТСС могут быть найдены, с одной стороны, в нормативных лексикографических источниках, а также, с другой стороны, в авторских грамматиках. В итоге, конструирование словарей-тезаурусов предметных областей гуманитарно-социальных наук может получить практическое применение в качестве дидактического комплекса в ходе проектирования инновационных образовательных программ и в частности – ОЛА «Тезаурус лингвиста» [15; 16].

В этой связи заслуживает также серьезного внимания исследование Ю. В. Ведерниковой, которое посвящено тезаурусному моделированию терминологии когнитивной лингвистики [17]. Основное внимание исследователя сконцентрировано на описании принципов организации тезауруса английской терминологии когнитивной лингвистики, к которым отнесены

ориентация и адресация тезауруса, его функциональная характеристика и уровень языка, отраженный в описании терминов. Тщательно разработанная автором структура словаря способствует систематизации терминосистемы ПО «Когнитивная лингвистика». В работе Ю. В. Ведерниковой впервые исследована английская когнитивно-лингвистическая терминология и построен тезаурус английской когнитивной лингвистики, который представляет собой сложную информационно-семиотическую модель, включающую терминопole и терминосистему английской когнитивной лингвистики, который может быть использован в качестве лингводидактического инструмента формирования лингвистической компетентности студентов-лингвистов.

В свою очередь И. И. Жучкова выдвигает гипотезу о том, что терминология ПО «Англоязычная лингвистика текста» может быть также систематизирована и унифицирована с помощью методики тезаурусного моделирования, которая обеспечивает распознавание взаимосвязей между терминами. В результате сравнительного анализа структуры и содержания словарных статей различных современных лексикографических источников по лингвистической терминологии и лингвистике текста, применив метод компонентного анализа словарных дефиниций, ей удалось построить тезаурусную модель ПО «Англоязычная лингвистика текста» в английском языке [18; 19]. Практическая значимость данного исследования состоит в том, что опыт тезаурусного моделирования дает возможность:

- 1) создать такой тип словаря, тезауруса и указателя, который как комплекс лингвистических знаний может стать вспомогательным средством для глубокого изучения и усвоения не только терминологии, но и самой предметной области в процессе профессиональной подготовки специалистов;

- 2) получить формализованную и детальную сравнительно-сопоставительную картину английской и русской терминологии современной лингвистики текста;

- 3) использовать модели, полученные в результате исследования, для систематизации других разделов языкознания;

- 4) внедрить в лексикографическую практику опыт тезаурусного моделирования при составлении терминологических словарей на материале различных языков.

В целом в данной работе дается характеристика современного метаязыка языкознания и выявляются его особенности по сравнению с естественным языком. Рассматриваются также процессы становления и современное развитие метаязыка ПО «Англоязычная лингвистика текста». Проводится обзор существующих взглядов по проблемам термина и терминологии, выбирается рабочее определение термина, используемого в исследовании. На основе информационно-семиотической теории рассматривается понятие «языкового знака» и компоненты его структуры. Выясняется, как видоизменяется модель термина-знака по отношению к общей модели языкового знака.

Кроме того особое внимание уделяется рассмотрению особенностей тезаурусного моделирования терминологии. Приводится ряд преимуществ компьютерной реализации моделируемого словаря-тезауруса, одним из которых является возможность его организации в виде гипертекстового доку-

мента. В конечном итоге моделируется с помощью тезауруса нечетко структурированная система терминологии ПО «Англоязычная лингвистика текста» с ориентиром на знаковую природу термина. С помощью применения методики компонентного анализа значений проводится компонентное описание объемов интенционалов терминов рассматриваемой ПО с целью выявления и упорядочения глубинных расхождений в их денотатах [20].

Все эти работы, несомненно, имеют явную практическую направленность: их последовательное внедрение в учебный процесс в составе ОЛА «Тезаурус лингвиста» способствует эффективному формированию профессиональной компетентности будущих бакалавров лингвистики. Несомненна также теоретическая значимость указанных работ, поскольку результаты исследования французской и английской лингвистической терминологии свидетельствуют о новых достижениях в изучении информационно-семиотической природы метаязыка современных направлений зарубежного языкознания.

ЛИТЕРАТУРА

1. *Пиотровская, К. Р.* Современная компьютерная лингводидактика / К. Р. Пиотровская // НТИ. Серия 2. Информационные процессы и системы. – 1991. – № 4. – С. 26–29.
2. *Беляева, Л. Н.* Применение ЭВМ в лингвистических исследованиях и лингводидактике : учеб. пособие к спецкурсу / Л. Н. Беляева. – Л. : Изд-во Ленингр. гос. пед. ин-та им. А. И. Герцена, 1986. – 84 с.
3. *Зубов, А. В.* Информационные технологии в лингвистике : учеб. пособие для студ. лингв. фак-тов высш. учеб. заведений / А. В. Зубов, И. И. Зубова. – М. : Издательский центр «Академия», 2004. – 208 с.
4. *Пиотровский, Р. Г.* Методы автоматического анализа и синтеза текста : учеб. пособие для ин-тов и фак. иностр. яз. / Р. Г. Пиотровский, В. Н. Билан, А. К. Боркун. – Минск : Выш. шк., 1985. – 222 с.
5. *Пиотровский, Р. Г.* Лингвистический автомат (в исследовании и непрерывном обучении) : учеб. пособие. – СПб. : Изд-во РГПУ им. А. И. Герцена, 1999. – 256 с.
6. *Беляева, Л. Н.* Лингвистические автоматы в современных информационных технологиях : монография / Л. Н. Беляева. – СПб : Изд-во Российского гос. пед. ун-та им. А. И. Герцена, 2001. – 130 с.
7. *Горбунов, Ю. И.* Французско-русский грамматический тезаурус : учеб. пособие. / Ю. И. Горбунов. – Тольятти : Изд-во ТГУ, 2003. – 152 с.
8. *Горбунов, Ю. И.* Французская грамматическая терминология: опыт тезаурусного исследования / Ю. И. Горбунов. – СПб. : РГПУ им. А.И. Герцена, 2004. – 187 с.
9. *Горбунов, Ю. И.* Тезаурусное моделирование французской грамматической терминологии : автореф. дис. ... д-ра филол. наук / Ю. И. Горбунов. – СПб. : РГПУ им. А.И. Герцена, 2004. – 28 с.
10. *Горбунов, Ю. И.* Тезаурусное моделирование метаязыка французской фонетики и фонологии / Ю. И. Горбунов // Научное мнение. – 2012. – № 9. – С. 46–52.
11. *Сенницкая, Е. Ю.* Семантическая область французских социолингвистических терминов / Е. Ю. Сенницкая // Гуманитарные, социально-экономические и общественные науки. – 2014. – № 12-3. – С. 244–247.
12. *Сенницкая, Е. Ю.* Лингвистическое конструирование тезауруса французской социо-лингвистической терминологии / Е. Ю. Сенницкая // Социология. – 2014. – № 1. – С. 144–157.

13. Горбунов, Е. Ю. Тезаурусный подход к моделированию английской морфологической терминологии / Е. Ю. Горбунов // Вестн. Самарского гос. ун-та. – 2006. – № 10-2. – С. 82–89.
14. Горбунов, Е. Ю. Тезаурусное моделирование английских грамматических терминов морфологии и морфосинтаксиса : автореф. дис. ... канд. филол. наук / Е. Ю. Горбунов. – СПб. : РГПУ им. А. И. Герцена, 2007. – 24 с.
15. Горбунов, Е. Ю. Тезаурусно-сетевое моделирование «мягкой» предметной области как инновационная гуманитарная технология / Е. Ю. Горбунов // Вестн. Орловского гос. ун-та. Серия Новые гуманитарные исследования. – 2014. – № 1. – С. 138–140.
16. Горбунов, Е. Ю. Семиотико-синергетическая природа английских морфологических терминов, обозначающих части речи / Е. Ю. Горбунов // Вестн. Сургутского гос. пед. ун-та. – 2014. – № 5. – С. 149–155.
17. Ведерникова, Ю. В. Тезаурусный подход к моделированию английской когнитивно-лингвистической терминологии (на примере семантического поля термина Cognitive Grammar (Когнитивная грамматика) / Ю. В. Ведерникова // Вектор науки Тольяттинского гос. ун-та. – 2011. – № 3. – С. 262–266.
18. Жучкова, И. И. Тезаурусный подход к моделированию английской терминологии лингвистики текста (на примере семантического поля термина «cohesion» (связность)) / И. И. Жучкова // Вестн. чувашского гос. пед. ун-та им. И. Я. Яковлева. – 2014. – № 1. – С. 53–59.
19. Жучкова, И. И. Тезаурусное моделирование терминологического поля предметной области «Text Linguistics» / И. И. Жучкова // Вестн. Волгоградского гос. ун-та. Серия 2. Языкознание. – 2014. – № 2. – С. 53–59.
20. Жучкова, И. И. О сущности и специфике использования терминов «cohesion» (связность) и «coherence» (цельность) в англоязычном и русском языкознании / И. И. Жучкова // Вектор науки Тольяттинского гос. ун-та. – 2014. – № 3. – С. 155–159.

А. А. Ефремова

ДИСТАНЦИОННОЕ ОБУЧЕНИЕ ИНОСТРАННОМУ ЯЗЫКУ ЮРИСТОВ-ЗАОЧНИКОВ

Согласно Болонской декларации студентам предоставляется больше самостоятельности в выборе изучаемых дисциплин и, как следствие, в формировании ими индивидуальной образовательной траектории.

Индивидуальная образовательная траектория формируется на основе индивидуального учебного плана, а также активизации и мотивации регулярной самостоятельной работы.

Для курсанта-заочника, будущего юриста, важно не просто усвоить новую информацию, но и научиться применять ее на практике. Такой подход требует использования интерактивных форм работы курсантов в сочетании с полноценной самостоятельной работой, что можно осуществить через систему дистанционного обучения.

