ИНТЕЛЛЕКТ ЯЗЫК КОМПЬЮТЕР

Выпуск 18
Том I

ТРУДЫ

МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ПО КОМПЬЮТЕРНОЙ И КОГНИТИВНОЙ ЛИНГВИСТИКЕ

TEL–2018

Казань, 31 октября – 3 ноября 2018 г.

ИНТЕЛЛЕКТ. ЯЗЫК. КОМПЬЮТЕР

ВЫПУСК 18

ТРУДЫ МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ ПО КОМПЬЮТЕРНОЙ И КОГНИТИВНОЙ ЛИНГВИСТИКЕ

TEL-2018

Том 1

КАЗАНЬ
31 октября – 3 ноября
2018

УДК 004.8+81’32
ББК 81.1

Академия наук Республики Татарстан
Институт прикладной семиотики

Казанский (Приволжский) федеральный университет
Российский фонд фундаментальных исследований

Издание осуществлено при финансовой поддержке
Российского фонда фундаментальных исследований совместно
с Правительством Республики Татарстан
(проект №1 8-47-161001)

Научные редакторы:
доктор физико-математических наук **А. М. Елизаров**,
доктор технических наук **Н. В. Лукашевич**

Т78 **Труды международной конференции по компьютерной и когнитивной лингвистике TEL-2018.** – В 2-х томах. Т 1. – Казань: Изд-во Академии наук РТ, 2018. – 438 с.
ISBN 978-5-9690-0477-1

Сборник содержит материалы международной конференции по компьютерной и когнитивной лингвистике TEL-2018 (Казань, 31 октября – 3 ноября 2018 г.).

Для научных работников, преподавателей, аспирантов и студентов, специализирующихся в области компьютерной и когнитивной лингвистики и ее приложений.

УДК 004.8+81’32
ББК 81.1

ISBN 978-5-9690-0477-1

XV Международная конференция
по компьютерной и когнитивной лингвистике
TEL'2018
(31 октября – 3 ноября 2018 г., Казань, Россия)

Программа конференции включала пленарные и секционные заседания, круглые столы, демонстрации программных систем.

Направления работы конференции:

– лингвистическая семантика и семантический анализ текста;
– семантические технологии Web;
– разработка и применение компьютерных лингвистических ресурсов;
– тезаурусы, онтологии;
– машинный перевод;
– корпусы национальных языков;
– семиотическое моделирование;
– речевые технологии.

Программный комитет:

Сулейманов Джавдет Шевкетович (Казань, Татарстан, Россия) – председатель
Дыбо Анна Владимировна (Москва, Россия)
Голенков Владимир Васильевич (Минск, Беларусь)
Елизаров Александр Михайлович (Казань, Татарстан, Россия)
Желтов Валериан Павлович (Чебоксары, Чувашия, Россия)
Кобозева Ирина Михайловна (Москва, Россия)
Ландэ Дмитрий Владимирович (Киев, Украина)
Мосин Сергей Геннадьевич (Казань, Татарстан, Россия)
Поляков Владимир Николаевич (Москва, Россия)
Соловьев Валерий Дмитриевич (Казань, Татарстан, Россия)
Соснин Петр Иванович (Ульяновск, Россия)
Татевосов Сергей Георгиевич (Москва, Россия)
Хасьянов Айрат Фаридович (Казань, Татарстан, Россия)
Шарипбаев Алтынбек Амирович (Астана, Казахстан)
Широков Владимир Анатольевич (Киев, Украина)

Организационный комитет:

Невзорова Ольга Авенировна – председатель (АН РТ, Казань, Татарстан, Россия)

Гильмуллин Ринат Абрекович – зам. председателя (АН РТ, Казань, Татарстан, Россия)

Гатиатуллин Айрат Рафизович – ученый секретарь (АН РТ, Казань, Татарстан, Россия)

Галимянов Анис Фуатович (КФУ, Казань, Татарстан, Россия)

Галиева Альфия Макаримовна (АН РТ, Казань, Татарстан, Россия)

Аюпов Мадехур Масхутович (АН РТ, Казань, Татарстан, Россия)

Хакимов Булат Эрнстович (КФУ, Казань, Татарстан, Россия)

Хусаинов Айдар Фаилович (АН РТ, Казань, Татарстан, Россия)

Гатауллин Рамиль Раисович (АН РТ, Казань, Татарстан, Россия)

Курманбакиев Марат Ильдарович (АН РТ, Казань, Россия)

Компьютерная лингвистика – современное научное направление, ориентированное на разработку компьютерных моделей, методов и технологий в лингвистике и смежных областях. В настоящем сборнике Трудов конференции (том 1) представлены статьи по актуальным проблемам когнитивной и компьютерной лингвистики, включая формальные модели синтаксиса и семантики, системы поиска и классификации, онтологии, корпусы национальных языков и лингвистические базы данных, программные системы обработки ЕЯ и др.

Тематика конференции находится в постоянном развитии. В 2014 году возникло новое направление, связанное с задачами тюркской компьютерной и корпусной лингвистики, в частности с актуальной задачей унификации систем разметки на уровнях морфологии, синтаксиса и семантики для электронных корпусов тюркских языков. Это направление было активно представлено и в программе конференции TEL-2018. В работе конференции приняли участие ведущие разработчики основных тюркских корпусов из Казахстана, Якутии, Кыргызстана, Башкортостана, России.

Впервые в рамках конференции был успешно организован международный семинар Computational Models in Language and Speech (CMLS 2018). Статьи, представленные на этот семинар и отобранные по результатам рецензирования Программным комитетом семинара, опубликованы в материалах CEUR Workshop Proceedings (CEUR-WS.org, ISSN 1613-0073), vol.2303, а также в сборнике Трудов конференции (том 2).

Научные редакторы
Сулейманов Д.Ш., Невзорова О.А.

УДК 004.652.5

ONTOLOGICAL MODEL OF UZBEK LANGUAGE (as example morphology)

Nilufar Abdurakhmonova, PHD
Tashkent state university of Uzbek language and literature, Tashkent
abdurahmonova.1987@mail.ru

This article analyzes the ontological models of the Uzbek language in one of the five languages within the project. Ontological models have their own advantages over the processing of texts on the computer, particularly for rule-based machine translation, corpus linguistics, and morphological analysis. The issue of grammatical rules was analyzed in the Uzbek language, which is one of the most effective ways to classify grammatical categories and analyze their case. The taxonomic classification of grammatical categories in the creation of metalanguage in Turkic languages is reflected in the article.

Key words: ontology, morphological analyzer, Protégé, grammatical categories, rule based machine translation.

ОНТОЛОГИЧЕСКАЯ МОДЕЛЬ УЗБЕКСКОГО ЯЗЫКА (как пример морфологии)

Нилуфар Абдурахмонова, PHD
Ташкентский государственный университет узбекского языка и литературы, Ташкент
abdurahmonova.1987@mail.ru

В этой статье анализируются онтологические модели узбекского языка на одном из пяти языков проекта. У онтологических моделей есть свои преимущества, особенно для машинного перевода на основе правил, лингвистики корпусов и морфологического анализа. Вопрос о грамматических правилах был проанализирован на узбекском языке, который является одним из наиболее эффективных способов классификации грамматических категорий и анализа их дела. В статье отражена таксономическая классификация грамматических категорий на турецком языке.

Ключевые слова: онтология, морфологический анализатор, Protégé, грамматические категории, машинный перевод на основе правил.

The article prepared under the project of "Development of Turkic languages electronic thesauri for creation of multilingual search and knowledge extraction systems № AP05132249" (2018–2020, Eurasian national university).

I. Introduction

Now the flow of information is significantly increasing, it has become necessary to search for new ways of its storage, presentation, formalization and systematization, as well as automatic processing. Thus, there is a growing interest in comprehensive knowledge bases that can be used for various practical purposes. Of great interest are systems that can, without human intervention, extract any information from the text. As a result, against the background of newly emerging needs, new technologies are being developed to solve the stated problems. Along with the World Wide Web, its extension appears, the Semantic Web, in which hypertext pages are provided with additional markup that carries information about the semantics of the elements included in the pages. An integral component of the Semantic Web is the concept of ontology, which describes the meaning of semantic markup. In general terms, ontology is understood as a system of concepts of a certain subject domain, which is represented as a set of entities connected by various relationships. Ontologies are used to formally define concepts and relationships that characterize a particular area of knowledge.

The advantage of ontologies as a way of representing knowledge is their formal structure, which simplifies their computer processing. We can speak of the implicit use of ontologies as systems of concepts in the natural sciences (biology, medicine, geology, and others), where they serve as a kind of foundation for the construction of theories. Since the classification structure (taxonomy) is an integral part of any ontology, one can speak of the presence of ontology elements in special classifications and indexing systems (for example, in library classification codes). Explicitly, ontologies are used as data sources for many computer applications (for information retrieval, text analysis, knowledge extraction and other information technologies), allowing for more efficient processing of complex and diverse information.

An ontology is used for formal and specialized concepts and relationships related to the exact domain. Having the advantage of ontology in the NPL for creating metalanguages in the field of machine translation (mainly machine-based translation of rules) or for other purposes (information retrieval system, text analysis, text annotation). Thanks

to ontology, creating the structure of information based on system and hierarchical data, it helps to simplify the computational processing of natural language. An effective way to create an ontology is OWL. “There are several types of ontologies. The word “ontology” can mean different objects of computer science, depending on the context. For example, an ontology could be:

– a thesaurus in the field of information retrieval or
– a model presented in OWL in the field of related data or
– XML schema in the context of databases
– and etc.”.

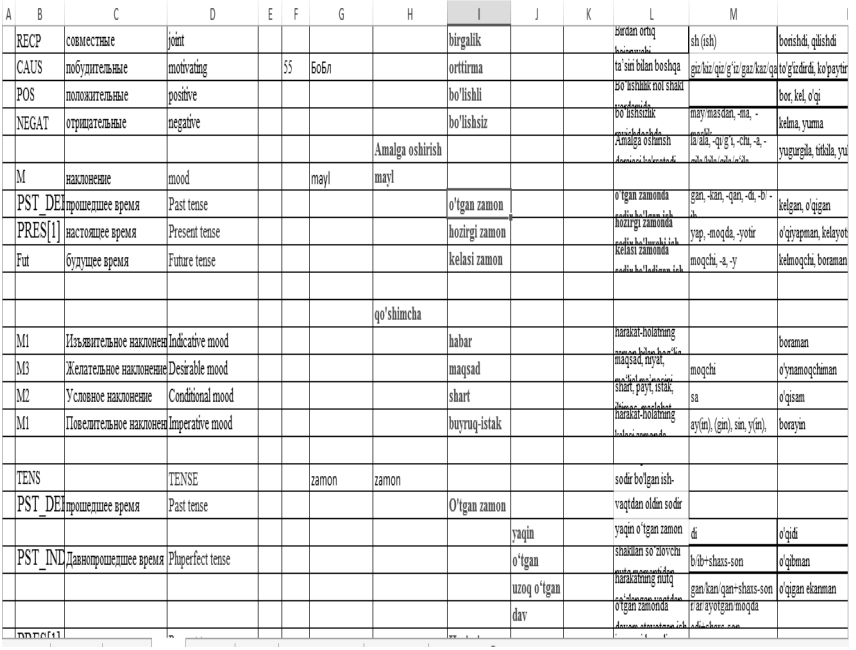
II. Ontology is one of stage morphoanalysis

Each language has systemicity, that grammar follow initial rules in order to make sense for understanding. Nevertheless, not every time the grammatical serve to compose text for understanding without semantics. This means that all domains of language are core for all speech activity. If we focus on computational morphology or syntax, either the smallest items are considered the basic function of path. Computational approaches to morphology and syntax are generally concerned with formal devices, such as grammars, and stochastic models, and algorithms, such as tagging or parsing [1]. Morphological analyzer is one splits up the word forms so that implement previous stage of process.

This issue has been studies as a focus of NLP for Turkic languages since 1960s. According to variances of natural languages, structures and forms there are typical different algorithms to deal with some problems in CL. In case any agglutinative language has no any full computational description, then it is difficult to get enough information from database so that the procedure of parsing as well.

Morphology is part of grammar, but it may be subpart other domain of language. According to some authors there are three types of morphological analysis: based on procedural method (systematization of morphological knowledge about a natural language and development of morphological information assignment algorithms to a separate word form), formal morphological rules, using the ontological models and the hypergraphs [2].