Мы придерживаемся взгляда, что «дистанционная форма получения образования – это совокупность информационных технологий, обеспечивающих доставку обучаемому основного объема изучаемого материала, интерактивное взаимодействие обучаемых и преподавателей в процессе обучения, представления студентам возможности самостоятельной работы по освоению изучаемого материала» [1, с. 125].

Технологии дистанционного обучения различным дисциплинам, в том числе и иностранным языкам, становятся с каждым днем все более популярными. Форма дистанционного обучения дает возможность обучаться мобильно, используя новейшие технологии, в удобном для обучаемого (курсанта-заочника) темпе, в любое время (свободный и гибкий график обучения), в любом географически удаленном от учебного учреждения месте, там, где есть доступ к Интернету, эффективно организовывая процесс усвоения знаний и предоставляя возможность для творчества.

Рассмотрим один из способов организации обучения юристов-заочников неязыковых вузов с использованием системы управления обучением Moodle (Learning Management System – LMS), виртуальной образовательной среды, позволяющей создавать эффективные сайты для онлайн-обучения [2]. Доступ к обучению на образовательном веб-сайте курсанты получают после регистрации на этом сайте в сети.

Объектами для представления электронных учебных материалов платформы Moodle являются *Ресурсы* (файлы любого формата) и *Занятия* (интерактивные страницы), а также *Глоссарий* (словарь терминов и понятий, используемых в курсе). Контроль знаний осуществляется с помощью *Тестов* и *Заданий*, а взаимодействие с преподавателем и другими курсантами (групповые консультации, семинары) – с помощью *Форумов* (асинхронное взаимодействие) и *Чатов* (синхронное взаимодействие). При затруднениях в изучении материала используются *Рабочие тетради*.

Основными пользователями Moodle являются студенты (курсанты-заочники) и преподаватель, а также создатель курса. Студенту (курсанту-заочнику) доступно обучение с обратной связью, просмотр учебных материалов, участие в форумах и чатах, выполнение тестов, заданий, просмотр своих оценок.

Платформу Moodle можно использовать для преподавания иностранного языка при дистанционном обучении курсантов-заочников, будущих юристов. В этом случае преподавателем совместно с создателем курса разрабатываются учебные материалы, например, по курсу обучения английскому подязыку юриспруденции. Создается учебная группа из курсантов-заочников, заинтересованных в изучении данного курса. В группу могут входить сотрудники любых подразделений МВД, которые хотят повысить свой уровень владения иностранным языком.

Рассмотрим процесс обучения на примере одной из изучаемых тем *Traffic Police (GAI / DAI)*. Преподавателем и создателем курса разрабатывается и загружается на сайт учебная текстовая информация по теме *Traffic Police (GAI / DAI)*. Это может быть следующий *текст в формате текстовой страницы* с минимальным форматированием (без ссылок и иллюстраций):

Everything a traveler needs to know as a motorist (especially if he/she is a foreigner) in Belarus includes traffic regulations, car hiring and insurance.

Traffic safety in the country is controlled by the traffic police called GAI (abbr. in Russian) and DAI (in Belarusian). The uniform of Belarusian road police officers has yellow fluorescent stripes. You should pull over when a road police officer waves a striped wand or shows you a red retro-reflector. If you are pulled over by the traffic police, you will be expected to show your driving and insurance documents. ...

Это также может быть *веб-страница с иллюстрациями*, аудио и видео, ссылками на другие ресурсы (файлы, каталоги и веб-страницы).

Основной составляющей работы в Moodle является *Занятие*, посредством которого реализуется процесс программированного обучения. *Занятие* состоит из нескольких веб-страниц и таблиц переходов в виде ветвей в конце каждой страницы. Во время занятия учебный материал подается по частям, затем задаются вопросы и, в зависимости от ответов, изучение материала продолжается по той или иной ветви.

Предположим, что первое задание, которое получают курсанты-заочники по работе с текстом в режиме удаленного доступа – это *Прочестъ и перевести текст Traffic Police (GAI / DAI) с использованием сетевого электронного словаря Lingvo*. После прочтения и перевода текста они либо получают задания к данному тексту *Traffic Police (GAI / DAI)*, которые следуют сразу за текстом, например, *Закончите следующие предложения*:

1) The rules of traffic regulation in Belarus are the following: 2) If a tourist is planning a long stay in Belarus 3) Checkpoints are common on ... ,

либо задания по работе с ресурсами *Интернет*, например, *Используя энциклопедию Википедия, найдите определения термина “radar” и сокращений “DUI/DWI” и поместите их в Глоссарий*.

Найдя информацию в *Википедии*, курсанты добавляют в *Глоссарий* следующую информацию:

1. Слово – radar (noun).
2. Определение – radio wave transmission that helps police track cars and map crime scenes.
3. Пример: – According to our radar you were driving well above the speed limit.
4. Слово – DUI/DWI.
5. Определение – Driving Under the Influence/Driving while Intoxicated/Driving after drinking too much alcohol.
6. Пример – Your record says that you have two previous DUIs.

При выполнении задания на выработку навыков диалогической речи и изучение лексики и по теме *Traffic Police (GAI / DAI)*, необходимой для проведения беседы с иностранными гражданами, въезжающими на территорию Республики Беларусь – *Используя следующие вопросы, попросите водителя предъявить вам водительское удостоверение и документы на автотранспортное средство*, можно перейти по ссылкам на различные веб-страницы:

- Can I see your driver's license and passport? (<i>Ссылка №1</i>)	- Certainly. Here you are.
- Have you been drinking tonight?	- Sure. Here you are.
- How fast do you think you were going?	- I had one or two drinks. I'm okay to drive, though. I know my limit.
- Do you know what the speed limit is?	- No, sorry. Perhaps, 90 k/h.
- You were driving too fast.	- No, sorry. It's a rental car. I'm not used to it.
- You broke the speed limit.	

Осуществив переход по *Ссылке №1* (список ссылок заранее подготовлен преподавателем и создателем курса), курсанты могут получить дополнительную информацию, например, об особенностях оформления документов, удостоверяющих личность иностранцев в разных странах:

ID is different in every country

It is important to keep in mind that every country requires different documentation that proves someone's identification (ID). In many countries such as Greece and Spain a national ID card is compulsory. Citizens are expected to have these cards on them at all times. However, in other countries ID cards are not required. Police and officials in these countries often use another system to identify people, such as asking for two pieces of ID.

Information/security items that may appear on documentation

Bar code	– a series of thick and thin black lines that holds computerized information
Eye colour	– blue, brown, green, black, grey
Height	– how tall a person is in centimeters or feet and inches
Fingerprints	– markings of person's thumbs or finger tips

Курсанты-заочники обсуждают изучаемую тему через средства общения участников курса *Форум* и *Чат*. *Форум* (асинхронное взаимодействие) позволяет вести дискуссию в процессе изучения всего курса в течение нескольких дней. Например, курсантам предлагаются для обсуждения различные ситуации:

Один из Ваших знакомых собирается приехать к Вам в гости из-за границы на автомобиле. Расскажите ему:

а) какие документы необходимы при пересечении границы с Республикой Беларусь на собственной машине, или лучше ему взять машину по приезду на прокат;

б) какие правила дорожного движения он должен соблюдать на территории Республики Беларусь;

в) какие курсы по вождению автомобиля ему следует пройти, если он собирается оставаться на территории Республики Беларусь более 3-х месяцев.

Курсанты отправляют на *Форум* сообщения по обсуждаемой теме, которые могут храниться неограниченно долго. Из-за того, что форум является асинхронным видом общения, курсантам дается шанс не спешить с формулировкой ответа, дважды проверить сообщение, перед тем как его отправить.

Форумы предназначены для обмена информацией между всеми участниками процесса дистанционного обучения, а также для проведения дискуссий. Создав *Форумы* для своих курсантов, преподаватель управляет ими в процессе обучения через расписание, где он оповещает курсантов о периодичности ответов на вопросы и о контроле за сообщениями. Устанавливаются критерии оценивания активности учеников в форумах. Конечной стратегической задачей поощрения участия курсантов является самостоятельное проведение форумов.

Чат (синхронное взаимодействие) предлагает проведение как совместных обсуждений в реальном времени через *Интернет*, так и ролевых игр. Например, один участник ролевой игры – офицер ГИБДД, а другой – иностранный водитель (участники ролевой игры могут находиться в разных концах Беларуси и выходить на аудио и видео связь):

A traffic police officer (участник игры №1) pulls over a foreign driver (участник игры №2) on the highway for speeding.

Traffic police officer: Can I see your driver's license please?

Driver: Ah, yeah. I think it's in my pocket or my wallet. Uh, here!

Traffic police officer: Do you know how fast you were driving?

Driver: No, sorry. It's a rental car. I'm not used to it.

Traffic police officer: Have you been drinking tonight, Mr. Davidson?

Driver: I had one or two drinks. I'm okay to drive, though. I know my limit.

Traffic police officer: How long have you been on the territory of Belarus?

Driver: A few weeks, why?

Traffic police officer: You are unaware of our zero tolerance for drinking and driving.

Driver: I'm not drunk. I'll blow into a breathalyzer.

Traffic police officer: In Belarus you cannot operate a vehicle after consuming any alcohol.

Driver: Oh, I didn't know that.

Traffic police officer: You were also going over the speed limit, Mr. Davidson. Step out of your vehicle and get into my car. ...

Ключ для успешного проведения *Чата* лежит в регулировании преподавателем процесса общения. У каждого учебного *Чата* имеется своя цель, которая определяет содержание *Чата*. Дискуссия может проходить линейно (по четкому плану от и до), циклично (периодически возвращаясь к одному и тому же вопросу, но рассматриваемому уже с иной точки зрения) или фрагментарно (когда нет общей обсуждаемой темы, а вопросы возникают спонтанно). *Чат* также используется для проведения онлайн-консультаций с преподавателем. В этих консультациях могут принимать участие другие участники группы.