What is interesting to note is that ontological as ready devices like other sketch engines over and beyond investigation based in order to build data structure in different spheres. Ontology is a powerful and widely used tool for modeling relationships between objects which belong to the different subject area. As given definition for ontology “The Artificial-

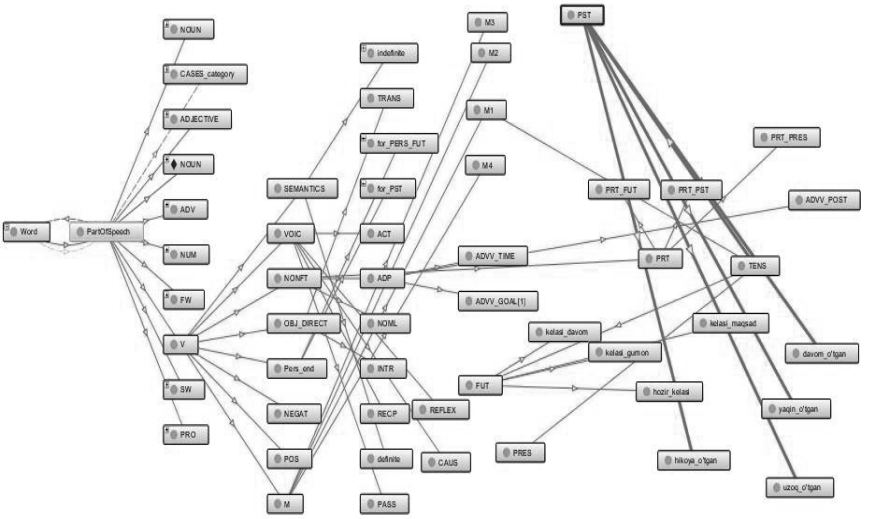


	RECP	совместные	joint					birgalik			birdan orhq	sh (ish)	borishdi, qilishdi
	CAUS	побудительные	motivating		55	Бобл		orttirma			ta`sir bilan boshqa	giz/kiz/qiz/g`iz/gaz/kaz/qa	to`g`irdirdi, ko`paytir
	POS	положительные	positive					bo'lishli			Bo`lishlik not shakl		bor, kel, o`qi
	NEGAT	отрицательные	negative					bo'lishsiz			Bo`lishsizlik	may/masdan, -ma, -	kelma, yurma
							Amalga oshirish				Amalga oshirish	la/ala, -qi/g`i, -chi, -a, -	yugurgila, tirkila, yu
	M	наклонение	mood			mayl	mayl						
	PST_DE	прошедшее время	Past tense					o'tgan zamon			o`tgan zamonda	gan, -kan, -qan, -di, -b/ -	kelgan, o'qigan
	PRES[1]	настоящее время	Present tense					hozirgi zamon			hozirgi zamonda	yap, -moqda, -yotir	o'qiyapman, kelayot
	Fut	будущее время	Future tense					kelasi zamon			kelasi zamonda	moqchi, -a, -y	kelmoqchi, boraman
							qo'shimcha						
	M1	Изъявительное наклонен	Indicative mood					habar			harakat-holatning		boraman
	M3	Желательное наклонение	Desirable mood					maqsad			maqsad, niyat,	moqchi	o'ynamoqchiman
	M2	Условное наклонение	Conditional mood					shart			shart, payt, istak,	sa	o'qisam
	M1	Повелительное наклонен	Imperative mood					buyruq-istak			harakat-holatning	ay(in), (gin), sin, y(in),	borayin
	TENS		TENSE			zamon	zamon				sodir bo'lgan ish-		
	PST_DE	прошедшее время	Past tense					O'tgan zamon			vaqtdan oldin sodir		
									yaqin		yaqin o`tgan zamon	di	o'qidi
	PST_IND	Давнопрошедшее время	Pluperfect tense						o'tgan		shaklan so`ziovchi	b/ib+shaxs-son	o'qibman
									uzoq o'tgan		harakatning muhq	gan/kan/qan+shaxs-son	o'qigan ekanman
									dav		o'tgan zamonda	r/ar/ayotgan/moqda	

Intelligence literature contains many definitions of an ontology; many of these contradict one another. For the purposes of this guide an ontology is a formal explicit description of concepts in a domain of discourse (classes (sometimes called concepts)), properties of each concept describing various features and attributes of the concept (slots (sometimes called roles or properties)), and restrictions on slots (facets (sometimes called role restrictions)). An ontology together with a set of individual instances of classes constitutes a knowledge base. In reality, there is a fine line where the ontology ends and the knowledge base begins. Classes are the focus of most ontologies. Classes describe concepts in the domain" [3].

In our point of view, for agglutinative languages morphology is strong side for all natural language processing by computer. We used the editor Protégé (http://protege.stanford.edu) to input grammatical classes and relations of Turkic languages (Uzbek, Kazakh, Tatar, Kyrgyz, and Turkish) into the ontological framework. It is a free open source ontology editor and a framework in order to generate knowledge bases so far.

The ontological model of the Uzbek parts of speech allows (1 fig.) us to work easier with programming languages like Java. It includes the morphological rules and the relationships of categories as hierarchic



system. Ontology gives to opportunity to map all chain with relations, properties, subclasses also.

For this reason, the principal purpose of our project to create meta-language for Turkic languages.

III. Development of a meta-language version of the concepts of morphological rules of the Uzbek language

The semantics of ontological models in Excel format and their structure were analyzed in the Uzbek language in independent and useful word genres. The morphological categories of the Uzbek language include subcategories, explanations, grammatical questions and related examples. When inserting affixes in Uzbek, there is almost no singularity. This is mainly due to the horse category. In the case of verbal words there is a difference in the amount of various bases. “Metalanguage is one of key concepts of system of the description of an object of science and is defined as artificial language of «the second order» in relation to which natural human language acts as «language object», that is as a subject of a linguistic research. In our case a natural language are the Kazakh and Turkish languages entering into the Turkic group of languages. The unified symbol system (UNIFIED TAG) was developed based on the idea of creating unified metalanguage for Turkic Languages. Firstly, the idea

of creating metalanguage was proposed at the 1st so-called International Conference on Computer processing of the Turkic Languages (TurkLang-2013) which was held in Astana on 3–4 October, 2013. A group of famous professors of technical sciences A.A. Sharipbay (Astana, Kazakhstan), D.SH. Suleimenov (Kazan, Tatarstan, Russia), Eşref Adalı (Istanbul, Turkey) is working on the creation of metalanguage" [4].

IV. Comparative analysis of ontologies

Tagging considered the core to understand each element of Noun. Therefore, we have such general results of languages morphological categories of Noun. In fact, there is not syngormonism in Uzbek, exception with other allomorphs; we can see more this position in those cases in Kazakh language. However there is unique similarities of grammatical point of view {SIMP, CMPL, FUSW, PAIR, CMPN, ABBR, UNDR, DRVT, COMP, ANIM, INAM, CMMN, PRPR, CNCR, ABST, Num., SG, COL} is common both of languages. Variables of case of Kazakh differ from Uzbek:

CASES TAG	Uzbek	Kazakh
NOM	Ø	Ø
GEN	-ning	-ның, -нің, -ың, -дің, -тың, -тің
DIR	-ga, -ka, -qa	-ға, -ге, -қа, -ке, -на, -не, -а, -е
ACC	-ni	Ны
LOC	-da	-да, -де, -та, -те, -нда, -нде
ABL	-dan	-нан, -нен, -дан, -ден, -тан, -тен
INST	Ø	-мен, -бен, -пен, -менен, -бенен, -пенен

Language	Plural	Tag
Tatar	-ЛІАр лар ләр нар нәр	PL
Kazakh	-ЛІАр лар лер дар дер тар тер	PL
Turkish	-lAr lar ler	PL
Kyrgyz	-Лар лар лер лор лөр дар дер дор дөр тар тер тор төр	PL
Uzbek	lar	PL

When we compare the ontological logic of the Uzbek, Kazakh, Tatar, Kyrgyz and Turkish languages in the Turkic languages, we identified similarities and differences between the languages:

difference	similarity
Morphological variability (allomorphs)	General morphology (derivation)
Internal categories of grammatical categories	Types of word categories (independent and auxiliary word categories)
Grammatical approach to analysis	Types of grammar analysis
Different names in Turkish	

V. Conclusion

In spite of diversity of languages, there is commonness of grammatical rules among the Turkic languages. Entities inputted in Protégé as classes including object properties, data properties, individuals, annotation etc. We hope that Ontology grammatical rules of Turkic languages (Uzbek, Kazakh, Tatar, Turkish, Kyrgyz) will be implemented correspondingly and it will service for computational language processing in perspectives. We think that ontological approached analysis is useful partanomy and taxonomy for creating semantic relations of the words also. Because it clearly shows distinction and connection of the words.

REFERENCES

1. Brian Roark, Richard Sproat. Computational approaches to morphology and Syntax. Oxford university press. 2007. – P. 1.

2. А.А. Шарипбаев, Г.Т. Бекманова, Г. Алтынбек, Е. Адалы, Л. Жеткенбай, У. Каманур Единый морфологический анализатор для казахского и турецкого языков // Turklang. 2017. – P. 234.

3. https://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

4. А.А. Шарипбаев, Г.Т. Бекманова, Г. Алтынбек, Е. Адалы, Л. Жеткенбай, У. Каманур Единый морфологический анализатор для казахского и турецкого языков // Turklang. 2017. – P. 240.

УДК 81’33, 81’374

MODERN LEXICOGRAPHY: DIRECTIONS FOR DEVELOPMENT

Liliya Bezborodova
Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic, Kazan
lilitt888@gmail.com

The Central problem of modern lexicography is the problem of lexicographical description of new words. In turn, the existence of many systems of classification of dictionaries signals the existence of the problem of systematization of dictionary reference literature. The solution of the designated problems through the definition of parameters (coordinates, reference points, differential features), which is strictly logically possible to build a classification scheme, will advance us on the way to create a modern actual lexicographic product in the form of a fundamentally new universal metalexicographic system.

Keywords: lexicography, language, dictionary, lexicographic parameter, classification system.

СОВРЕМЕННАЯ ЛЕКСИКОГРАФИЯ: НАПРАВЛЕНИЯ РАЗВИТИЯ

Л. В. Безбородова
Институт прикладной семиотики Академии наук Республики Татарстан, Казань
lilitt888@gmail.com

Центральной проблемой современной лексикографии остается проблема лексикографического описания новых слов. В свою очередь, существование множества систем классификации словарей сигнализирует о существовании проблемы систематизации словарно-справочной литературы. Решение обозначенных задач через определение параметров (координат, точек отсчета, дифференциальных признаков), по которым строго логически можно построить классификационную схему, продвинет нас на пути создания современного актуального лексикографического продукта в виде принципиально новой универсальной металексикографической системы.

Ключевые слова: лексикография, язык, словарь, лексикографический параметр, классификационная система.

Наш язык – это своего рода «лингвистический портрет» современного общества со всеми его плюсами и минусами, в нем отражена идеология, система ценностей и предпочтений современного человека, его общий уровень образованности и культуры. Вполне закономерно, что столь сложная и неоднозначная проблема описания и оценки тенденций развития современного русского языка представляется крайне актуальной.

Язык как продукт постоянного коллективного возвратного наблюдения за миром его носителей был, есть и будет самым надежным средством диагностики социальных, политических, культурных процессов, происходящих в современном российском обществе.

Новые явления в лексике, словообразовании, грамматике и прагматике русского языка последних лет важно осмыслять через призму изменений в русской языковой картине мира. Анализ современных словарей и данные различных исследований демонстрируют, что, несмотря на обилие лексических заимствований и внедрение в российскую речемыслительную сферу инокультурных моделей речевого поведения, русский язык в целом сохраняет традиционные способы восприятия мира сквозь призму слова. Это связано с тем, что исконно русские традиционные модели языковой концептуализации мира закодированы в языке и слабо поддаются серьезной трансформации несмотря на современные условия функционирования языка.

Русский язык начала 21 века характеризуется освоением новых заимствований также, как русский язык конца 20 века. В целом для настоящего периода развития русского языка характерно продолжение «периода языковой нестабильности», связанного с активным внедрением в отечественный дискурс массива иноязычной лексики. Однако общее направление адаптационных процессов «чужих слов» заключается в стремлении заимствованных лексических единиц органично влиться в систему заимствующего языка.

Новейший этап развития русского языка характеризуется также медийным словотворчеством, эти выводы можно сделать, проанализировав репрезентативный словообразовательный материал в дискурсе масс-медиа. Масс-медиа наших дней стали относительно автономной подсистемой общенационального языка, в которой формируются языковые изменения. Именно эта сфера развивается сегодня особенно активно и динамично, демонстрируя наиболее характерные модели новообразований. Богатый и разнообразный языковой материал, предоставляемый этой сферой, исследователи берут

за основу для выделения таких тенденций современного медийного словотворчества, как интернационализация и демократизация. Это богатый материал данных для анализа словообразовательной структуры новообразований и ее нарушений как средства экспрессии в вопросах формирования текста.

Ну и конечно же, центральной проблемой современной лексикографии остается проблема лексикографического описания новых слов. Очевидно, что любой словарь национального языка – это своеобразное зеркало материальной и духовной культуры народа. В первую очередь мы имеем в виду словари новой лексики, поскольку описание десятков тысяч неологизмов позволяет выделить главные тенденции в развитии языка того или иного периода.

В настоящее время существуют десятки систем классификации словарей, что очередной раз доказывает существование проблемы систематизации словарно-справочной литературы.

Как и в любой классификации, важно выработать определенные ее параметры (координаты, точки отсчета, дифференциальные признаки), по которым строго логически можно построить классификационную схему.

Так, Л.В. Щерба выделял шесть противопоставлений словарей. В его классификации учитывается не только языковой аспект, в ней заложен теоретически тот тип идеального словаря, в котором был бы описан весь словарный состав национального языка во всех его аспектах.

Классификации остальных ученых лишь дополняют и уточняют противопоставления словарей Л.В. Щербы, однако в этом уточнении шла выработка основных параметров типологии, охватывающих все аспекты описания слов в словарях.

В последнее время сложилось несколько направлений, или параметров классификации словарей, которые мы и возьмем за основу: **лингвистический, семиотический, формальный и социально-прагматический**.

Данная классификация (О. Л. Рублева, Дальневосточный университет) касается только словарей лингвистических (филологических), которые только и можно называть словарями в подлинном смысле этого слова.

В каждом из упомянутых 4-х параметров, в свою очередь, можно выделить по четыре (можно и больше) частных **ДП (дифференциальных признака):**

1. Семиотический параметр.
1.1. Единицы описания
1.2. Принцип описания.
1.3. Способ описания.
1.4. Язык описания.
2. Формальный параметр.
2.1. Объем словаря.
2.2. Состав словаря.
2.3. Форма словарной статьи.
2.4. Порядок расположения.
3. Прагматический параметр.
3.1. Национальность читателя.
3.2. Возраст читателя.
3.3. Категория читателя.
3.4. Назначение словаря.
4. Лингвистический параметр.
4.1. Отношение к норме.
4.2. Отношение к объекту.
4.3. Отношение ко времени.
4.4. Отношения к аспекту языка.

В ходе анализа справочных изданий мы обнаруживаем тенденцию создания классификационных систем металексикографического характера, которые могут включать в себя комплексные лексикографические системы лингвистических, психологических, семиотических, социологических координат.

Перспективность детальной разработки подобных систем координат и последовательный анализ содержания разработанных словарей и справочных изданий все еще очевидна, как и много лет назад, на заре компьютерной лексикографии. В качестве современного актуального лексикографического продукта мы видим создание принципиально новой **универсальной металексикографической системы**, которая могла бы решить существующие проблемы систематизации словарного знания.

Словарная фиксация – сложная категория, отражающая принципиальные взгляды лексикографа на выстраивание концепции словаря, принципы отбора слов, способы представления лексикографической информации, архитектуру словаря и т. д. Деятельность лексикографа, по мнению Валерия Вениаминовича Морковкина, – «это инженерно-филологическая деятельность, состоящая в изобретении словарей, словарных систем и словарных серий, установлении их

оптимального вида и компонентного состава, а также в разработке процедур, позволяющих с помощью композиционных преобразований сообщить словарю (словарям) максимальную информативность и удобство использования» [Морковкин 90:47].

В работе лексикографа основным термином мы считаем термин **лексикографический параметр** как «некоторый квант информации о языковой структуре, который в экстремальном случае может представлять для пользователя самостоятельный интерес, но, как правило, выступает в сочетании с другими квантами (параметрами) и находит специфическое выражение в словарях; иными словами – это особое словарное представление структурных черт языка». Юрий Николаевич Караулов, давший такое определение параметра еще в 1981 году в книге «Лингвистическое конструирование и тезаурус литературного языка», выделил на основе анализа словарей 67 параметров, различное количественное сочетание которых составляет и определяет тип словаря. Для лексикографа параметрическая характеристика определяет компетентный анализ характера и объема информации, представленной в словаре.

Такое лингвистическое конструирование расширяет возможности создания словарей современного типа, а создание лексикографических произведений с уникальным набором параметров становится возможным.

УДК 801

MACHINE TRANSLATION IN A TRANSLATOR WORKFLOW: PRACTICAL VIEW

L. N. Beliaeva
Herzen State Pedagogical University of Russia, S.Petersburg
lauranbel@gmail.com

The paper presents the peculiarities of using a modern MT system in the technological chain of scientific and technical translation. MT systems are considered as a part of translator's workstation, specifics of both text pre- and postediting and automatic dictionaries management are analyzed. Modern paradigm 4.0 (Industry 4.0 and Information 4.0) dictates a new trend in automation and data exchange in manufacturing, such that will greatly influence both technology and science. That obviously leads to a serious and qualitative change in translator training, since a translator's work today is, to a large extent, in the technical communication domain.

Keywords: machine translation, automatic dictionary, preediting, postediting, translator's workstation, translator's competences, technical communication.

МАШИННЫЙ ПЕРЕВОД В РАБОТЕ ПЕРЕВОДЧИКА: ПРАКТИЧЕСКИЙ АСПЕКТ

Л. Н. Беляева
РГПУ им. А. И. Герцена, Санкт-Петербург
lauranbel@gmail.com

В статье описываются особенности современного процесса работы с практической системой машинного перевода в рамках технологической цепочки обработки и перевода научных и технических текстов. Рассматриваются системы машинного перевода в составе автоматизированного рабочего места переводчика, особенности пред- и постредактирования текстов, анализируются особенности ведения пользовательских словарей. Учитывается современная парадигма 4.0 (Промышленность 4.0 и Информация 4.0), диктующая условия, при которых состояние технологии и науки определяются потенциалом процессов автоматизации в промышленности и соответствующими способами представления информации к разрабатываемым проектам. В связи с этим качественно изменяются требования к подготовке переводчиков, чья деятельность сегодня все больше переносится в область технической коммуникации.

Ключевые слова: машинный перевод, автоматический словарь, предредактирование, постредактирование, автоматическое рабочее место переводчика, компетенции переводчика, техническая коммуникация.

Многолетнее использование систем машинного перевода (МП) не только специалистами в различных областях знаний, но и переводчиками научных и технических (специальных) текстов определяет необходимость предварительного подведения итогов, касающихся собственно процедуры работы с такими системами, минимизации объема постредактирования результатов МП, а также связи методов постредактирования с полнотой и точностью автоматического словаря (АС) соответствующей системы.

Основными профессиональными пользователями систем МП являются специалисты, получившие в современной англоязычной литературе термин *language worker,* который приблизительно можно перевести как *специалист в области переработки текстов*; под этим термином объединяются все лингвисты, работающие в области переработки научных и технических текстов: терминологи, переводчики, технические писатели, специалисты в области передачи технической информации [1; 2]. Современный переводчик специальной литературы является участником технологического процесса обработки текста, поэтому использование специализированной предметно-ориентированной системы МП, выбранной лично им или компанией, в которой он работает, сегодня обязательно. Огромный выбор систем машинного перевода, автоматизированных словарей, баз данных и знаний определяет необходимость формирования автоматизированного рабочего места (АРМ) переводчика [3].

Известно, что автоматизированное рабочее место переводчика, как правило, включает специализированную систему машинного перевода с настроенными пользовательскими словарями, средства переводческой памяти, онлайновые словари, доступные средства автоматизации работы с терминологией [4: 5], комплекс резидентных словарей, тезаурусов, систем проверки орфографии, систем доступа к информации по различным сетям передачи данных, средства формирования моделируемых текстов определенной структуры [5; 6]. В переводческих системах, создаваемых в больших производственных компаниях, используется описание специально создаваемого контролируемого языка и средства работы с ним. Наиболее распространенные и активно реализуемые АРМ предназначены для непосредственного использования профессиональными переводчиками, знающими как исходный язык, так и язык перевода, лексическое описание которых включено в словарное обеспечение; такие АРМ позволяют переводчику сохранить полный контроль над продуцированием собственных переводов. Системы машинного перевода,

составляющие неотъемлемую часть АРМ, обеспечивают получение рабочего варианта перевода, жестко ориентированного на конкретную предметную область, задачи пользователя и тип документации. Любая система машинного перевода, выбранная и настроенная на необходимую предметную область, дает вариант перевода, который требует анализа и постредактирования.

В процессе любого перевода выделяются 3 основных этапа: ознакомление с текстом, создание перевода, его редактирование [7:109] или постредактирование в случае работы с системой МП. Само постредактирование, входящее в работу с результатами МП как обязательный этап, теоретически не должно вызывать затруднений, однако при его проведении необходимы определенные знания особенностей работы системы МП, а также структуры и состава АС.

Поскольку умение перевести специальный текст вырабатывается тогда, когда человек способен создать этот текст на родном языке, то профессиональные переводчики, терминологи, технические писатели должны обладать базовыми компетенциями в области создания специальных текстов на родных и иностранных языках, а также в области их перевода и обработки. В качестве такой обработки может рассматриваться извлечение информации, а также создание вторичных текстов любого типа и назначения.

Выполнение всех этих видов работ требует от специалистов в области обработки текстов:

1) знания типологии специальных и технических текстов на родном (русском) языке и иностранных языках, их различий и особенностей;

2) умения создавать все типы специальных текстов на родном языке;

3) умения создавать все типы специальных текстов на иностранном языке;

4) умения переводить тексты с учетом различий в требованиях к специальным текстам в различных культурах.

В то же время современное развитие науки и техники во многом определяется не только скоростью и качеством переработки постоянно расширяющегося потока научной и технической информации на разных языках, в большой степени поддерживаемое качественным переводом, но и степенью внедрения информационных технологий при реализации новых научных проектов и/или при разработке и внедрении конкретной научной и/или технической продукции. Успешность реализации этих процессов также во многом зависит

от квалификации переводчика, его умения не только быстро и качественно переводить предлагаемый материал, но и активно участвовать в его разработке и структурировании. Последнее требование непосредственно связано с особым подходом к процессам создания документации и обмена информацией, т. е. технической коммуникацией, что определяется новым подходом к автоматизации и обмену информацией в промышленном производстве − *Промышленностью 4.0* (*Industry 4.0*) [9].

В основе представления и перевода научной и технической документации в рамках этого нового подхода лежит понятие *авторской разработки структурированного контента* (*structured content authoring)*, которая предполагает предварительное разделение текста на небольшие части, называемые *тематическими разделами* (*topics*); для создания окончательного варианта конкретного документа эти разделы далее объединяются на основе *карт* (*maps*).

Подход к формированию текста в этих новых условиях опирается на особые требования к представлению информации, которая должна быть: молекулярной, т. е. формируемой из информационных молекул, а не из готовых документов, динамической, т. е. непрерывно обновляемой, предлагаемой, а не поставляемой в готовом виде, универсальной, т. е. интерактивной, доступной и удобной для поиска, спонтанной, т. е. вызываемой конкретными контекстами, профилируемой автоматически [10].

При этом сами молекулы рассматриваются как завершенные крупицы информации, а тематические разделы должны соответствовать темам текста. Тогда подобные молекулы могут алгоритмически маркироваться и использоваться для создания текстов разных типов, этот процесс в свою очередь также может быть автоматизирован. Известно, что различные инструментальные средства разрабатывались и применялись для того, чтобы оптимизировать продуцирование и поддержание больших массивов текстовых документов на основе систем, которые позволяют создавать тексты параллельно, избегая дублирования контента за счет повторяющихся тематических разделов. Тем самым облегчается модификация текстов, связанная с разработкой новых версий изделия, уменьшаются расходы на услуги переводчиков и т. д.

В основе нового подхода к формированию документации лежит анализ продуктивности (*productivist approach*), при котором степень детализации тематических разделов определяется задачами создания научной и технической документации и потенциально отделена

от самого содержания, т. е. от тех тем, которые реально обсуждаются в тексте [11].

Специалисты в области разработки технической документации остро необходимы сегодня, они должны обладать рядом стандартных компетенций в области:

1. планирования своей работы, учитывая

• особенности адресатов текста и их профессиональный уровень,

• конкретное предназначение текста и собственное владение материалом,

• бюджет времени, отведенный на создание текста, включая оценку времени на написание текста, его пересмотр и редактирование;

2. создания специального текста, учитывая такие требования как ясность, краткость, простота выбираемых выражений, использование корректной терминологии, активного залога, полных синтаксических конструкций, отказ от использования синонимических терминов;

3. анализа и редактирования получаемого результата [12].

Однако Информация 4.0 требует новых компетенций, к которым в самом общем виде относятся следующие:

• способность собирать, анализировать и отбирать подходящую информацию, чтобы разрабатывать информационный продукт,

• способность выбирать такую стратегию разработки продукта, чтобы получать соответствующие информационные продукты для различных целей и потребителей,

• способность гарантировать, что информация является извлекаемой и доступной, представляет связную ментальную модель и согласуется по продуктам и средам,

• умение выбирать аппаратные средства и программное обеспечение, необходимое для использования в научной и технической коммуникации,

• способность разрабатывать и оценивать модули электронного обучения,

• знание процесса издания информационного продукта и его стадий,

• достаточное понимание предметных областей, которые являются релевантными для специалистов по распространению технической информации (информатика, машиностроение, физика и т. д.), чтобы быть способными сотрудничать с экспертами в предметной области,

• знание основных принципов и методов терминоведения,

• способность формировать ресурсные и лексикографические базы данных и корпуса текстов для решения профессиональных задач [ср. 13].