Курсанты-заочники выполняют задания в электронном виде и отправляют в виде файлов на проверку преподавателю. Проверка усвоенного материала в Moodle осуществляется через *Тесты*, включающие различные типы заданий, например, на *Сопоставление английских и русских эквивалентов*:

- 1) traffic safety
- 2) to pull over
- 3) a border crossing

- а) пограничный переход
- б) безопасность дорожного движения
- в) съехать на обочину

или на *Выбор правильных ответов*:

Traffic police officer: Welcome to the Republic of Belarus. I see you have driven to our country by your own car for the first time. I want you to pass a short test on our traffic rules. The first question: are road signs in Belarus in Cyrillic or Latin script?

Traveler: _____

- a) In Cyrillic.
- b) Road signs are in Cyrillic, although signs at major junctions are also in Latin script.

c) In Latin script.

Traffic police officer: What are the speed limits in the Republic of Belarus?

Traveler: I think _____

a) speed limits are: built-up areas 90km/h, unless specified otherwise; outside built-up areas 100km/h. The motorway speed limit is 150 km/h.

b) speed limits are: built-up areas 60km/h, unless specified otherwise; outside built-up areas 90km/h. The motorway speed limit is 120 km/h (75 mph).

c) there are no special speed limits.

Проверка усвоения материала также проводится с помощью *Рабочей тетради*. В рабочей тетради курсант-заочник высказывается на заданную тему (например, в виде доклада или реферата) или отвечает на конкретные вопросы преподавателя по изучаемой теме, а преподаватель может конфиденциально комментировать и оценивать его записи.

Одним из эффективных способов для управления элементом курса *Рабочей тетрадью* является обратная связь в интерактивном режиме, использование которой является хорошим стимулом для активного выполнения заданий. Используя обратную связь в интерактивном режиме, преподаватель поощряет курсантов творчески подходить к выполнению заданий, высказывать свои взгляды и разрабатывать идеи.

На первый план в системе дистанционного обучения иностранному языку выходят педагогическая и содержательная его организация. Это не только отбор учебного материала определенного содержания для усвоения, но и его структурная организация и методы обучения.

Дистанционное обучение иностранному языку предполагает тщательное и детальное планирование деятельности юристов-заочников, организацию этой деятельности, четкую постановку задач и целей обучения, доставку необходимых учебных материалов, которые должны обеспечивать систематическую и эффективную интерактивность между курсантом-заочником и преподавателем, обратную связь между курсантом-заочником и учебным материалом, предоставление возможности группового обучения.

При этом реализуются дидактические принципы организации обучения иностранному языку, а именно: принципы научности, системности и систематичности, активности, принципы развивающего обучения, наглядности, дифференциации и индивидуализации обучения. Процесс реализации дидактических принципов обусловлен спецификой новой формы дистанционного обучения, возможностями информационной среды Интернет, огромным спектром ее услуг.

Выбрав дистанционную форму обучения иностранным языкам, курсант-заочник может без отрыва от службы в органах внутренних дел самостоятельно совершенствовать свой уровень владения иностранными языками.

ЛИТЕРАТУРА

1. Андреев, А. А. Дистанционное обучение: сущность, технология, организация / А. А. Андреев, В. И. Солдаткин. – М. : Изд-во МЭСИ, 2000. – 350 с.
2. Брезгунова, И. В. Программная платформа LMS Moodle / И. В. Брезгунова, С. И. Максимов, В. М. Шульганова. – Минск : Изд-во РИВШ, 2010. – 52 с.

**СОЗДАНИЕ И РАЗВИТИЕ СРЕДСТВ ЭЛЕКТРОННОЙ ПОДДЕРЖКИ
ОБУЧЕНИЯ ИНОСТРАННОМУ ЯЗЫКУ
НА БАЗЕ ВЫЧИСЛИТЕЛЬНОГО ЭКСПЕРИМЕНТА**

Любое прикладное программное средство разрабатывается применительно к ситуациям, которые оно должно обслуживать. А ситуации, в свою очередь, определяют цель и ограничения на постановку задачи, решение которой это программное средство должно реализовывать. Достаточно очевидно, что электронная поддержка иностранным языкам, как компонент системы языкового обучения в вузе, чаще всего трактуется как атрибут самостоятельной работы студента. Абсолютно очевидно, что содержание этой поддержки, значительная часть которой составляет выполнение упражнений типа «стимул – реакция», должно быть согласовано по содержанию с аудиторной работой и скоординировано с ней по времени.

Не менее важным аспектом является управление самостоятельной работой студента с электронной поддержкой. Из множества стоящих перед ним задач студент в первую очередь выбирает те, которые четко очерчены в своих границах, оставляя выполнение остальных, более сложных и не понятых им, задач «на потом». Ему необходимы программы, которые бы брали управление процессом обучения на себя, сводя его роль в ходе обучения к последовательному выполнению упражнений, предлагаемых ему программой, и ожиданию сообщения о том, что сеанс обучения успешно завершен. Естественно, что реализуемое программой управление процессом должно гарантировать требуемый результат, т.е. выполнение целей обучения.

Начиная с замечательной работы Л. А. Растригина и М. Х. Эренштейн [1], в качестве целей обучения, как правило, рассматриваются две. Первая определяет желаемое состояние обучаемого (или группы обучаемых) на момент завершения процесса обучения. В качестве второй цели, которая является экстремальной, рассматривается время достижения желаемого состояния обучаемого.

С позиции преподавателя, ведущего практический курс иностранного языка, желаемое состояние обучаемого может быть представлено неравенством, которое задает нижнюю границу доли неверных ответов обучаемых в тестовой работе, проводимой по окончании обучения

$$\frac{k}{m} \leq q,$$

где m – число стимулов изучаемых лексических единиц, представленных в тестовой работе, k – количество ошибочных ответов (или реакций) обучаемого, а q – пороговое значение допустимой доли таких ответов.

Выраженная в такой форме цель обучения естественна для языка преподавателя и легко проверяема. Однако в плане организации управления как метода решения задачи обучения, где целевое неравенство выступает как условие или ограничение, такая форма выражения цели неприемлема.

В упомянутой выше работе [1], посвященной решению задачи бесконтекстного изучения иностранной лексики, ее авторы описывают состояние обучаемого (по отношению к объекту обучения) в форме вектора

$$p(t) = (p_1(t), p_2(t), p_3(t), \dots, p_{n-1}(t), p_n(t)),$$

где n – общее число изучаемых лексических единиц, $p_i(t)$ – функция, позволяющая прогнозировать значение вероятности ошибочной реакции обучаемого на стимул i -ой лексической единицы (или просто вероятность ее незнания) на момент времени t . Вид или структура функций $p_i(t)$ полагается одинаковой и трактуется как модель обучаемого. Перевод задачи на язык теории вероятностей, принятый в стохастической теории обучения, естественным образом преобразует целевое неравенство к виду

$$Q(p) = \sum_{i=1}^n p_i < q,$$

где $Q(p)$ трактуется как показатель (или критерий) качества обучения, а q – как его пороговое значение (или критерий завершения процесса обучения).

Попытка решения представленной в такой форме задачи обучения порождает два ключевых вопроса. Первый связан с идентификацией модели обучаемого, как структурной, так и параметрической, а второй – с реализацией связи между двумя обозначенными выше целями.

В качестве альтернатив для выбора модели обучаемого при решении задачи бесконтекстного изучения иностранной лексики рассматривались «относительно близкие» по своей сути модели, ранее разработанные в рамках исследований по экспериментальной психологии, а для выбора наилучшей – использовалась методика Р. Буша и Ф. Мостеллера [2]. Для подавляющего большинства реальных задач по изучению грамматики иностранного языка такие модели отсутствуют. Здесь необходимы (по крайней мере, на данном этапе) частные исследования, связанные с изучением процессов обучения различным языковым явлениям.

Что же касается второго вопроса, то учитывая опосредованный характер связи между целью, выраженной неравенством $Q(p) < q$, и временем ее достижения, поиск решения задач, подобных данной, обычно переводится на язык стратегий. Другими словами, стараются отгадать, «какую локальную задачу, связанную с каждым отдельным сеансом, должна решать обучающая система в ходе его проведения, чтобы стратегия, построенная на последовательном решении локальных задач, была «интуитивно близка» к гипотетически оптимальной стратегии» [3]. При этом успешность поиска естественным образом зависит от наличия априорной информации об исследуемом процессе.

В теории управления взаимную обусловленность цели и модели процесса иногда называет парадоксом цели. В работе [1] парадокс цели для сформулированной задачи компьютерного обучения формулируется следующим образом. «Для выбора стратегии решения задачи и создания обучающей программы, обеспечивающей ее реализацию, необходима модель обучаемого, для построения которой необходимо с помощью той же самой обучающей программы провести массовый эксперимент, собрать и обработать его результаты».

Одним из путей разрешения парадокса цели в условиях отсутствия априорных данных об исследуемом процессе является вычислительный эксперимент. В методологическом плане под термином вычислительный эксперимент понимается «циклический процесс, реализующий эволюционную стратегию построения модели исследуемого объекта или процесса, каждую

итерацию которой составляют три последовательно реализуемые стадии: анализ, синтез и оценка, а проверка адекватности модели, синтезированной в ходе каждой очередной итерации, осуществляется путем сравнения результатов натуральных экспериментов с результатами, полученными с помощью компьютера путем вычислений по модели» [4].

Для реализации вычислительного эксперимента еще в 2006 году на филологическом факультете РГПУ им. А. И. Герцена была разработана и в последующие годы активно развивалась программа языкового компьютерного обучения и тестирования, реализующая четыре основных типа письменных упражнений типа «стимул – реакция». Разработанная программа снабжена механизмом сбора данных о реакциях обучаемого в ходе обучения с целью: а) последующей обработки полученной информации и формирования новых знаний о процессах компьютерного обучения иностранным языкам; б) использования этих знаний для постановки новых задач исследования, создания новых и корректировки существующих алгоритмов управления процессами обучения.

Воплощенное в структуре программы решение позволило отделить:

а) те части программы, с помощью которых реализуются операции выполнения языковых упражнений, от тех ее частей, которые реализуют управление процессом обучения;

б) инвариантные (по отношению к реализуемому виду языковых упражнений) части программы от неинвариантных.