Работа большинства систем МП осуществляется на нескольких иерархически соподчиненных уровнях автоматического предредактирования текста; лексико-морфологического анализа; контекстного анализа и анализа групп; анализа функциональных сегментов; анализа предложений; синтеза выходного текста; автоматического постредактирования. Для полноценного использования системы МП переводчик должен представлять себе в общем виде общую процедуру анализа текста в системе, что позволит заранее готовить текст так, чтобы результат МП требовал минимального редактирования. Анализ текста системой МП начинается с уровня формального анализа, результатом решения этой задачи является предварительная разметка текста: установление границ отдельных разделов, заголовков, оглавления, таблиц, рисунков, формул. Кроме того, особым образом обрабатывается и запоминается формально-графическая структура текста, что необходимо для ее восстановления при синтезе перевода. Алгоритмы морфологического, синтаксического и семантического анализа реализуются в системах машинного перевода на разных уровнях: слов, функциональных групп, предложений. Результат их работы определяется тем, насколько однозначно могут быть приняты решения на основе вариативности результатов анализа на каждом уровне.

Сегодня системы машинного перевода делятся на несколько типов: предметно-ориентированные бинарные системы, системы, работающие на основе использования накопленных примеров (example-based) и системы статистического машинного перевода. Последние два типа систем основаны на использовании результатов переводческой памяти. По сути, все современные варианты систем МП являются гибридными, поскольку сочетают использование архивов систем переводческой памяти с процедурами, реализующими МП для тех фрагментов текста, которые в этих архивах не найдены (ср. [14; 15]).

При условии использования в таком гибридном варианте предметно-ориентированных бинарных систем МП с трансфером, которые являются практическими системами и основаны на предварительном терминологическом анализе соответствующей предметной области, пользователю необходимо учитывать, что автоматический

словарь является ядерной частью любой системы, он предназначен не только для преобразования текста на лексическом уровне, что является нижним уровнем анализа и трансфера, но и для обеспечения работы алгоритмов автоматического синтаксического анализа (парсинга). При анализе результатов работы системы следует иметь в виду, что парсинг осуществляется в рамках одного предложения, а не в пределах сверхфразового единства и тем более не текста как целого. Поэтому с каждым новым предложением система МП как бы начинает анализ заново, теряя информацию о границах именных и глагольных групп, функциональных сегментах, установленную при анализе предыдущего предложения. Разработчики систем МП прекрасно осознают ущербность такого подхода, но он жестко определяется требованием перехода на начальном этапе работы от конкретных лексических единиц к их кодовым обозначениям. Эти семантико-синтаксические коды, суть и разнообразие которых зависит от системы и заданных в ней алгоритмов, являются основой применения универсальных для конкретного языка алгоритмов анализа и синтеза.

Практическая работа переводчика с системой машинного перевода предусматривает:

- подготовку исходного текста (массива текстов) к переводу – ручное предредактирование текста;
- редактирование результатов работы системы МП – ручное постредактирование переводов;
- ведение собственного (пользовательского) словаря, фиксирующего результаты работы с машинными переводами и определяющего настройку системы МП на задачи конкретного переводчика.

При реализации работы на этих этапах следует учитывать ограничения, которые накладываются на результаты работы любой системы МП:

- вследствие локального перевода (перевода по предложениям), эта особенность приводит к тому, что в системе затруднен анализ связей внутри сверхфразового единства и поиск антецедентов, что приводит к неверному переводу местоимений-заместителей. Следовательно, при предварительном редактировании исходного текста необходимо обратить внимание на использование таких заместителей и по возможности заменить их соответствующими знаменательными словами;
- вследствие особенностей работы со словами, отсутствующими в словарях системы (геоназваниями и именами собственными, фир-

менными знаками и редкими словами), что приводит к возможным нарушениям в синтаксическом анализе входного предложения. Кроме того, возможны ситуации неправильного опознавания имен собственных как имен нарицательных и, соответственно, их перевода. При предварительном редактировании следует обратить внимание на использование таких имен и маркировать их так, чтобы не допустить их перевода;

• вследствие вариативности использования терминов в исходном тексте, что может нарушить унификацию перевода терминологии в рамках одного и того же текста. При предварительном редактировании следует проанализировать наиболее частотные номинации (используемые термины), окказиональные аббревиатуры, которые могут совпадать в различных терминологических системах и языках для специальных целей, а также способы использования дефисных конструкций;

• вследствие того, что в реальных текстах встречаются очень длинные предложения, а в системах введены ограничения на длину предложения, при которой синтаксическая структура распознается достаточно устойчиво. Это ограничение может быть снято за счет предварительного редактирования очень длинных предложений. Опыт показывает, что при средней длине предложения в 12 слов результат МП оптимальный, однако в реальном тексте этот показатель очень часто превышается;

• вследствие линейности распознавания устойчивых коллокаций (машинных оборотов), которые составляют большую часть словарного обеспечения любой системы машинного перевода.

Предредактирование текста позволяет заранее снять некоторые ограничения систем МП, оно необходимо для установления единства используемой терминологии, например, в системах извлечения данных (*data mining systems*), в которых часто неверные результаты возникают в результате расхождения между данными, извлекаемыми из текста, и номинацией соответствующих объектов в словарном обеспечении (базах данных или онтологиях). Предредактирование должно использоваться для исправления ошибок и в целом для упрощения текста в связи с решением задач перевода и инженерии знаний. Предредактирование предполагает выполнение следующих действий:

• введение в иноязычный текст артиклей там, где это необходимо или грамматически оправдано;

• повторение элементов при сочинительной связи словосочетаний в предложении;

• введение союзов при использовании бессоюзной связи между предложениями;

• устранение конструкций в скобках в середине именной группы или в середине предложения;

• замена окказиональных аббревиатур на полные наименования либо введение специальных символов, предотвращающее их перевод как обычных слов;

• устранение эллипсисов, неформальных конструкций и метафор;

• приведение к единому виду конструкций, которые могут иметь разное написание.

Лингвистическое обеспечение систем МП обычно реализуется как скоррелированная система автоматических словарей (АС) и грамматических правил. В соответствии с таким подходом автоматический словарь системы МП функционально можно разделить на 4 составные части:

1) словарные статьи так называемых стоп-слов, т.е. служебной лексики, которая определяет привлечение конкретных алгоритмов парсинга;

2) терминологические словарные базы, ориентированные на фиксацию терминов-универбов или многокомпонентных терминов, характерных для использования в конкретных предметных областях или подобластях;

3) словарные статьи общенаучной лексики, используемой практически во всех научных и технических текстах;

4) словарные статьи лексических единиц (слов и словосочетаний), добавляемых пользователем в так называемый пользовательский словарь. Эта часть АС формируется переводчиком и/или терминологом в рамках собственного АРМ и обеспечивает его более тонкую настройку на лексический спектр текстов, предназначенных для перевода.

Каким бы полным и ориентированным на узкую подобласть не был АС, результат МП требует постредактирования как на уровне синтаксической структуры предложения, так и на уровне уточнения и/или изменения переводов отдельных слов и словосочетаний, а также изменения морфологических характеристик рода, числа, падежа, уточнения форм времени и залога, изменения пунктуации. При оценке трудоемкости этого процесса внесение стилистических изменений обычно не рассматривается. Как ни парадоксально, именно этот процесс вызывает неприятие переводчиков и отрицательное отношение к результатам МП в целом. Проведенные исследования

[16] показали, что такое неприятие больше свойственно профессиональным переводчикам, чем тем, кто еще только получает эту профессию. Возможно, это связано еще и с уровнем компьютерной грамотности испытуемых, а также с небольшим объемом перевода. Многолетний опыт собственной работы автора показывает, что работа с постредактированием результатов МП оставляет простор для решения творческих и лингвистических задач, однако обучение постредактированию результатов МП должно составлять обязательную часть подготовки переводчиков.

Постредактирование на лексическом уровне требует уточнения и изменения переводов конкретных лексических единиц, на синтаксическом – преобразования структуры предложения. Например, в случаях перевода с английского языка на русский необходимы: проверка согласования по роду, числу и падежу, уточнение места подлежащего, иногда полная перестройка предложения или переход к непрямой структуре типа *we have → мы имеем → у нас есть*.

Постредактирование результатов МП и получение окончательного варианта перевода текста требует обращения к словарным и энциклопедическим базам данных, выбранным переводчиком и входящим в состав АРМ, а также к заранее выбранным корпусам текстов. В результате работы на этапе собственно перевода формируется пользовательский словарь, уточняющий терминологические особенности конкретного текста. Этот словарь на этапе поддержки выбранной системы машинного перевода включается в ее лингвистические ресурсы.

Таким образом, после завершения перевода конкретного текста должна происходить перенастройка лингвистических ресурсов: пополняться корпус параллельных текстов за счет исходного текста и его перевода, формироваться и/или пополняться пользовательский словарь, включающий терминологию, выявленную и проверенную переводчиком, пополняться база словарей. Только постоянное ведение собственной системы машинного перевода позволяет использовать ее с максимальным эффектом, настраивая словари на необходимую терминологию и выбирая удобные средства и методы постредактирования.

Рассмотрим особенности процесса постредактирования и ведения пользовательского словаря при работе переводчика с результатами МП. Первая часть словаря включает особую строевую лексику, которая задает опорную информацию для реализации алгоритмов трансфера, поэтому эта часть словаря – словарь стоп-слов – является

«неприкосновенной» в том смысле, что никакие ее единицы (вспомогательные и модальные глаголы, союзы, предлоги или омонимы с ними) не должны включаться в пользовательский словарь даже в том случае, если переводчика не устраивает выбранный в системе вариант перевода. Этот вариант при постредактировании может быть исправлен в режиме замены во всем тексте так, как это предпочитает переводчик, но не в словаре.

Вторая (терминологическая) часть словаря включает словарные статьи с выверенными описаниями терминологических единиц и их семантико-синтаксические коды. Эта терминологическая база ориентирована на предметную область и за ее ведение как правило отвечают разработчики системы или терминологи, которым это специально поручается в команде тех, кто готовит и переводит тексты.

К третьей части словаря относятся слова широкой семантики, которые и вызывают самый большой объем постредактирования. Дело в том, что значение и перевод этих слов задается в АС лексическими единицами, определяющими самые обобщенные значения, входящие в объем соответствующего понятия. Значения слов широкой семантики частично уточняются за счет введения в АС фразовых глаголов и словосочетаний. Поскольку в реальном тексте у автора есть большая свобода формирования новых уточняющих словосочетаний, использования низкочастотных или несвойственных научному стилю выражений, то никакой АС не в состоянии включить их все, соответственно, при постредактировании именно эти лексические единицы требуют особого внимания и решения креативных задач.

Пользовательский словарь формируется в результате работы на этапе постредактирования, этот словарь фиксирует терминологические особенности конкретных текстов, с которыми работает переводчик. Исследование результатов МП, научных и технических текстов, а также реального объема постредактирования позволяет рекомендовать особую осторожность при выборе новых лексических единиц и их переводов, включаемых в пользовательский словарь. Необходимо проанализировать весь текст в целом, чтобы понять, насколько эти переводы ему (и не только ему) соответствуют. Только в случае, если соответствие установлено, можно в режиме замены отредактировать все употребления подобных слов и словосочетаний, а затем ввести их в пользовательский словарь для использования при переводе других текстов из той же предметной области.

Сегодня использование машинного перевода в научных, технических и исследовательских проектах, а также в коммерческих це-

лях постоянно растет. Серьезные достижения в качестве результатов машинного перевода привели к широкому использованию МП непрофессионалами для извлечения сути текстов, написанных на незнакомых языках. Соответственно, возникли особые требования к процедурам и технологиям постредактирования [16]. В то же время, можно утверждать, что для полноценного использования системы МП профессиональный переводчик должен представлять себе в общем виде процедуру анализа текста в системе, что позволит ему заранее подготовить текст так, чтобы минимизировать объем постредактирования. Кроме того, пользователи системы МП должны хорошо понимать, что качество результатов машинного перевода зависит от настройки системы АС на задачи конкретного пользователя. Учет спектра и возможностей и ограничений выбранной системы перевода позволит переводчику получать результат, легко редактируемый с помощью современных лингвистических технологий. Корректное использование всего спектра этих технологий сегодня приобретает особую важность.

ЛИТЕРАТУРА

1. Vasiljevs, A., Pinnis, M., Gornostay, T. Service model for semi-automatic generation of multilingual terminology resources // Terminology and Knowledge Engineering 2014, Jun 19–21, 2014. – pp. 67–76.

2. Беляева Л.Н. Лингвистические технологии в современном сетевом пространстве: *language worker* в индустрии локализации. − СПб.: Изд-во ООО «Книжный дом», 2016. – 134 с.

3. Беляева Л.Н., Джепа Т.Л., Зак Г.Н., Камшилова О.Н., Нымм В.Р., Разумова В.В. Автоматизированное рабочее место филолога в структуре образовательного пространства современного вуза. − СПб.: Изд-во ООО «Книжный дом», 2013. − 123 с.

4. Steinberger R. Language Engineering Technologies and their use for TF−UCLAF. A Report on JRC's Institutional Support Activities // электронный ресурс http://langtech.jrc.it/Documents/Report-98_Steinberger_LangTech4OLAF.pdf]

5. Rychtyckyj N. An Assessment of Machine Translation for Vehicle Assembly Process Planning at Ford Motor Company// S. D. Richardson (ed): AMTA 2002, Lecture Notes in Computer science, Vol. 2499, Berlin Heidelberg: Springer-Verlag, 2002. – pp. 207–215.

6. Knebel M., Ralf F. DITA Customization – Create Your Own Flavor // tekom-Jahrestagungund tcworld conference in Stuttgart. Zusammenfassungen der Referate. Stuttgart: tcworld GmbHVerantwortlich. 2016. – pp. 51–53.

7. Погосов А.А. Развитие переводческого процесса: подход современных ученых // Вестник Военного университета, 2009, № 4 (20). – с. 109–114

8. Ермаков А.Е. Язык семантических трансформаций для компьютерной интерпретации текста // Информационные технологии, 2017, Том 23, № 6, – с. 403–412.

9. Gollner, J.Information 4.0 for Industry 4.0 // Towards a European Competence Framework // tekom-Jahrestagungundtcworld conference in Stuttgart. ZusammenfassungenderReferate – Stuttgart:tcworldGmbHVerant wortlich, 2016. – pp. 93–94.

10. *Gallon R*. Information 4.0, the Next Steps // Towards a European Competence Framework // tekom-Jahrestagungundtcworld conference in Stuttgart. Zusammenfassungen der Referate – Stuttgart:tcworld GmbH Verantwortlich, 2016. – Pp. 95–97.

11. Lacroix, F. Writing for the 21st Century // Towards a European Competence Framework // tekom-Jahrestagungundtcworld conference in Stuttgart. Zusammenfassungen der Referate– Stuttgart:tcworldGmbHVeran twortlich, 2016. − pp. 102–106.

12. Беляева Л.Н., Гейхман Л.К., Камшилова О.Н. Компетентностный потенциал современного переводчика: проблемы лингвообразования // Профессиональное лингвообразование: материалы девятой международной научно-практической конференции. Июль 2015 г. – Нижний Новгород: НИУ РАНХиГС, 2015. – с. 337–350.

13. Meex, B., Karreman, J. TecCOMFrame. Towards a European Competence Framework // Towards a European Competence Framework // tekom-Jahrestagungund tcworld conference in Stuttgart. Zusammenfassungen der Referate – Stuttgart: tcworld GmbHVerantwortlich, 2016. – pp. 486–489.

14. Dandapat, S., Morrissey S., Way A., Forcada M. L. Using example-based MT to support statistical MT when translating homogeneous data in a resource-poor setting // Proceedings of the 15th annual meeting of the European Association for Machine Translation (EAMT 2011), 2011 – pp. 201–208.

15. Dandapat, S., Morrissey S., Way A., van Genabith J... Combining EBMT, SMT, TM and IR technologies for quality and scale // Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra), 2012 – pp. 48–58.

16. Moorkens J., O'Brien S. Post-Editing Evaluations: Trade-offs between Novice and Professional Participants // EAMT 2015. Proceedings of the 18th Annual Conference of the European Association for Machine Translation. Antalya, Turkey, May 11–13, 2015. – pp. 75–81.

УДК 81-13

NONVERBAL ELEMENTS IN EVERYDAY RUSSIAN SPEECH: AN ATTEMPT AT CATEGORIZATION

N. V. Bogdanova-Beglarian, E. M. Baeva
St Petersburg state university, St Petersburg
n.bogdanova@spbu.ru, e.baeva@spbu.ru

The article provides an attempt at systematization of the elements of oral discourse which are not related to the text content but are nonetheless very frequent in everyday speech and thus essential for its understanding and decoding.

Keywords: modern Russian, everyday speech, nonverbal vocalizations, paralinguistic elements, speech corpus, hesitation phenomena.

НЕРЕЧЕВОЕ В ПОВСЕДНЕВНОЙ РУССКОЙ РЕЧИ: ОПЫТ КАТЕГОРИЗАЦИИ

Н. В. Богданова-Бегларян, Е. М. Баева
Санкт-Петербургский государственный университет, Санкт-Петербург
n.bogdanova@spbu.ru, e.baeva@spbu.ru

В статье приводится систематизация таких элементов устного дискурса, которые никак не связаны с содержанием текста, но очень частотны и, как правило, заполняют хезитационные паузы, маркирующие «точки сбоя». Описывается, какие именно элементы неречевой коммуникации можно обнаружить в повседневной русской речи, и предлагается способ их категоризации с целью дальнейшего изучения.

Ключевые слова: русская речь, повседневная речь, неречевые элементы, паралингвистические элементы, хезитационные явления.

«Неречевые» элементы звуковой цепи выделяются на фоне полнозначных речевых, передающих основное содержание текста, а также во многом прагматических условно-речевых единиц, характеризующихся высокой употребительностью и повторяемостью в речи разных говорящих, помогающих структурировать текст, но не связанных напрямую с его содержанием [1]. К условно-речевым единицам, помимо вводных и служебных слов, относятся весьма частотные в устной речи прагматические маркеры, такие как поисковые вербальные хезитативы (*это самое, как его (её, их), ну*

это и под.), дискурсивы разного типа (*значит, вот, ну*), маркеры-ксенопоказатели (*типа (того что), такой, грит/гыт*), маркеры-аппроксиматоры (*то-сё, пятое-десятое, и все дела*), рефлексивы (*так скажем, или как его*) и ряд других [2; 3]. Неречевые элементы звуковой цепи представляют собой разновидности речевых сбоев, которые нарушают плавное развертывание речевого потока, беглость речи [4] и, как показал анализ корпусного материала, могут быть отнесены к невербальным прагматическим маркерам, так как выполняют в устном дискурсе вполне определенные функции. Из «неречевых» элементов наиболее распространенными считаются хезитационные паузы, заполненные нефонемными звуками (так называемые вокализации – э-э, *а-а, м-м*), паралингвистические элементы (смех, вздох, кашель), а также разного рода звуковые артефакты [5], также часто имеющие хезитационный характер.

В научной литературе удалось найти примеры исследований неречевых элементов дискурса, в основном на материале английского языка [6-8]. Наиболее подробно изучен феномен хезитационных пауз; в работах по устной речи и фонетике также встречаются упоминания о паралингвистических элементах [9-11], цоканье [12-14], в том числе на материале китайского и некоторых африканских языков [15-16], и причмокивании [17]. Что касается русской устной речи, то последнее время особенно популярны исследования прагматических единиц [18-22], однако невербальная часть устной повседневной коммуникации остается освещенной крайне скудно.

Задачей данного исследования было обозначить, какие именно элементы неречевой коммуникации можно найти в повседневной русской речи, и предложить способ их категоризации, с целью дальнейшего изучения.

Источником материала стали два модуля Звукового корпуса русского языка: корпус повседневной русской речи «Один речевой день» (ОРД) и сбалансированная аннотированная текстотека (САТ). Корпус ОРД (130 информантов, более 1000 их коммуникантов, 1250 часов звучания и 1 млн словоформ в расшифровках) фиксирует наиболее естественную речь носителей русского языка и содержит по преимуществу повседневные диалоги и полилоги, записанные по методике непрерывного суточного речевого мониторинга [23]. САТ (800 текстов, 50 часов звучания) содержит менее естественную экспериментальную речь – монологи, записанные от носителей русского языка из разных профессиональных групп и построенные в рамках серии коммуникативных сценариев, типичных для повседневного

общения: чтение, пересказ, описание изображения, рассказ. Помимо речи носителей русского языка, CAT включает также несколько блоков интерферированной русской речи американцев, французов, китайцев и голландцев [24].

В результате анализа в материале удалось выделить следующие неречевые элементы: заполненные и незаполненные паузы хезитации [25], цоканье языком, причмокивание, шумное втягивание воздуха (хлюпанье), различные паралингвистические элементы (смех, кашель, вздох, зевок, чихание и др.) [26]. Представляется, что данные элементы можно анализировать с учетом их прагматических функций. Одной из основных в данном случае оказывается функция хезитационного поиска (конкретной единицы или вообще продолжения речи), при этом часто неречевые элементы идут в соседстве с другими – вербальными – хезитативами. Несомненная поисковая функция невербальных вокализаций осложняется иногда другими – рефлексивной функцией или функцией начала/конца речевого фрагмента, ср. (**Ц* – цоканье, **В* – шумный вдох, *мп* – причмокивание), ср.:

ой / <смех> я чего-то / я чего-то / я чего-то запомнила только конец / как они кота накормили ***это самое*** */ он начал* ***это самое*** */* ***э-э ну это как его // э-э ну з… ж…*** */ ну жареной свининой /* ***значит / э*** *окунями*;

ой я не могу с ней ****В*** *// накаталась на велике //* ****Ц*** *о(:)*;

н-нужно / мне нужно / ***мп*** *а собира... / собирать там / материал для / анализа*.

Паралингвистические элементы устной речи – и в целом, как класс элементов, и каждый по отдельности – достойны, вне всякого сомнения, специального описания, однако уже приведенные примеры демонстрируют и их явный хезитационный характер, и также некоторую полифункциональность.

Проведенное исследование на обширном корпусном материале подтверждает существование в устном дискурсе большого количества элементов, которые, с одной стороны, не претендуют на статус значимых, речевых, и не могут быть описаны как вербальные, а с другой стороны, обладают явным прагматическим значением и так или иначе помогают говорящему структурировать порождаемый текст. Основной функцией таких неречевых фрагментов звуковой цепи следует признать хезитационно-поисковую, которая часто осложняется другими: функциями дискурсивного маркера (стартового или финального), рефлексива или маркера «нетривиального»,

которого довольно много в повседневной речи. Корпусный подход к анализу устной речи позволяет не только выявить все такие неречевые элементы, но и систематизировать их определенным образом, а также использовать во многих прикладных целях: от преподавания русского языка в иноязычной аудитории до автоматического распознавания речи и лингвистической экспертизы.

Благодарности. Работа выполнена при финансовой поддержке РНФ (проект № 18-18-00242 «Система прагматических маркеров русской повседневной речи»).

ЛИТЕРАТУРА

1. Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 1. Чтение. Пересказ. Описание / Отв. ред. Н. В. Богданова-Бегларян. СПб, 2013. 532 с.

2. Богданова-Бегларян Н. В. Прагматемы в устной повседневной речи: определение понятия и общая типология. Вестник Пермского университета. Российская и зарубежная филология. 2014. № 3 (27). С. 7–20.

3. Bogdanova-Beglarian N. V., Filyasova Yu. A. Discourse vs. Pragmatic Markers: a Contrastive Terminological Study. 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts. SGEM 2018. Vienna ART Conference Proceedings, 19-21 March, 2018, vol. 5, 2018. Pp. 123–130.

4. Verdonik D., Rojc M., Stabej M. Annotating Discourse Markers in Spontaneous Speech Corpora on an Example for the Slovenian Language. Language Resources and Evaluation. The Netherlands. Iss. 41 (2), 2007. Pp. 147–180.

5. Кипяткова И. С., Верходанова О. В., Ронжин А. Л. Сегментация паралингвистических фонационных явлений в спонтанной русской речи // Вестник Пермского университета. Российская и зарубежная филология. Вып. 2 (18), 2012. С. 17–23.

6. Crystal D. Prosodic Systems and Intonation in English. Cambridge: Cambridge University Press. 1969.

7. Trouvain J., Truong K. Comparing non-verbal vocalizations in conversational speech corpora. Proc. 4th Int'l Workshop on Corpora for Research on Emotion Sentiment & Social Signals, Istanbul, vol. 36–39, 2012.

8. De Jong N. H., & Bosker H. R. Choosing a threshold for silent pauses to measure second language fluency. In The 6th Workshop on Disfluency in Spontaneous Speech (DiSS), 2013, pp. 17–20.

9. Scherer, K. R. Affect bursts. Emotions. New York: Psychology Press. 2014, pp. 175–208.

10. Schröder, M. Experimental study of affect bursts. Speech communication, 40(1-2). 2003, pp. 99–116.

11. Trouvain J. Laughing, breathing, clicking – The prosody of nonverbal vocalisations. Proc. Speech Prosody. 2014, pp. 598–602.

12. Wright M. On clicks in English talk-in-interaction. Journal of the International Phonetic Association 41(2), 2011, pp. 207–229.

13. Wright M. Clicks as markers of new sequences in English conversation. 16th International Congress of the Phonetic Sciences (ICPhS XVI), Saarbrucken. 2007, pp. 1069–1072.

14. Trouvain J., & Malisz Z. Inter-speech clicks in an Interspeech keynote. INTERSPEECH 2016. International Speech Communication Association. 2016, pp. 1397–1401.

15. Maddieson I. Patterns of sounds. Cambridge: CUP. 1984.

16. Li A., Zheng,F., Byrne W., Fung P., Kamm T., Liu Y., ... & Chen X. CASS: A phonetically transcribed corpus of Mandarin spontaneous speech. In Sixth International Conference on Spoken Language Processing. 2000.