А это, в свою очередь, на достаточно долгий период времени обеспечило необходимые условия для нормального сопровождения и развития программы, т.е.:

а) открытость системы в плане включения в ее состав функциональных блоков, реализующих новые типы языковых упражнений;

б) достаточно простую (т.е. без сколько-нибудь существенных перестроек программы) замену одних алгоритмов управления процессом обучения другими.

Осуществление вычислительного эксперимента подразумевает, что в любой момент времени программа работает на основе алгоритма, соответствующего знаниям и представлениям о процессе обучения. Использование программы сопровождается сбором данных о ходе процесса обучения. Обработка полученных данных служит основой для корректировки существующего алгоритма или его замены новым алгоритмом.

Обеспеченная структурой программы отделимость тех ее частей, которые реализуют управление процессом обучения от частей, которые реализуют операции выполнения языковых упражнений, позволяет выполнить замену одного алгоритма управления другим достаточно «безболезненно», т.е. практически не затрагивая остальных частей программы.

Более того, то обстоятельство, что доля программного кода, которая приходится на реализацию алгоритма управления, достаточно невелика, позволяет одновременно использовать и сопровождать несколько программ, использующих разные алгоритмы управления, т.е. вести исследования по широкому спектру задач, составляющих объем проблемы.

В организационном плане вычислительный эксперимент рассматривается как часть высокотехнологичной образовательной среды, в рамках

которой осуществляется интеграция учебной и созидательной деятельности студентов, направленная на создание компонентов информационного и методического обеспечения систем языкового компьютерного обучения, с одной стороны, с другой, – на то, чтобы в процессе этой деятельности студенты научились: а) создавать (на базе существующей программы) собственные системы компьютерного обучения; б) выполнять на их базе научные и научно-педагогические исследования; в) обрабатывать полученные данные о ходе обучения, интерпретировать результаты.

Объектом одного из таких исследований являлось обучение лексике английского научного текста на базе алгоритма, в основе которого лежит гипотеза голографического устройства человеческой памяти. Общая идея алгоритма состояла в том, чтобы вынуждать обучаемого выучивать «правильные реакции» в контексте стимулов заданий, создавая тем самым в его памяти ядро контекстов (или примеров), образующих некий коллективный образ изучаемого языкового явления, который впоследствии используется (путем «наведения соответствующих ассоциативных связей») при выполнении последующих упражнений, формировании фрагментов речи и текста. Детальное описание алгоритма приведено в статье [5].

Не останавливаясь на детализации выполненной работы, отметим лишь результаты эксперимента, которые позволяют дать предварительную оценку разработанного алгоритма управления.

В эксперименте принимала группа из 50 студентов четвертого курса филологического факультета. Материал изучения включал 100 языковых клише (устойчивых словосочетаний, вводных конструкций и т.п.). Так как реализация алгоритма обеспечивает полную передачу функций управления обучающей программной системе, роль участников эксперимента в ходе сеанса сводилась к выполнению заданий, предлагаемых ему программой, и ожиданию сообщения о том, что сеанс обучения успешно завершен. После достижения обучаемым заданного критерия качества обучения, которое отслеживалось программой, он может быть допущен к контрольному тестированию. Технически это реализовывалось путем блокировки пункта меню «Обучение» и снятия блокировки с пункта меню «Контрольное тестирование».

В результате обработки данных были получены следующие результаты. Среднее число предъявлений стимула (по группе в целом), которое потребовалось для завершения обучения (по одному упражнению), оказалось равным 4,64. Разброс этого показателя для отдельных обучаемых достаточно велик: от 2.72 до 7.81, стандартное отклонение – 1.57.

Среднее число сеансов, в ходе которых эти предъявления реализовывались, оказалось достаточно малым – всего 1.91. Это означает, что «выучивание» правильного ответа в среднем осуществлялось примерно за два сеанса так, что на третьем сеансе обучаемый мог воспроизвести правильный ответ с первого раза. Разброс значений этого показателя так же достаточно мал: от 1.51 до 2.37, стандартное отклонение – 0.32.

Результаты обучения оценивались в ходе контрольного тестирования, к которому обучаемый допускался программой после полного завершения процесса обучения. Несмотря на то, что процедура тестирования была достаточно утомительной (она проводилась по всему материалу обучения,

а не по выборке из него, как это делается обычно), результаты обучения оказались более чем оптимистичными. Доля правильных ответов для разных обучаемых оказалась в диапазоне от 0.61 до 0.9, среднее значение – 0.779, стандартное отклонение – 0.0832.

В настоящее время ведутся совместные работы филологического факультета, лаборатории инженерной лингвистики и института компьютерных наук и технологического образования по созданию сетевой версии описанной выше электронной поддержки и разработанного программного обеспечения.

ЛИТЕРАТУРА

1. *Растригин, Л. А.* Адаптивное обучение с моделью обучаемого / Л. А. Растригин, М. Х. Эренштейн. – Рига : Зинатне, 1988. – 160 с.
2. *Буш, Р.* Сравнение восьми моделей / Р. Буш, Ф. Мостеллер // Математические методы в социальных науках. – М. : Физматгиз, 1973.
3. *Беляева, Л. Н.* Автоматизированное рабочее место филолога в структуре образовательного пространства вуза / Л. Н. Беляева, Т. Л. Джепа, Г. Н. Зак, О. Н. Камшилова, В. Р. Нымм, В. В. Разумова // Монография. – СПб : Книжный Дом, 2013, – 127 с.
4. *Нымм, В. Р.* Эволюционный подход к созданию и развитию CALL-технологий на базе вычислительного эксперимента / В. Р. Нымм // Изв. Рос. гос. пед. ун-та им. А. И. Герцена. – 2015. – № 173. – С. 137–147.
5. *Нымм, В. Р.* CALL – технологии: управление процессом обучения языку / В. Р. Нымм, К. Р. Пиотровская, В. А. Быстрая // Высокотехнологичная образовательная среда : сб. ст. Междунар. науч.-практич. конф. – СПб. : ООО «Книжный дом», 2015. – С. 172–179.

К. Р. Пиотровская

ТЕХНОЛОГИИ ТЕКСТ-МАЙНИНГА КАК ЭЛЕМЕНТ ИНФОРМАЦИОННОЙ И МАТЕМАТИЧЕСКОЙ КУЛЬТУРЫ СТУДЕНТОВ-ФИЛОЛОГОВ

Неуклонное реформирование российской школьной и вузовской системы привело к неожиданным результатам. С одной стороны, повсеместное закрытие кафедр прикладной лингвистики, сокращение числа ученых советов по специальности **прикладная** и математическая **лингвистика**, отсутствие должной подготовки в области информационных технологий и математических основ обработки информации, а также базовых знаний в области высшей математики, а также изолировали многих филологов-исследователей от возможности грамотно, на доказательном, современном уровне проводить профессиональную исследовательскую работу. Обострению создавшейся ситуации способствовал объективный перенос исследовательских работ с чисто лингвистических площадок на математические и информационные, что было связано с коммерциализацией и стремительным развитием аппарата информатики, математики и квантитативной лингвистики. А высоты, достигнутые в области популяризации применения аппарата математики и теоретической информатики в филологических исследованиях советскими серьезными исследовательскими группами, такими как Всесоюзная группа «Статистика Речи», сегодня во

многим утрачены: хорошие учебники по математике для лингвистов давно не переиздавались в России и постепенно превратились в библиографический раритет [1; 2; 3].

Хотя значительная интенсификация информационного пространства, увеличение его объемов, рассеяние, дублирование и быстрое старение информации усиливают роль информационной и математической культуры для современного студенчества, особенно для студентов-филологов, чья деятельность неразрывно связана с поиском и осмысленной переработкой информации, в то же время, опыт преподавателей свидетельствует, что обращение студентов-филологов к электронным информационным системам и Интернет-ресурсам с целью извлечения знаний путем поиска, синтеза и семантической обработки текстовой информации зачастую имеет негативные результаты. Это связано, в первую очередь, с отсутствием должной подготовки в области оснований математики и математической статистики, а также навыков использования средств, которые предоставляет современная наука для решения этих проблем. К одним из востребованных современной исследовательской деятельностью средств относят технологии дата-майнинг (Data Mining) и текст-майнинг (Text Mining). Эти технологии, основаны на аппарате математической статистики, искусственного интеллекта, экспертных систем, нейронных сетях и др. и предназначены для выявления в текстах скрытой от непосредственного наблюдения информации и ранее неисследованных закономерностей. Оформившись в конце XX века, как направление анализа неструктурированной текстовой информации, технология текст-майнинг стала логическим продолжением дата-майнинга и объединила в себе как классические методы извлечения данных (например, кластеризация) так и методы контент-анализа, статистического анализа и др. [4]. Принципиальное отличие технологии текст-майнинг от дата-майнинга заключается в том, что последняя работает с базами данных, в то время как текст-майнинг позволяет исследователю анализировать обычные тексты, представленные на естественном языке.

На практике использование технологий глубинного анализа текстов открывает для студентов-филологов следующие возможности:

- мониторинг ресурсов Интернет (контент-мониторинг), семантический поиск информации в Интернет и существенное сужение границ поиска за счет включения методов текст-майнинга в современные поисковые системы;
- создание семантических сетей текстов больших объемов, реферирование, классификация и кластеризация текстов, поиск по тексту, интегрирование неструктурированной текстовой информации с существующими структурированными данными, наглядная визуализация кластеризированной текстовой информации.