17. Li Y., He Q., Li T., & Wang W. A detection method of lip-smack in spontaneous speech. In Audio, Language and Image Processing, 2008. ICALIP 2008. International Conference on. IEEE. (2008, July), pp. 292–297.

18. Дискурсивные слова русского языка: опыт контекстно-семантического описания / Ред. Киселева К. Л., Пайар Д. Метатекст. 1998.

19. Дискурсивные слова русского языка. Контекстное варьирование и семантическое единство/ Ред. Киселева К. Л., Пайар Д. М: Азбуковник. 2003.

20. Bolden G. B. Reopening Russian conversations: The discourse particle–to and the negotiation of interpersonal accountability in closings. Human Communication Research, 34(1), 2008, pp. 99–136.

21. Sherstinova T. Macro episodes of Russian everyday oral communication: towards pragmatic annotation of the ORD speech corpus. International Conference on Speech and Computer, Springer, Cham 2015, pp. 268-276.

22. Dobrovol'skij D., & Pöppel L. Corpus perspectives on Russian discursive units: semantics, pragmatics, and contrastive analysis. Yearbook of Corpus Linguistics and Pragmatics 2015, Springer, Cham, 2015, pp. 223–241.

23. Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation.

In: Matoušek, V., Mautner, P. (eds.) TSD 2009. LNAI, vol. 5729. Springer, Berlin-Heidelberg, 2009. Pp. 250–257.

24. Богданова-Бегларян Н.В., Шерстинова Т.Ю., Зайдес К.Д. Корпус “Сбалансированная аннотированная текстотека”: методика многоуровневого анализа русской монологической речи // Анализ разговорной речи (AR3-2017): труды седбмого междисциплинарного семинара / Ред. Кочаров Д.А., Скрелин П.А. СПб: Политехника-Принт, 2017, С. 8–13.

25. Баева Е.М. Хезитационные явления в устных монологах низкой степени спонтанности // Комуникативные исследования. Вып. 1 (15), 2018. С. 75–84.

26. Богданова-Бегларян Н.В. «Нетривиальное» в повседневной устной коммуникации: опыт систематизации // Коммуникативные исследования. Вып. 4 (14), 2017. С. 9–30.

УДК 81.32

DYNAMICS OF THE NUMBER OF SYNTACTIC DEPENDENCIES IN RUSSIAN AND ENGLISH

V. V. Bochkarev, V. D. Solovyev, A. V. Shevlyakova
Kazan Federal University, Kazan
vbochkarev@mail.ru, maki.solovyev@mail.ru, anna_ling@mail.ru

In this work, we use the Google Books Ngram diachronic corpus to study the dynamics of the number of syntactic dependencies and 2-grams in Russian and English. We counted the total number of 2-grams and syntactic dependencies detected in Google Books Books Ngram at least once in a given year, as well as stable and effective dependencies. It was found that quantitatively unchanged core and rapidly growing periphery can be distinguished among the syntactic dependencies of words. The obtained results also show that the effective number of dependencies doubles approximately every 250 years.

Keywords: Google Books Ngram, syntactic dependencies.

ДИНАМИКА ЧИСЛА СИНТАКСИЧЕСКИХ СВЯЗЕЙ В РУССКОМ И АНГЛИЙСКОМ ЯЗЫКАХ

В. В. Бочкарев, В. Д. Соловьев, А. В. Шевлякова
Казанский федеральный университет, Казань
vbochkarev@mail.ru, maki.solovyev@mail.ru, anna_ling@mail.ru

В работе исследуется динамика количества семантико-синтаксических связей и 2-грамм в русском и английском языках на материале диахронического корпуса Google Books Ngram. Проведены подсчеты общего числа 2-грамм и связей, отмеченных в базе Google Books Ngram хотя бы один раз в тот или иной год, а также только устойчивых. Подсчитано эффективное число связей, выражающееся через перплексию распределения частот 2-грамм. Данная величина по смыслу представляет собой характерное число интенсивно использующихся словосочетаний. Показано, что среди синтаксических связей слов выделяется некое устойчивое, медленно изменяющееся в количественном отношении ядро и быстро растущая периферия. Для русского языка получены оценки, показывающие что удвоение эффективного числа связей происходит приблизительно за 250 лет.

Ключевые слова: Google Books Ngram, синтаксические связи.

Появление сверхбольших корпусов текстов и разработка новых алгоритмов и методов лингвистических исследований открывает

широкие возможности для изучения динамических процессов, происходящих в языке, и позволяет проследить эволюцию языковых явлений.

Наиболее интересной в контексте исследования динамики языка и нашей работы представляется является статья [1], в которой исследовался рост числа уникальных словосочетаний в английском языке на материале корпуса Google Books Ngram. Увеличение числа словосочетаний и связей в данной работе интерпретируется как следствие увеличения сложности культуры.

Принимая во внимание выводы [1], мы проанализировали динамику количества синтаксических связей и 2-грамм. Априори можно ожидать, что число связей слов увеличивается с течением времени благодаря двум факторам: 1) возрастанию сложности человеческой культуры и появлению новых слов, обеспечивающих рост числа семантико-синтаксических связей; 2) процессам метафоризации, которые также увеличивают количество связей между словами. Также число зафиксированных в корпусе связей растет вследствие увеличения объема корпуса. Задачей исследования было выявить, каким образом происходит прирост количества указанных связей, проследить влияние каждого из указанных факторов.

Для анализа мы использовали общий корпус английского языка и корпус русского языка из состава Google Books Ngram. Анализировались база данных частот 2-грамм и база частот синтаксических связей в составе этих корпусов.

При анализе 2-грамм мы использовали два подхода. В первом случае в подсчет включались не все сочетания и связи, а только часто повторяющиеся. В качестве меры ассоциативной связи мы использовали величину, называемую в компьютерной лингвистике mutual information (MI). Второй подход состоит в подсчете числа словосочетаний с учетом их информационного содержания. Для этого мы можем использовать такую характеристику распределения частот как перплексия. Наш подход близок к используемому в [1], однако использование перплексии вместо энтропии позволяет представить результаты более наглядно, а также провести сопоставления с оценками, полученными другими способами.

Число связей между словами, зафиксированными в корпусе Google Books Ngram, растет чрезвычайно быстро, увеличившись с 1800 по 2000 год в 160 раз для общего корпуса английского языка и в 66 раз для корпуса русского языка. Ясно, что большая часть этого роста связана не с действительным усложнением языка, а с

экстенсивным ростом объема корпуса. Для того, чтобы изучать факторы, обуславливающие усложнение языка, удобнее пользоваться не общим числом связей, а числом устойчивых связей (с показателем ассоциативной связи выше заданного порога), либо эффективным числом связей (вычисляемым как перплексия распределения частот). Последняя характеристика демонстрирует гораздо более плавное и закономерное изменение по сравнению с общим числом связей. Кривая эффективного числа связей практически не реагирует на исторические события и при расчете по всему объему лексики показывает для английского языка рост по закону, близкому к экспоненциальному. В то же время число связей между словами, фиксируемых в корпусе каждый год в течении достаточно длинного интервала времени (1750–2008 для английского и 1920–2008 для русского) меняется очень медленно. Это может быть индикатором того, что среди синтаксических связей слов выделяются некое устойчивое, слабо изменяющееся в количественном отношении ядро и быстро растущая периферия.

Установлено, что из факторов, влияющих на рост числа связей между словами, доминируют эффекты, связанные с появлением новых слов. Зависимость общего числа синтаксических связей и словосочетаний от числа уникальных слов близка степенной. Разумеется, степенной закон следует рассматривать лишь как некоторую аппроксимацию эмпирических данных. Однако нужно отметить, что степенная зависимость в данном случае лучше соответствует эмпирическим данным, чем для зависимости числа связей от объема корпуса (что ожидается в соответствии с законом Хипса). То же можно сказать и для числа устойчивых связей (с MI>0). При этом показатели степени для общего числа связей для исследуемых языков несколько больше единицы (1.1-1.17), а для числа только устойчивых связей – меньше (0.79-0.96). Эти факты должны учитываться при построении моделей роста сети синтаксических связей в естественных языках.

Для русского языка удается получить оценку темпа роста эффективного числа синтаксических связей в языке. При неизменном объеме корпуса удвоение числа связей должно происходить приблизительно за 250 лет. Для английского языка на протяжении длительного периода характерны близкие темпы роста эффективного числа синтаксических связей, которые, однако увеличиваются примерно после 1950 года. По-видимому, это является следствием роли английского языка как глобального.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект №17-29-09163).

ЛИТЕРАТУРА

1. Juola P. Using the Google N-Gram corpus to measure cultural complexity / P. Juola // Literary and Linguistic Computing. – 2013. – № 28(4). P. 668–675.

УДК 004.912.021:025.4.026:81'322.2

AN OVERVIEW OF THE AVAILABLE CORPORA FOR EVALUATION OF THE AUTOMATIC KEYWORD EXTRACTION ALGORITHMS

A. S. Vanyushkin[1], L. A. Graschenko[2]
[1]Pskov State University, Pskov
[2]Institute of Mathematics named A. Juraev of the Academy of Sciences of the Republic of Tajikistan, Dushanbe
alexmandr@mail.ru, graschenko@mail.ru

The article discusses the evaluation of automatic keyword extraction algorithms (AKWEA) and points out AKWEA's dependence on the properties of the test collection for effectiveness. As a result, it is difficult to compare different algorithms who's tests were based on various test datasets. It is also difficult to predict the effectiveness of different systems for solving real-world problems of natural language processing (NLP). We take in to consideration a number of characteristics, such as the text length distribution in words and the method of keyword assignment. Our analysis of publicly available analytical exposition text which is typical for the keywords extraction domain revealed that their length distributions are very regular and described by the lognormal form. Moreover, most of the article lengths range between 400 and 2500 words. Additionally, the paper presents a brief review of eleven corpora that have been used to evaluate AKWEA's.

Keywords: text corpus, corpus linguistics, keyword extraction, text length distribution, natural language processing, information retrieval.

ОБЗОР ДОСТУПНЫХ КОРПУСОВ ДЛЯ ОЦЕНИВАНИЯ АЛГОРИТМОВ АВТОМАТИЧЕСКОГО ИЗВЛЕЧЕНИЯ КЛЮЧЕВЫХ СЛОВ

А. С. Ванюшкин[1], Л. А. Гращенко[2]
[1]Псковский государственный университет, Псков
[2]Институт математики им. А. Джураева Академии наук Республики Таджикистан, Душанбе
alexmandr@mail.ru, graschenko@mail.ru

В статье представлен обзор одиннадцати доступных в открытом доступе текстовых корпусов, которые использовались в различных работах по автоматическому извлечению ключевых слов (АИКС). Произведено их сравнение с естественными коллекциями текстов аналитической направ-

ленности, для которых такая задача представляется актуальной. Показано, что для естественных коллекций наблюдается логнормальное распределение длин текстов с рабочим диапазонов длин в пределах 400-2500 слов. В связи с этим, для сравнения между собой различных алгоритмов АИКС предлагается использовать наиболее близкий по параметрам распределения длин естественным коллекциям текстов корпус DUC-2001, а также разработать новый двуязычный корпус, удовлетворяющий рассмотренным в статье требованиям.

Ключевые слова: текстовый корпус, корпусная лингвистика, извлечение ключевых слов, распределение длин текстов, обработка текстов на естественных языках, информационный поиск.

1. Introduction

The number of digital documents available is growing on a daily basis at an overwhelming rate. As a consequence, there is a need to increase the complexity of the structure and software solutions in the field of NLP which are based on a number of basic methods and algorithms. The algorithms of automatic keyword and key phrase (KW) extraction are among them. This task has been analyzed over the past sixty years from different perspectives. There has been a significant increase in the number of research that took place in the last twenty years, of which many have been publications of different AKWEA's [1]. The reason for this is the increasing amount of computing research, data resources and especially the development of internet services. It also simplifies the development and evaluation of new algorithms. This trend is clearly illustrated in Figure 1, obtained using Google Books Ngram Viewer.

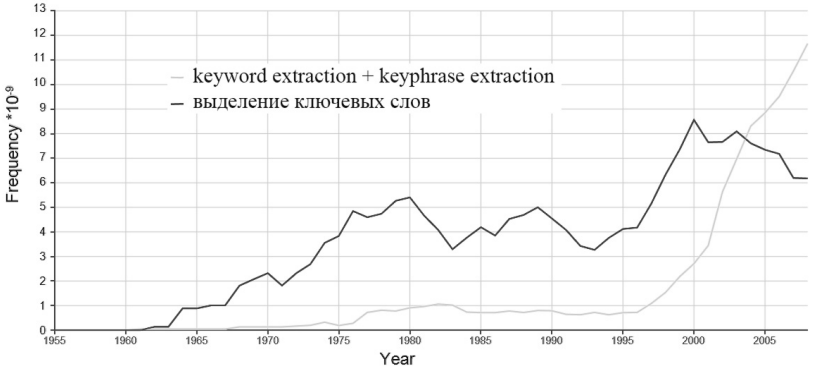


Figure 1. Usage of phrases 'keyword extraction', 'keyphrase extraction' (Russian: 'выделение ключевых слов') found in the Google Books Dataset

The term “keyword” is interdisciplinary and above all, is used in works on psycholinguistics and Information Retrieval [2] that causes the existence of different approaches to its definition. Summarizing the numerous opinions, we can conclude that the keywords (phrases) are words (phrases) in the text that are especially important, commonly understood, capacious and representative of a particular culture. The set of which can give a high-level description of its content for the reader and providing a compact representation and storage of its meaning in mind [1]. In practice, the terms keyword and key phrase have the same meaning.