Довольно большой набор программных продуктов, предоставляет как пробные бесплатные, так и свободно распространяемые версии для проведения исследований:

- QDA Maining, входящий в известный пакет Word Stat, разработанный канадской исследовательской группой Provalis Research, <http://provalisresearch.com/>;
- RapidMaining, разработка Дортмундского технического университета, Германия, <http://rapidminer.com/>;

- Carrot2, разработка Познаньского технического университета, Польша, <http://project.carrot2.org/>;
- Gate, разработка Шеффелдского университета, Англия, <https://gate.ac.uk/>;
- семейство программ открытого доступа AntConc, разработка Японского университета г. Васеда, <http://www.laurenceanthony.net/~software/antconc/>;
- Textometrica, разработка университета Умео, Швеция, <http://www.simonlindgren.com/textometrica>;
- Juxta, разработка университета штата Вирджиния, США, <http://www.juxtasoftware.org/>;
- SEASR, разработка университета в Иллинойсе, США, <http://www.seasr.org/>;
- TAPoR Tools, разработка университета Альберта, Канада, <http://www.tarpor.ca/>;
- Textpresso, разработка технологического университета Калифорнии, США, <http://www.textpresso.org/>;
- Vivisimo/Clusty, разработка университета Карнеги Меллона, США, <http://yippyinc.com/about/>;
- VisualText, разработка калифорнийской инкорпорации анализа текста, США, <http://www.textanalysis.com/>;
- Word Hoard, разработка Северо-западного университета, США, <http://wordhoard.northwestern.edu/userman/index.html>;
- WordSmith, разработка Оксфордского университета, Англия, <http://www.lexically.net/wordsmith/index.html>;
- свободные текст-майнинговые библиотеки для языков программирования Java, Python, R и C++.

Разработка методики применения этих программных продуктов может послужить основой для составления новых современных учебных пособий по практическому применению студентами-филологами аппарата количественной лингвистики, текст-майнинга и автоматической переработки текста.

Активизация их применения в учебной и исследовательской деятельности позволит студентам-филологам ориентироваться в больших информационных потоках, осуществлять более плодотворную деятельность, связанную с поиском, анализом, синтезом электронной текстовой информации, оценивать ее полезность с меньшими временными и энергетическими затратами, формировать систему информационных понятий. Данные обстоятельства делают технологии текст-майнинга необходимым элементом информационной культуры современного студенчества, что обуславливает необходимость преподавания основ этого направления на филологических факультетах.

ЛИТЕРАТУРА

1. *Пиотровский, Р. Г.* Математическая лингвистика / Р. Г. Пиотровский, К. Б. Бектаев, А. А. Пиотровская. – М. : Высш. шк., 1977. – 383 с.
2. *Лесохин, М. М.* Введение в математическую лингвистику / М. М. Лесохин, К. Ф. Лукьяненко, Р. Г. Пиотровский. – Минск : Наука и техника, 1982.
3. *Пиотровский, Р. Г.* Методы автоматического анализа и синтеза текста / Р. Г. Пиотровский, В. Н. Билан, А. К. Боркун. – Минск : Выш. шк., 1985.
4. Miner G et al Practical Text Mining and Statistical Analysis for Non-structured Text Data. – N.Y. : Elsver, 2012.

ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ ГИБРИДНЫХ СИСТЕМ РАСПОЗНАВАНИЯ РЕЧИ (ПРИМЕНИТЕЛЬНО К ОБУЧАЮЩИМ СИСТЕМАМ)

Автоматическое распознавание речи – междисциплинарная область, в которой работают специалисты различного профиля: инженеры, программисты, лингвисты, математики и многие другие. Подходы к решению этой задачи обусловлены как требованиями к разрабатываемому продукту, так и компетенцией разработчиков, включая лингвистов [1; 2].

Широко применяемыми методами распознавания речи являются скрытые модели Маркова, которые доказали свою эффективность и гибкость при реализации систем распознавания речи. Обычно используются три базовых алгоритма, основанные на скрытых марковских моделях: алгоритм прямого распространения (применяется для распознавания изолированных слов); алгоритм Витерби, применяемый для распознавания слитной речи; алгоритм прямого – обратного распространения, используемый для тренинга моделей, в частности, интонационного и звукового строя различных языков.

Сложность алгоритма поиска зависит от природы пространства, в котором производится поиск, и, следовательно, от ограничений, накладываемых на лингвистические модели. При поиске используются следующие алгоритмы: общий поиск на графе; поиск в глубину; поиск в ширину; эвристический лучевой поиск; иерархический поиск. В задачах, связанных с распознаванием и соотносением образов, чаще всего используется иерархический вид поиска как позволяющий находить минимальные различия, и следовательно, подбирать максимальные соответствия для различных языковых уровней.

Общая схема работы системы распознавания речи выглядит следующим образом (см. рис. 1).

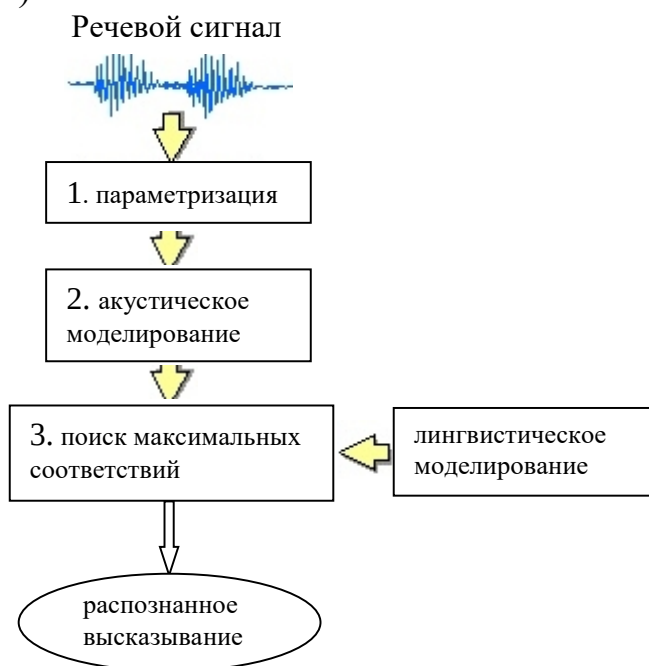


Рис. 1. Принципиальная упрощенная схема работы системы распознавания устной речи

В настоящее время активно разрабатываются алгоритмы распознавания речи, основанные на применении нейронных сетей. Теория нейронных сетей в целом строится на гипотезе, что сложные операции вычисления, выполняемые в рамках одного высокоэффективного модуля системы, можно заменить на ряд параллельно выполняемых простейших операций, производимых в отдельных компонентах системы. Все нейронные сети работают по примерно одинаковым принципам: они состоят из узлов (нейронов), куда поступают данные либо от входов сети, либо от других узлов сети. Входящие сигналы суммируются и посылаются другим элементам по взвешенным связям. Весовой коэффициент ослабляет или усиливает сигнал, идущий по данной связи от одного элемента к другому. Таким образом, нейронная сеть представляет собой соединенные друг с другом простейшие процессоры (элементы, нейроны, узлы), выполняющие элементарные преобразования с входящими сигналами с учетом весовых коэффициентов (коэффициентов, характеризующих связь между элементами), и формирующих выходной сигнал, который поступает на другие элементы, связанные с ним. Каждая связь имеет свой весовой коэффициент. Если он положительный, то входящий сигнал усиливается, в противном случае – подавляется.

Преимущество такого подхода состоит в том, что количество разных функций, по которым вычисляются активность и суммарный вход, ограничено, а диапазон задач, которые могут решать нейронные сети, достаточно велик. Настройка сети на решение новой задачи состоит только в изменении архитектуры сети. Под архитектурой понимается топология сети, структура связей, правила распространения сигналов в сети, правила комбинирования входящих сигналов, правила вычисления активности, правила обучения, т.е. изменения весовых коэффициентов. Как показал ряд исследований, нейронные сети могут быть достаточно эффективны для обработки речи и, в частности, для распознавания речи. Среди задач, с которыми наиболее эффективно справляются нейронные сети, в первую очередь следует назвать классификацию и распознавание образов. Распознавание речи есть приложение теории распознавания акустических и перцептивных слуховых образов [3; 4; 5; 6; 2]. Определенный образ, который может представляться в качестве сегмента речевого сигнала или набора акустических параметров, характеризующих этот сегмент, должен быть отнесен к классу, которому он принадлежит (в качестве класса в рамках различных подходов может быть взята отдельная реализация фонемы, дифон или трифон, слог, или другой элемент, на основе которого выделяются характеризующие параметры класса), что есть задача классификации. Следовательно, алгоритмы распознавания речи, основанные на применении нейронных сетей, могут быть достаточно эффективны. Алгоритмы распознавания речи требуют выполнения огромного объема сложных вычислений. Нейронная сеть, моделируя работу нервной системы человека, сводит этот массив сложнейших вычислений к набору параллельно выполняющихся простейших операций, что ускоряет обработку входящей информации и принятие решения.

Еще одно преимущество использования нейронных сетей для распознавания речи заключается в их эффективности при решении задач классификации и кластеризации. Под классификацией имеется в виду прямое обу-

чение сети – относить ли тот или иной образ к определенному классу в соответствии с инструкциями. То есть выходной образец дает всей сети информацию о том, к какому классу следует научиться относить образец. Если же таких инструкций нет при обучении без управления, то сеть должна научиться группировать образцы самостоятельно. В этом случае принято считать, что сеть должна научиться проводить кластеризацию образов самостоятельно. Преимущество нейронных сетей заключается в том, что они могут сегментировать образцы на группы даже в том случае, если невозможно разделить образцы на классы с помощью линейной зависимости. Задача линейной классификации образцов легко поддается алгоритмизации. Однако при обработке устной речи, к сожалению, такого идеального варианта не встречается. Приходится иметь дело с нелинейными зависимостями. А если к этому добавить проблему динамического изменения значений элементов векторов признаков и размера векторов признаков во времени, то алгоритмизация из сложной превращается в практически невозможную.

В качестве наиболее практичного подхода к распознаванию слитной речи в настоящее время принято использовать **гибридные системы**, представляющие интеграцию нейронных сетей в рамках подхода, основанного на марковских статистических моделях.

Гибридные сети имеют ряд преимуществ при разработке систем распознавания речи. Во-первых, повышается точность моделей, так как нейронные сети формируют более точные акустические непараметрические модели, так как предположения о форме дистрибуции в таких моделях отсутствуют (а это один из основных источников ошибок при акустическом моделировании нейронными сетями). Во-вторых, гибридные сети отличаются повышенной чувствительностью к контексту. Цепи Маркова исходят из предположения о том, что участки обрабатываемого сигнала не зависят друг от друга, и каждый фрейм обрабатывается изолированно. Для того, чтобы иметь возможность использовать информацию о коартикуляции, интерференции и т.д. (т.е. о взаимовлиянии соседних сегментов речи друг на друга), необходимо искусственно расширять границы фрейма и включать в обрабатываемый сегмент соседние. Нейронные сети, с другой стороны, способны естественным образом обрабатывать сегменты любого размера (лишь бы они были одинаковые). Поэтому нейронные сети более чувствительны к фонетическому контексту (окружению). В-третьих, нейронные сети лучше разделяют образцы по группам, ибо классификация и кластеризация входных данных – естественное свойство нейронных сетей. В-четвертых, нейронные сети имеют преимущество в том случае, если данных для тренировки недостаточно, так как им необходимо меньше параметров для деления образцов на классы [7].

На эффективность и точность работы распознающей и обучающей системы влияют следующие факторы [8; 9; 10]: размер словаря и наличие в нем схожих слов (число ошибок, как правило, сильно увеличивается с увеличением словаря). Наличие слов, которые легко спутать, также представляют определенную проблему; манера произнесения: система может быть нацелена на распознавание изолированно произнесенных слов, пословной речи, когда каждое слово в предложении отделяется паузой, или слитной (естественной) речи. Последний случай наиболее сложный, но он же – необ-

ходимое условие, если мы хотим построить обучающую систему, занимающуюся не только дриллингом отдельных сегментов речи в изолированном окружении, но именно обучением произношению как на сегментном, так и на супraseгментном уровнях в естественной речи; темпоральная вариативность: проблема вариативности в темпе произнесения и в определении момента начала артикуляции очередного сегмента (на темп влияет время артикуляции ударных слогов и время физических пауз); акустическая вариативность – проблема различий в произношении (акцент, индивидуальные особенности речи, громкость, акустические внешние условия и т.д.).

Проблема темпоральной вариативности решается с помощью алгоритмов динамического программирования, чаще всего используются DTW и алгоритм Витерби. Вариативность, которую приходится учитывать при разработке систем автоматического распознавания речи, вызвана междикторской, внутридикторской, ситуативной и т.д. вариативностью речи и вариативностью сигнала, в том числе искажениями сигнала [1].

В настоящее время уровень систем автоматического распознавания речи таков, что точное определение места артикуляции того или иного звука имеет первостепенное значение, поэтому распознавание иноязычной речи не настолько надежно, как распознавание речи говорящего на родном языке диктора.

Для моделирования иностранного произношения в обучающих системах предлагаются следующие подходы: использование репрезентативных массивов данных для обучения системы; применение независимых от диктора моделей по алгоритму Витта и Янга; проведение лингвистического моделирования; расширение словаря за счет вариантов произнесения после обработки правил произношения в родном языке диктора и выявления закономерностей интерференции; использование звукового словаря специфических иноязычных акцентов; классификация иноязычных акцентов на основе их акустических характеристик.

Специальная целенаправленная фонетическая практика позволяет избавиться от возможных произносительных ошибок, увеличить стабильность реализации отдельных звуков в определенных контекстах.

Таким образом, для разработки обучающих фонетических систем наиболее продвинутого вида оптимальным способом является метод повышения устойчивости к иноязычному акценту. При его применении система способна распознать интерферируемую речь, и в то же время, используя алгоритм Витта, выделить ошибку реализации речевого сегмента, после чего возможно применение блока формирования правил коррекции произнесения [8; 9; 1].

Таким образом, для повышения эффективности работы, например, лингвистической обучающей системы (фонетический аспект) представляется целесообразным рассмотреть вопрос использования не одного, а комплекса алгоритмов для реализации базового метода распознавания речи. Инкорпорирование элементов нейронных сетей, а также гибридного метода для распознавания и обучения иноязычной речи представляется более перспективным, нежели использование традиционных методов. Для моделирования иноязычного акцента оптимальным является расширение словаря за счет вариантов произнесения после обработки общих правил произношения на родном языке обучающегося и выявления закономерностей интерязыковой интерференции.

ЛИТЕРАТУРА

1. *Потапова, Р. К.* Приоритетные направления современной прикладной лингвистики / Р. К. Потапова // Конверсия в машиностроении. – 2004. – № 3–4. – С. 92–100; 97–103.
2. *Потапова, Р. К.* Новые информационные технологии и лингвистика / Р. К. Потапова. – 10 изд. – М., 2014.
3. *Потапова, Р. К.* Звучащая речь как объект исследования в фундаментальной и прикладной лингвистике / Р. К. Потапова, В. В. Потапов // Ежегодник «Акустика речи и прикладная лингвистика»: сб. тр. РАО и МГЛУ. – М., 2002. – С. 6–28.
4. *Потапова, Р. К.* Речь: коммуникация, информация, кибернетика / Р. К. Потапова. – 4 изд. доп. – М., 2010.
5. *Потапова, Р. К.* Основы речевой акустики / Р. К. Потапова, В. Г. Михайлов. – М.: Рема, 2012.
6. *Потапова, Р. К.* Речевое управление роботом / Р. К. Потапова. – 2 изд., доп., перераб. – М., 2012.
7. *Потапова, Р. К.* Междисциплинарность в исследовании речевой полиинформативности / Р. К. Потапова, В. В. Потапова, Н. Н. Лебедева, Т. В. Агибалова. – М.: Языки славянской культуры, 2015.
8. *Потапова, Р. К.* Некоторые подходы к реализации речевых компонентов в компьютерных лингвистических обучающих системах / Р. К. Потапова, М. Ю. Ордин // Акустика речи. Медицинская и биологическая акустика: матер. 13 сессии Рос. акустического об-ва. – М., 2003. – Т. 3. – С. 145–149.
9. *Potapova, R. K.* Articulation Models in Educational Software with Embedded ASR Components / R. K. Potapova, M. Yu. Ordin // Proceedings of Intern. conf. “SPECOM’ 2003”. – Moscow, 2003. – P. 360–364.
10. *Potapova, R. K.* Modern CALL systems with elements of acoustic feedback / R. K. Potapova // Proceedings of Intern. conf. “SPECOM’ 2003”. – Moscow, 2003. – P. 53–60.

Роман Хайчук

КОМПЬЮТЕРНАЯ ЛИНГВОДИДАКТИКА

(несколько указаний по вопросу о глоттодидактическом приоритете)

Овладеть всерьез иностранным языком (не в порядке билингвизма), даже родственным или близкородственным – задача из самых трудоемких. Ее осуществление требует затраты неимоверного количества времени и усилий, а также хорошей организации языкового материала в процессе обучения. Программы преподавания языков в школах и вузах, предусматривающие уроки размером в общей сложности даже в несколько сот часов в больших, да и в небольших, группах (не говоря уже о раньше обычном регламенте, который предусматривал меньше или намного меньше 200 часов), надо считать смехотворными с точки зрения эффективности (разве что предполагается, что обучающиеся получают лишь некоторое пассивное знакомство с какой-то ограниченной областью терминологии). Тем не менее, пропагандировать существующую систему – это у нас непреложный закон; говорить же о ее плачевных результатах просто не подобает (и неудивительно: громадные деньги, растраченные в течение десятилетий впустую и все время растрачиваемые таким же образом вещь не очень благопристойная).

Профессор Александр Васильевич Зубов, блестящий, опытнейший специалист в области компьютерной глоттодидактики, мог бы высказаться по этому поводу гораздо компетентнее, чем я. Больше распространяться на эту тему не стану.

В настоящей статье я хочу рассмотреть один специальный вопрос, относящийся к более широкой области средств усовершенствования глоттодидактического процесса; притом имеется в виду процесс преподавания одного языка, именно русского, полякам. Независимо от того, насколько неэффективна основная система обучения иностранным языкам, этот процесс в тех или иных формах, хотя бы индивидуальных, осуществляется и будет осуществляться. Поэтому забота о нем не должна нас покидать.

Несомненно, существенную роль в этом отношении может играть разумная дифференциация и иерархизация материала, с которым имеет дело обучающийся, а затем последовательное внедрение соответствующих данных на подходящих этапах живого процесса преподавания.

С одной стороны необходимо устранить из занятий все то, что сводится к пустой таксономии языковых явлений с ее многочисленными ярлыками, которые никак не способствуют подбору нужных вполне конкретных выражений для обслуживания текущих коммуникативных потребностей; соответствующие описания и аналитические упражнения приводят только к чистой потере времени. С другой стороны бесполезно тратить слишком много времени на подробное теоретическое усвоение того, чем любое существо, обладающее языковой способностью, овладевает практически без малейшего труда, того, что каждому нормальному говорящему дается совсем легко. Причем имеются в виду также весьма специфические явления, например такие, как тот факт, что общеупотребительным окончанием род. падежа ед. числа существительных мужского рода в русском языке является **-а** и что, стало быть, поляку, переходящему на русский язык нечего заботиться о возможном выборе из двух окончаний: **-а** и **-у** (я оставляю в стороне т. наз. паритив, заведомо факультативный); поляк мгновенно улавливает эту особенность и пускает ее в ход; может случиться и так, что он осознает это свое поведение только вторично, что не имеет уже никакого практического значения.

Зато существуют совсем не самоочевидные языковые данные, которые имеют особое значение с точки зрения базисных коммуникативных потребностей и частотности их обслуживания. Ясно, что усвоение их должно предшествовать усвоению более второстепенных элементов. Таким образом, возникает то, что можно назвать глоттодидактическим приоритетом. В данном конкретном соотношении его надо приписывать тем языковым явлениям, которые обладают, во-первых, признаком относительно большой производительности в роли компонентов или свойств текстов, считающихся ожидаемыми продуктами обучающихся, во-вторых, признаком относительной простоты и легкости усвоения (возрастающей громоздкости соответствующих формул должно сопутствовать передвижение их усвоения к более поздним стадиям процесса; первенство в таких случаях должно отводиться овладению элементами класса в индивидуальном порядке, в отличие от самого класса).