Despite the large amount of specialized and interdisciplinary work there has not been a consistent technique developed for detecting keywords yet. Experiments confirmed that this is done intuitively by people, and is personality, and even gender-based [3]. This implies the non-triviality of the development of formal methods and KW extraction algorithms for computing. Therefore, the current efforts of researchers are focused on the development and implementation of hybrid learning-based AKWEA’s which assumes the use a variety of linguistic resources. Thus, the accuracy of training and control datasets has great importance on the effectiveness of development.

Our analysis reveals number problematic areas. The author’s results in testing AKWEA’s are often different from those obtained by other researchers, since they use different control data in the evaluation of algorithms [1]. Independent testing of KW extraction algorithms is a difficult task because there is a lack of implemented system and source code of algorithms in open access. This problem is partially solved by carrying out workshops when the organizers propose test data for all participants. At the same time the number of available and well-proven corpora for KW extraction evaluation is small (10-20) and the criteria for their formation are not methodologically well enough investigated. The possibility of transferring the results of the algorithms in other languages remains an open question. The remarkable thing is that most of the known results are obtained for the English language, and the rules for the interpretation of them to the Slavic languages, especially to Russian, have not been established.

Indeed, preliminary empirical data show that for the graph-based algorithms with increased text size the precision of AKWEA’s might reduce. Therefore, the effectiveness of the algorithms depends on the type and parameters of the text lengths distribution (in words) that constitute research data. Homogeneity of the data by genre and text difficulty probably has some influence on the effectiveness of AKWEA’s too, Figure 2.

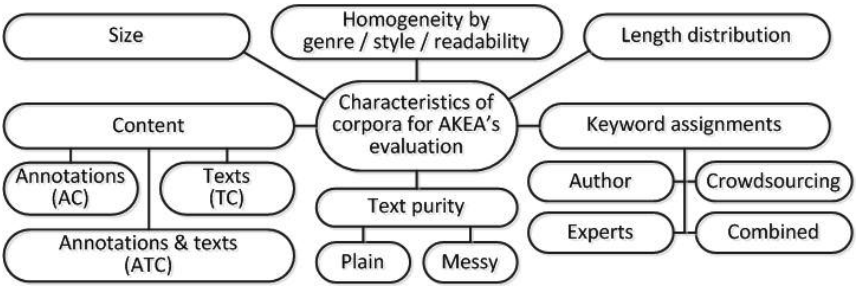


Figure 2. The specifications of research corpora for keyword extraction evaluation

A separate discussion is necessary to explore the characteristics of experimental corpora such as size, existence and the methods of KW assignment (who and how many authors assigned them), the subject and the type of text (abstracts and full articles). KW assignment can be performed by authors, experts on the topic or by crowdsourcing. In this case, questions arise such as what kind of assignment is considered optimal, is it possible to rely on public opinion and what is a minimum number of participants that must specify the word as a keyword to assign it as such. It should be noted that the quality of KW assignment depends on the size of a corpus. As the size increases, the complexity of assignment rises.

But first of all it is necessary to investigate existing text collections (those used for KW extraction) for the length distribution parameters (in words).

2. Methodology and Research Tools

Articles from six web sites were selected as the statistical and research database subset that contains a voluminous collection on various English topics. This choice is due to the assumption that the main sphere of work for KW extraction is mostly with topical or subject-based text, especially those that contain elements of analytical themes. The eleven corpora (test and trial), that were used in some or other research or scholarly articles, were found using a search engine.

Many sites block automatic downloading for article collection or don't have freely available archives for use at all. So sites with freely available resources were used. After downloading the collection of articles, automatically parsing of the pages was made and the text was extracted. Then the tokenization and a count of the number of words in each

article was made. Stanford Log-linear Part-Of-Speech Tagger[1] was used for tokenization of English texts, which is widely used in both research and commercial sectors [4].

The text lengths distributions in words were presented for every collection. We used Pearson's chi-squared test to evaluate the fitness of observed data to some theoretical distributions using advanced analytics software package *Statistica[2] and EasyFit[3] software. It is worth pointing out that the form distribution depends on the mode of data grouping [5]. Calculating the number of bins k* in different ways leads to a wide range of its possible values. For the expected Gaussian distribution, the *Sturges* formula is normally used, but if the data are not normal or there are more than 200 cases, it's poorly applied [6].

For the unification of the calculation the bin sizes in the histograms we used the *Freedman and Diaconis* rule, which gives the value agreed with the recommendations on standardization[4] and then convert it into the number of bins:

$$h = 2\left(IQ\right)n^{-\frac{1}{3}}, \tag{1}$$

where h is the bin size, IQ is the interquartile range of the data and n is the number of observations. At the same time according to the Pearson's chi-squared test (p-value = 0.05) we did not obtain a satisfactory fit of the results in all cases. Our hypothesis was confirmed by varying k in a small range with respect to the calculated value. To improve the accuracy of estimates of the form and parameters of the probability density function further research is needed. For example, the *Levenberg-Marquardt* algorithm was used by other researches to solve similar problems [7].

3. A Review of Existing Information Resources

3.1. Text Length Distributions in Analytical Articles Collections

The issue of natural length distribution and optimal lengths are taken into consideration by many researches. Most studies have been devoted to investigate blog post sizes [8-10], which describes the text length dis-

[1] http://nlp.stanford.edu/

[2] https://www.quest.com/

[3] http://www.mathwave.com/

[4] R 50.1.033-2001. Applied statistics. Rules of check of experimental and theoretical distribution of the consent. Part I. Goodness-of-fit tests of a type chi-square

tribution with fat tails. This is true for the user comments [7], e-mail messages [11] and for the length of the texts that are stored on users' computers [12]. It is proposed [13] to consider the length of the articles from Wikipedia encyclopedia as an indicator of their quality, and the overall length of the English papers described by the lognormal form [14]. Figure 3 presents the probability density function distributions for the six data sets.

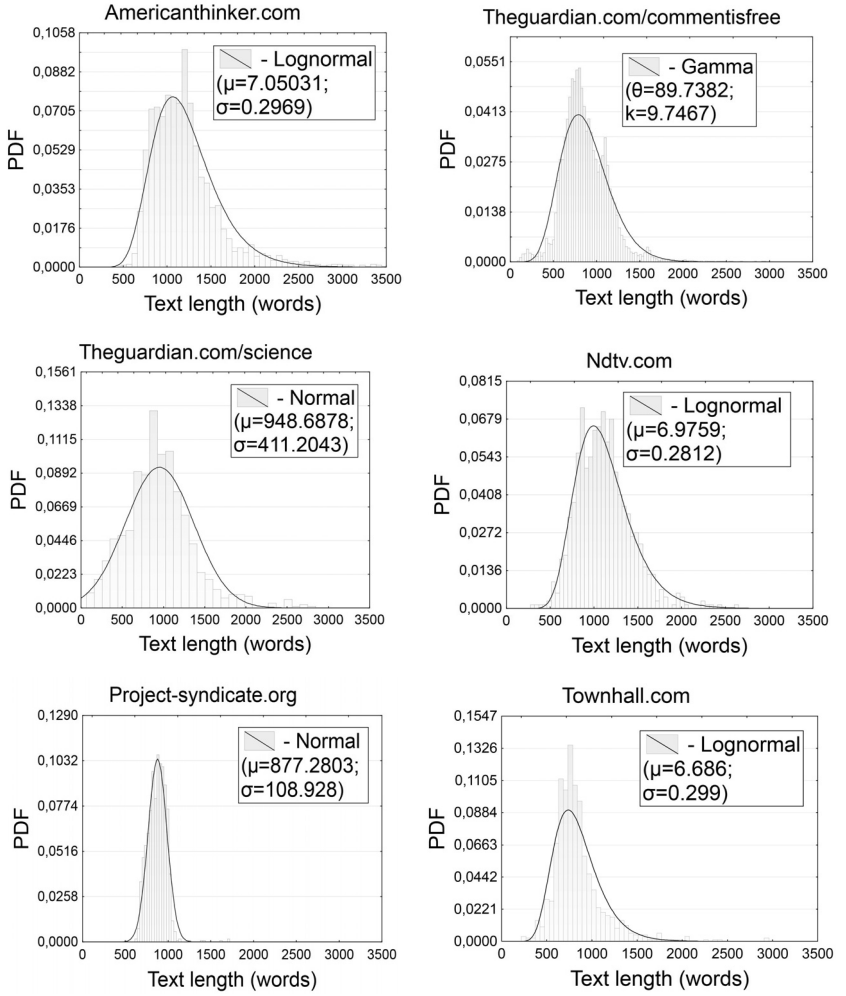


Figure 3. Distribution of analytical articles lengths in words

As can be seen from the graphs, the majority of the length distribution of analytical articles can be comparative to the normal or lognormal form. The majority of texts are in the range of 400 to 2500 words.

Table 1 presents general information and statistical characteristics of the reviewed text collections. Collection size ranges from 736 to 14529 articles and their publication dates cover the period from 2015 to 2016. Mean lengths of articles varies between 839-1212 words.

Table 1. **Characteristics of the analytical articles collections**

№	Source	Count	Text length				Publishing period
			Mean	Min.	Max.	Std. Dev.	
1	project-syndicate.org	1163	873,3	612	1721	108,9	01.15-12.15
2	ndtv.com	736	1112,5	274	2650	309,9	01.15-12.15
3	americanthinker.com	2268	1212,2	473	3703	410,4	01.15-02.15
4	townhall.com	905	839,5	217	2960	283,9	07.15-12.16
5	theguardian.com/ science	897	948,7	66	2848	411,2	01.15-12.16
6	theguardian.com/ commentisfree	14529	874,6	79	3045	278,8	01.15-12.16

It is worth pointing out that there are possible restrictions authors can have on the length of published articles. For example, on project-syndicate.org a recommended article length by their editorial team is 1000 words.

3.2. Existing Corpora for Keyword Extraction Evaluation

Despite the large number of works devoted to keyword extraction evaluation the number of specially trained and public corpora are much less so. Some of them are used multiple times in different studies. *Hulth-2003* [15] for example, consisting of abstracts of scientific articles, is one of the most popular and was used in the many academic papers [16-22]. Other datasets are used much less frequently, often only by their authors. One of the main drawbacks of such corpora is the “messy” texts, as many of them contain a bibliography, tables, captions and pictures in text files.

We surveyed eleven public corpora, which are significantly different from each other such as the text length distribution as well as other characteristics such as the size, themes and authorship of the keyword assignment. Table 2 summarizes the characteristics of reviewed corpora. The following are some explanations.

Table 2. **Characteristics of the available corpora for KW extraction evaluation**

№	Corpus	Year	Contents	KW assign	Type	Resource
1	DUC-2001 [23]	2001	News articles	E-2	AT	github.com
2	Hulth-2003 [15]	2003	Paper abstracts from Inspec 1998-2002	E-?	A	researchgate.net
3	NLM-500 [24], [25]	2005	Full papers of PubMed documents	E-?	AT	github.com
4	NUS [26]	2007	Scientific conference papers	A+E-?	AT	github.com
5	WIKI-20 [27], [28]	2008	Technical research reports of computer science	E-15	AT	github.com
6	FAO-30 [28], [29]	2008	Documents from UN FAO[1]	E-6	T	github.com
7	FAO-780 [28], [29]	2008	Documents from UN FAO	E-?	T	github.com
8	KRAPIVIN [30]	2009	ACM[2] full papers 2003-2005	A	AT	disi.unitn.it
9	CiteULike [28], [31]	2009	Bioinformatics papers	O-3	T	github.com
10	SemEval-2010 [32]	2010	ACM full papers	A+E-0,2	AT	github.com
11	500N KPCrowd-v1.1 [33]	2012	News articles	O-20	T	github.com

Note: notation of KW assignment: A-text authors, O-N – Crowdsourcing (N – number of people per one text, ? - n/a), E-experts.

Corpus type: A – annotation, AT – annotation + text, T – the main body of the text.

Let us explain the features of the KW assignment of the given corpora. *DUC-2001* was prepared for text summarization evaluation within the Document Understanding Conferences, but KW assignment was made by two only graduate students in 2008 for the study of AKWEA's [23]. A feature of the *Hulth-2003* assignment is the presence of two sets of KW – a set of controlled, i.e. terms restricted to the *Inspec* thesaurus, and a set of uncontrolled terms that can be any terms. *NLM-500* sets of key-

[1] Food and Agriculture Organization.
[2] Association for Computing Machinery.

words restricted to the thesaurus of Medical Subject Headings. *WIKI-20* assigned by 15 teams consisting of two senior computer science undergraduates each. These KW sets were restricted to the names of Wikipedia articles. *NUS* has the author's assigned KW lists as well as KW lists assigned by student volunteers.