Совершенно очевидно, что в глобальной трактовке языкового материала приоритет должен присваиваться готовым, цельным лексическим единицам, включая наравне с ними и фразеологизмы, вместе всеми их синтаксическими свойствами, т.е. с их полной синтаксической «проекцией» (в частных случаях она может быть и нулевой) или с их ролью заполнения определенных валентностных мест, а также устойчивым оборотам, которые состоят из обычных единиц языка, но которые идиоматичны по своей коммуникативной приуроченности (примером здесь может служить русское установившееся «успокоительное заявление» ничего страшного; оно могло бы соответствовать польскому *nic strasznego*, но на самом деле такого точного переводного соответствия нет).

Намеченный глобальный приоритет вытекает из следующих предпосылок. Главенствующей функцией языка является познавательная функция, т.е. функция передачи знания; следовательно, мы заинтересованы главным образом во взаимопонимании. Притом любому существу, обладающему языковой способностью, свойственна необычайная, чуть ли не беспредельная, гибкость и быстрота в необходимом восполнении формально отсутствующих элементов информации, при сопоставлении с остальными данными в его знании: язык, который сам всецело опирается на альтернативы, позволяет *нормальному* говорящему в огромном большинстве случаев правильно элиминировать представляющие ему, но неподходящие члены альтернатив. Поэтому тем, что в обычном случае достаточно для понимания, является основной «костяк» единицы языка или сочетания таких единиц; разные «грамматические» особенности и подробности оформления могут и отсутствовать. Правда, особо сложные, многоярусные, основанные на переплетающихся одно-, дву- и многосторонних зависимостях членов языковые конструкции могут становиться, при нарушении точных форм показателей этих зависимостей, неудобопонятными. Однако этим качеством сложные конструкции отличаются и сами по себе, даже в случае их полной корректности. Итак, в общем и целом, частные грамматические признаки играют в процессе взаимопонимания собеседников явно второстепенную роль: их нарушение приводит к информационному провалу в минимальном количестве случаев (если отталкиваться от всего истинного *океана* общения, где господствуют *простые* высказывания); впрочем, даже лексические ошибки (с точки зрения намерения говорящего) очень часто преодолеваются в понимании без труда.

Однако не следует сводить языковое *общение* к успешному *пониманию*. *Полный*, эстетический, процесс общения охватывает *полную* правильность высказываний. Причем грамматическая правильность, на пути к этой «полной правильности», безусловно, имеет первенство перед фонетической правильностью, т.е. правильностью с точки зрения т. наз. норм произношения (с тем, что этот последний вид правильности необходимо отличать от фонологической [смыслоразличительной] правильности, которая является неотъемлемой принадлежностью самих единиц языка).

Главным проявлением полной внефонетической правильности высказывания является, наряду с т. наз. порядком слов, соблюдение предусматриваемых данным кодом, всегда специфических для каждого кода, особенностей *словоизменения*. Это проявление имеет особый удельный вес в некоторых

языках; к этим, «флексивным», языкам относятся и русский и польский. Что касается вопроса о *сущности* «словоизменения», мы придерживаемся концепции профессора Анджея Богуславского, которую он изложил в своей книге [1].

Как отмечают современные методисты, количество факторов, влияющих на процесс овладения языком, составляет несколько сотен [2]. Эти факторы – связаны и со спецификой изучаемого языка, и способами его описания и овладения, и методам обучения и т.п. И вполне естественно, создавая тот или иной метод обучения языку, приходится упрощать сложный процесс овладения языком, сводя его к сравнительно небольшому числу факторов.

Тоже самое можно сказать и о методах обучения русскому языку как иностранному. Создавая соответствующие методы необходимо учитывать изменившийся портрет языковой личности обучаемого, современные технологии обучения, современные взгляды на процесс обучения как на технологический процесс. В наши дни такой процесс представляется интегрированным, объединяющим различные подходы, методы, модели и формы обучения. Использование современных информационных технологий позволит обеспечить иное взаимодействие, иную степень интерактивности всех участников образовательного процесса [3].

Если говорить более конкретно, такие технологии дают возможность организовать открытую учебную языковую среду, приближающуюся по аутентичности и интенсивности языкового общения к реальной речевой коммуникации. Еще одно важное преимущество современных информационных технологий заключается в том, что они позволяют организовать индивидуализированный процесс обучения. Это особенно важно при обучении русскому языку как иностранному, ибо контингент обучаемых здесь весьма различен по начальному уровню владения языком и по степени мотивации обучения. Значительно различается он и по психофизиологическим особенностям обучаемых и сформированности у них учебных умений и навыков.

Навстречу вышесказанному выходит предложение, высказанное в книге «Информационные технологии в обучении русскому языку как иностранному» [4].

Предлагаемая книга имеет целью познакомить педагогов с существующими компьютерными средствами обучения русскому языку как иностранному. Вторая задача – показать, как от конкретной учебной программы «Русский язык как иностранный» перейти к конечному числу конкретных учебных задач, которые могут быть выполнены компьютером. И третья задача – научить преподавателя русского языка создавать обучающие сценарии для будущих компьютерных обучающих программ.

Как отметил известный специалист по автоматической обработке текстов Терри Виноград, компьютеры – «это языковые машины: основа их могущества заключается в способности манипулировать лингвистическими знаками-символами, которым приписывается некоторый смысл» [5].

В вышеупомянутой книге сделана попытка показать те немногочисленные современные компьютерные обучающие средства, которые используются при обучении русскому языку как иностранному. Особый акцент

сделан на детальное описание процесса перехода от конкретной учебной программы дисциплины «Русский язык как иностранный» к конечному числу учебных задач, которые могут выполнить компьютеры.

Как отмечают в последние десятилетия многие ученые, интерес к русскому языку как иностранному постоянно растет. И сфере обучения русскому языку реальный интерес могут представлять только «живые ресурсы», содержание которых постоянно совершенствуется и оперативно обновляется при использовании самых совершенных информационных технологий. Создавать соответствующие учебные информационные ресурсы могут только преподаватели–русисты, владеющие современной методикой создания компьютерных средств обучения, и в них главное и второстепенное поможет обучаемым, с одной стороны, правильно и быстро постигать содержание текстов, а с другой стороны – станет руководством для педагогов при создании принципиально новых учебников и учебных пособий. В итоге будут выработаны новые модели интеллектуального взаимодействия человека и компьютера.

Вышеуказанная монография – это эффект многолетнего научного сотрудничества Института восточнославянской филологии Университета в Белостоке и кафедры информатики и прикладной лингвистики Минского государственного лингвистического университета.

ЛИТЕРАТУРА

1. *Boguslawski, A. Two Essays on Inflection / A. Boguslawski. – Warszawa, 1992.*
2. *Лебединский, С. И. Теория обучения русскому языку как иностранному: психолого-лингвистические и лингвометодические аспекты : учеб. пособие / С. И. Лебединский. – Минск : БГУ, 2005. – С. 46–47.*
3. *Николенко, Е. Ю. Современная информационно-коммуникационная парадигма и прагматические проблемы лингводидактики (типы учебных пособий по русскому языку как иностранному) / Е. Ю. Николенко // Вестн. Моск. ун-та. Сер. 9. Филология. – 2008. – № 5. – С.91–92.*
4. *Хайчук, Роман. Информационные технологии в обучении русскому языку как иностранному / Роман Хайчук, Александр Зубов. – Белосток, 2009. – 180 с.*
5. *Виноград, Т. Работа с естественным языком / Т. Виноград // Современный компьютер. – М. : Мир, 1986. – С. 84–96.*

АВТОРЫ СТАТЕЙ СБОРНИКА

Аникеева Татьяна Владимировна – преподаватель кафедры фонетики английского языка МГЛУ

Белюга Марина Александровна – магистрант кафедры информатики и прикладной лингвистики МГЛУ

Беляева Лариса Николаевна – доктор филологических наук, профессор кафедры образовательных технологий в филологии Российского государственного педагогического университета им. А. И. Герцена (г. Санкт-Петербург)

Березюк Дарья Владимировна – студентка МГЛУ

Бобков (Bobkov) Анатолий Калистратович – кандидат филологических наук, ст. научный сотрудник НИЧ БГУ, сотрудник лаборатории «Информационные технологии и компьютерная графика», ведущий бизнес-аналитик компании “Byelex”

Богданова Наталья Альбертовна – кандидат филологических наук, доцент МГЛУ

Богушевич Дмитрий Георгиевич – доктор филологических наук, профессор кафедры истории и грамматики английского языка МГЛУ

Болдак Ирина Александровна – кандидат филологических наук, доцент кафедры иностранных языков №2 Гродненского государственного университета им. Я. Купалы

Болдак-Бриал (Boldak-Brial) Анастасия Александровна – аспирант Университета Лион-3 имени Жана Мулена (г. Лион, Франция)

Бусел Татьяна Викторовна – кандидат филологических наук, доцент кафедры иноязычного речевого общения МГЛУ

Василевская Виктория Анатольевна – преподаватель кафедры информатики и прикладной лингвистики МГЛУ

Василенко Елена Сергеевна – преподаватель кафедры информатики и прикладной лингвистики МГЛУ

Верби́ненко (Verbynenko) Юлия Игоревна – мл. научный сотрудник Украинского языково-информационного фонда НАН Украины

Виссия (Vissia) Херман – кандидат технических наук, владелец компании “Byelex Multimedia Product BV” (Минск)

Гяцэвіч Юрась Станіслававіч – кандыдат тэхнічных навук, загадчык лабараторыі распазнавання і сінтэзу маўлення АПП НАН Беларусі

Головня Анастасия Ивановна – кандидат филологических наук, доцент кафедры прикладной лингвистики Белорусского государственного университета

Горбунов Юрий Иванович – доктор филологических наук, доцент, профессор кафедры теории и практики перевода Тольяттинского государственного университета

Горбунова Ольга Юрьевна – ст. преподаватель кафедры теории и практики перевода Тольяттинского государственного университета

Гафуров (Gafurov) Сергей – руководитель группы, разработчик программного обеспечения компании “Byelex” (Минск)

Грибач Татьяна Анатольевна – преподаватель кафедры иностранных языков Белорусского государственного университета транспорта, научный сотрудник фирмы “IBA Group Head Office”

Грязнухина Татьяна Александровна – кандидат филологических наук, ст. научный сотрудник Украинского языково-информационного фонда НАН Украины

Гурбик Екатерина Александровна – доцент кафедры английского языка и литературы Университета Файсалабада (Пакистан)

Дарчук Наталья Петровна – доктор филологических наук, профессор кафедры современного украинского языка Института филологии Киевского национального университета им. Тараса Шевченко

Детскина Раиса Владимировна – кандидат филологических наук, доцент кафедры информатики и прикладной лингвистики МГЛУ

Дзибук Татьяна Владимировна – магистрант Минского государственного лингвистического университета

Долматова Екатерина Дмитриевна – кандидат филологических наук, доцент, доцент кафедры фонетики английского языка МГЛУ

Еремеева Юлия Павловна – магистрант кафедры компьютерной инженерии и программной техники Российского государственного педагогического университета им. А. И. Герцена (г. Санкт-Петербург)

Ефремова Алла Александровна – кандидат филологических наук, доцент кафедры белорусского и иностранных языков Академии МВД Республики Беларусь

Жданович Алла Евгеньевна – преподаватель кафедры информатики и прикладной лингвистики МГЛУ

Жубанов Аскар Кудайбергенович – доктор филологических наук, профессор, зав. отделом прикладной лингвистики Института языкознания им. А. Байтурсынова НАН Республики Казахстан

Заика Наталья Михайловна – кандидат филологических наук, ст. научный сотрудник Украинского языково-информационного фонда НАН Украины

Зубань Оксана Николаевна – кандидат филологических наук, доцент кафедры современного украинского языка Киевского национального университета им. Тараса Шевченко

Зубов Александр Васильевич – доктор филологических наук, профессор, зав. кафедрой информатики и прикладной лингвистики МГЛУ

Зубова Ирина Ивановна – доцент кафедры информатики и прикладной лингвистики МГЛУ

Зыгмантович Наталья Викентьевна – ст. преподаватель кафедры информатики и прикладной лингвистики МГЛУ

Иванова-Мицкевич Ирина Валерьевна – кандидат филологических наук, доцент кафедры истории и грамматики английского языка МГЛУ

Ильина Елена Иосифовна – преподаватель кафедры информатики и прикладной лингвистики МГЛУ

Камшилова Ольга Николаевна – кандидат филологических наук, доцент кафедры образовательных технологий в филологии Российского государственного педагогического университета им. А. И. Герцена (г. Санкт-Петербург)

Карневская Елена Борисовна – доктор филологических наук, профессор кафедры фонетики английского языка МГЛУ

Карпилович Татьяна Павловна – доктор филологических наук, профессор, зав. кафедрой теории и практики английского языка МГЛУ

Карпова Ольга Михайловна – доктор филологических наук, профессор, проректор по связям с общественностью и воспитательной работе Ивановского государственного университета

Качан Яўгенія Сяргеёўна – мал. навуковы супрацоўнік, стажор лабараторыі распазнавання і сінтэзу маўлення АПП НАН Беларусі

Коваленко Анастасия Дмитриевна – преподаватель-стажер кафедры фонетики английского языка МГЛУ

Козеренко Елена Борисовна – кандидат филологических наук, зав. лабораторией федерального исследовательского центра «Информатика и управление» Российской академии наук

Копачева Нина Андреевна – кандидат филологических наук, доцент кафедры стилистики английского языка МГЛУ

Кошчанка Уладзімір Аляксандравіч – навуковы супрацоўнік Інстытута мовазнаўства імя Якуба Коласа Цэнтра даследаванняў беларускай культуры, мовы і літаратуры НАН Беларусі

Краснопрошин (Krasnoproshin) Виктор Владимирович – доктор технических наук, профессор, зав. кафедрой информационных систем управления ФПМИ БГУ

Кривальцэвіч Алена Віктараўна – мал. навуковы супрацоўнік, стажор лабараторыі распазнавання і сінтэзу маўлення АПП НАН Беларусі

Кунцевич Светлана Евгеньевна – кандидат филологических наук, доцент, зав. кафедрой стилистики английского языка МГЛУ

Лосик Георгий Васильевич – доктор психологических наук, главный научный сотрудник объединенного Института проблем информатики НАН Беларуси

Любченко Татьяна Петровна – кандидат технических наук, научный сотрудник Украинского языково-информационного фонда НАН Украины

Макарич Марина Васильевна – кандидат филологических наук, доцент кафедры английского языка № 2 Белорусского национального технического университета

Макогон Ирина Станиславовна – магистрант кафедры образовательных технологий в филологии Российского государственного педагогического университета им. А.И. Герцена (г. Санкт-Петербург)

Маракулина Полина Александровна – студентка МГЛУ

Марковская Елена Вадимовна – преподаватель кафедры информатики и прикладной лингвистики МГЛУ

Меньшенина Ирина Александровна – аспирант Минского государственного лингвистического университета

Молоткова Анастасия Викторовна – студентка МГЛУ

Москалева Мария Игоревна – выпускница МГЛУ 2016 года

Нагарнович Эллина Вацлавовна – студентка МГЛУ

Новикова Ксения Владимировна – магистрант Российского государственного педагогического университета им. А. И. Герцена (г. Санкт-Петербург)

Нымм Вольдемар Рихардович – кандидат педагогических наук, доцент кафедры образовательных технологий в филологии Российского государственного педагогического университета им. А. И. Герцена (г. Санкт-Петербург)

Останина-Ольшевская Юлия Александровна – кандидат филологических наук, адъюнкт Института прикладной лингвистики Варшавского университета

Остапова Ирина Вадимовна – кандидат технических наук, ст. научный сотрудник Украинского языково-информационного фонда НАН Украины

Панова Инна Ивановна – кандидат филологических наук, профессор кафедры фонетики английского языка МГЛУ

Пиотровская Ксения Раймондовна – доктор педагогических наук, профессор кафедры информатики Российского государственного педагогического университета им. А. И. Герцена (г. Санкт-Петербург)

Потапова Родмонга Кондратьевна – доктор филологических наук, профессор, зав. кафедрой прикладной и экспериментальной лингвистики, директор Института прикладной и математической лингвистики Московского государственного лингвистического университета

Рычкова Людмила Васильевна – кандидат филологических наук, доцент, зав. кафедрой общего и славянского языкознания Гродненского государственного университета им. Я. Купалы

Семак Александр Сергеевич – аспирант кафедры информатики и прикладной лингвистики МГЛУ

Сиразитдинов Зиннур Амирович – зав. лабораторией лингвистики и информационных технологий Института истории, языка и литературы Уфимского научного центра РАН

Скрябина Анна Владимировна – магистрант МГЛУ

Смалюк Антон Федаравіч – ст. навуковы супрацоўнік Навукова-даследчай лабараторыі дынамікі сістэм і механікі матэрыялаў БНТУ

Сосунова Татьяна Андреевна – магистрант Российского государственного педагогического университета им. А. И. Герцена (г. Санкт-Петербург)

Степанова Дарья Валерьевна – кандидат филологических наук, доцент кафедры теории и практики английской речи МГЛУ

Сянкевіч Ніна Міхайлаўна – навуковы супрацоўнік Цэнтра даследаванняў беларускай культуры, мовы і літаратуры НАН Беларусі, філіял Інстытута мовазнаўства

Товпинец Ксения Сергеевна – соискатель кафедры информатики и прикладной лингвистики МГЛУ

Третьяк Анастасия Олеговна – студентка МГЛУ

Филимонова Таисия Аркадьевна – ст. преподаватель кафедры информатики и прикладной лингвистики МГЛУ

Хайчук Роман – доктор филологических наук, профессор ГВПШ г. Сувалки (Польша)

Чапля Аркадий Иосифович – кандидат филологических наук, доцент, свободный исследователь (Израиль)

Шалимо Наталья Владимировна – преподаватель кафедры информатики и прикладной лингвистики МГЛУ

Шимуля Роберт – кандидат филологических наук, зам. директора Института восточнославянских языков Университета в Белостоке (Польша)

Шировов Владимир Анатольевич – доктор технических наук, академик НАН Украины, директор Украинского языково-информационного фонда НАН Украины

Шумская Мария Александровна – магистрант МГЛУ

Щербин Вячеслав Константинович – кандидат филологических наук, зав. сектором инновационного развития отраслей Центра системного анализа и стратегических исследований НАН Беларуси

Юдкин-Рипун Игорь Николаевич – доктор искусствоведения, член-корреспондент Национальной академии искусств Украины, зав. отделом Института искусствоведения, фольклористики и этнологии НАН Украины

Ющенко Светлана Сергеевна – мл. научный сотрудник Украинского языково-информационного фонда НАН Украины

Яковишин Владимир Семенович – кандидат филологических наук, выпускник Минского государственного лингвистического университета, свободный исследователь

Научное издание

**АКТУАЛЬНЫЕ ПРОБЛЕМЫ
СОВРЕМЕННОЙ ПРИКЛАДНОЙ ЛИНГВИСТИКИ**

Сборник научных статей

В авторской редакции

Ответственный за выпуск *А. В. Зубов*

Подписано в печать 15.02.2017. Формат 60х84 ¹/₁₆. Бумага офсетная. Гарнитура Таймс. Ризография. Усл. печ. л. 19,07. Уч.-изд. л. 12,15. Тираж 110 экз. Заказ 13.

Издатель и полиграфическое исполнение: учреждение образования «Минский государственный лингвистический университет». Свидетельство о государственной регистрации издателя, изготовителя, распространителя печатных изданий от 02.06.2014 г. № 1/337. ЛП № 02330/458 от 23.01.2014 г. Адрес: 220034, г. Минск, ул. Захарова, 21.