FAO-30 and *FAO-780* differ in size and composition of the experts, but both KW sets were restricted to the *Agrovoc*[1] *thesaurus. In KRAPIVIN* parts of the articles are separated by special characters, which makes it convenient to their separate processing. *CiteULike* KW's were assigned by 322 volunteers but the authors noted that for this reason the high quality of the KW assignment is not guaranteed. For assignment of 500N-KeyPhrasesCrowdAnnotated-Corpus (500N-KPCrowd-v1.1) the researchers used the crowdsourcing platform Amazon's Mechanical Turk[2].

SemEval-2010 has been specially prepared for the Workshop on Semantic Evaluation 2010, where 19 systems were evaluated by matching their KW's against manually assigned ones. It consists of three parts: trial, training and test data. The authors note that on average 15% of the reader-assigned KW and 19% of the author-assigned KW's did not appear in the papers.

Table 3 shows the statistical characteristics of text length distributions in the reviewed corpora.

Table 3. **Statistical characteristics for the datasets used in this paper**

№	Name	Count	Mean	Min.	Max.	Std. Dev.
1	DUC-2001	307	769,1	141	2505	435,1
2	Hulth-2003	2000	125,9	15	510	59,9
3	NUS	211	6731,7	1379	13145	2370,6
4	NLM-500	500	4805	436	24316	2943,3
5	WIKI-20	20	5487,8	2768	15127	2773,4
6	FAO-30	30	19714,3	3326	70982	16101,6
7	FAO-780	779	30106,5	1224	255966	31076,5
8	KRAPIVIN	2304	7572,8	144	15197	2092,3
9	CiteULike	180	6454,1	878	23516	3408,9
10	SemEval-2010	244	7669,1	988	13573	2061,9
11	500N-KPCrowd-v1.1	447	425,9	38	1478	311,7

[1] http://www.fao.org/agrovoc
[2] https://www.mturk.com/

Figures 4–8 shows the text length distribution of the reviewed corpora.

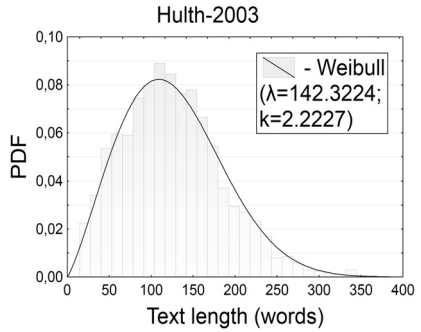


Figure 4. Distribution of annotation lengths in words

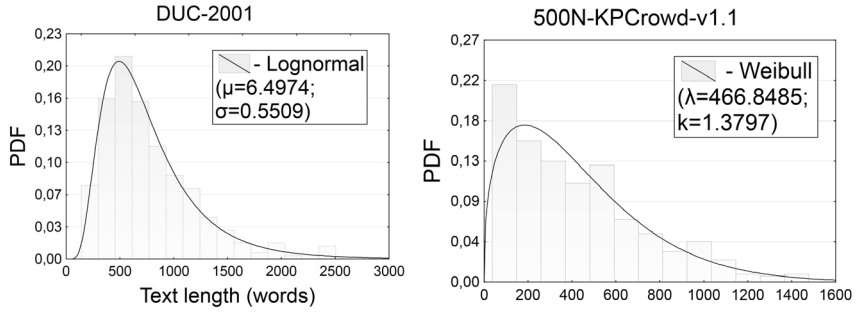


Figure 5. Distribution of news article lengths in words

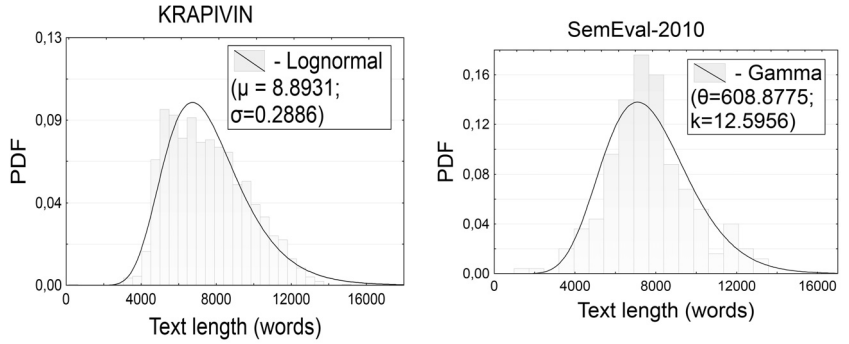


Figure 6. Distribution of ACM article lengths in words

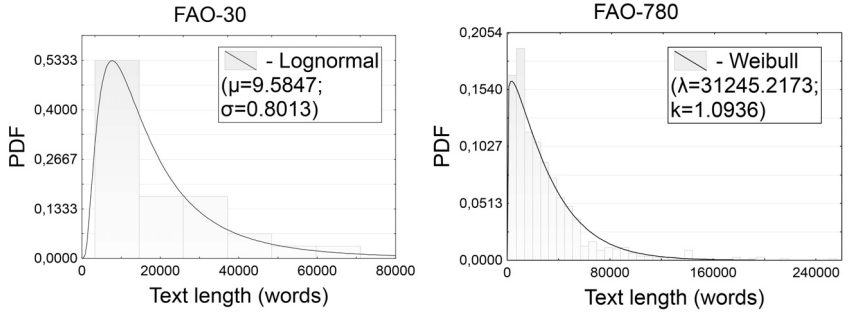


Figure 7. Distribution of FAO document lengths in words

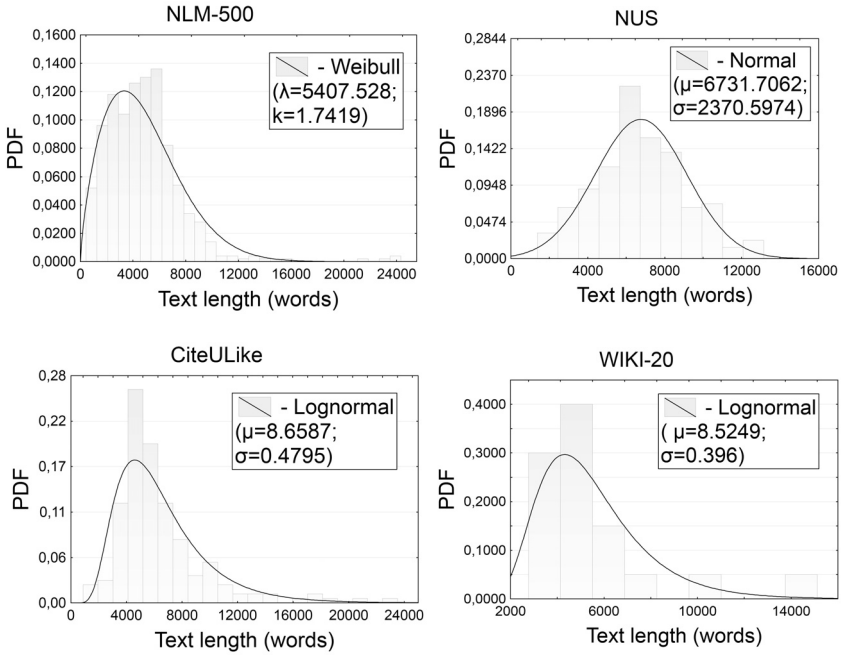


Figure 8. Distribution of Scientific paper lengths in words

A review of test corpora revealed that they differ significantly on the sizes, the themes, and the method of keyword assignment. The difference of text lengths for some couples is three orders of magnitude. The text

length in the tens of thousands of words questioned the possibility and the meaning of the use of AKWEA's at its entire length, without division into semantic parts. In contrast, annotation in definition contain a higher percentage of KW's than text containing a few thousand words.

The text length distribution histograms of the most reviewed corpora have outliers, and does not correspond to the established in Section 3.1 principles, that is their apparent drawback. *DUC-2001* has the most relevant form and distribution parameters (LN (6.49, 0.55)) but its disadvantage is the small number of experts participating in the KW assignment (only two). Moreover, all the above corpora are monolingual and do not allow carry cross-language study of KW extraction.

4. Conclusions

As can be seen from the above, the majority of the texts for which KW extraction is relevant are in the range of 400 to 2500 words and their text length distribution is quite well described by the lognormal form. Thus in practice it is advisable to use AKWEA's that show a good performance in certain text length ranges. However, in general a comparison of existing AKWEA's was performed on corpora with different characteristics. Moreover, the length of the manually assigned KW lists in them varies widely, and KW assignment was made by different categories of people such as students, volunteers and experts for example. Thus, for an objective comparison of existing AKEA, it is necessary to use corpora, whose characteristics are close to those of natural collections.

REFERENCES

1. A.S. Vanyush\kin and L.A. Graschenko, "Methods and algorithms of keyword extraction [Metody i algoritmy izvlecheniya klyushevyh slov]" New information technology in the automated systems [Novye informacionnye tekhnologii v avtomatizirovannyh sistemah], pp. 85–93, 2016.

2. E.V Yagunova, "Experiment and computation in the analysis of literary text's keywords [Eksperiment i vychisleniya v analize klyuchevyh slov hudozhestvennogo teksta]" Collection of scientific works of the department of foreign languages and philosophy of PSC of UB RAS. Philosophy of Language. Linguistics. Linguodidactics [Sbornik nauchnyh trudov kafedry inostrannyh yazykov i filosofii PNC UrO RAN. Filosofiya yazyka. Lingvistika. Lingvodidaktika], pp. 85-91, 2010.

3. T.G. Nozdrina, "Reconstructing original texts by key words [Osobennosti vosstanovlenija tekstov – originalov na osnove kljuchevyh slov]"

Modern problems of science and education [Sovremennye problemy nauki i obrazovanija], Vol. 1-2, pp. 167–174, 2015.

4. C.D. Manning, J.Bauer, J.Finkel, and S.J. Bethard, "The Stanford CoreNLP Natural Language Processing Toolkit" Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics, Baltimore, pp. 55–60. 2014.

5. B.Y. Lemeshko and S.N. Postovalov, "Limit distributions of the Pearson $\chi 2$ and likelihood ratio statistics and their dependence on the mode of data grouping" Industrial laboratory, Vol. 64, Iss. 5, pp. 344–351, 1998.

6. R.J. Hyndman, "The problem with Sturges' rule for constructing histograms" 1995. [Online]. Available at: http://robjhyndman.com/papers/sturges.pdf.

7. P. Sobkowicz, M. Thelwall, K. Buckley, G. Paltoglou, and A. Sobkowicz, "Lognormal distributions of user post lengths in Internet discussions - a consequence of the Weber-Fechner law?" EPJ Data Science, Vol. 2, pp. 1–20, 2013.

8. N. Kagan, "Why Content Goes Viral: What Analyzing 100 Million Articles Taught Us" 2013. [Online]. Available at: http://okdork.com/why-content-goes-viral-what-analyzing-100-millions-articles-taught-us.

9. N. Patel, "Why 3000+ Word Blog Posts Get More Traffic (A Data Driven Answer)." [Online]. Available at: http://neilpatel.com/blog/why-you-need-to-create-evergreen-long-form-content-and-how-to-produce-it.

10. T. Tunguz, "The Optimal Blog Post Length to Maximize Viewership" 2013. [Online]. Available: http://tomtunguz.com/content-marketing-optimizatio.

11. V. Paxson, "Empirically-Derived Analytic Models of Wide-Area TCP Connections" IEEE/ACM Transactions on Networking, Vol. 2, Iss. 4, pp. 316–336, 1994.

12. J.R. Douceur and W. J. Bolosky, "A Large-Scale Study of File-System Contents" Proceedings of the 1999 ACM SIGMETRICS international conference on Measurement and modeling of computer systems, Atlanta, pp. 59–70, 1999.

13. J.E. Blumenstock, "Size matters: word count as a measure of quality on wikipedia" Proceedings of the 17th International Conference on World Wide Web, Beijing, pp.1095–1096, 2008.

14. M.A. Serrano, A. Flammini, and F. Menczer, "Modeling statistical properties of written text" PLoS One, Vol. 4, Iss. 4, pp. 1–8, 2009.

15. A. Hulth, "Improved Automatic Keyword Extraction Given More Linguistic Knowledge" Proceedings of the 2003 conference on Empirical Methods in Natural Language Processing, Sapporo, pp. 216–223, 2003.

16. K.S. Hasan and V. Ng, "Conundrums in unsupervised keyphrase

extraction: making sense of the State-of-the-Art" Coling 2010 – 23rd International Conference on Computational Linguistics, Proceedings of the Conference, Beijing, pp. 365–373, 2010.

17. R. Mihalcea and P. Tarau, "TextRank: Bringing order into texts" Proceedings of EMNLP 2004,. Barcelona, pp. 404–411, 2004.

18. S.V Popova and I.A. Khodyrev, "Tag lines extraction and ranking in text annotation [Izvlechenie i ranzhirovanie kljuchevyh fraz v zadache annotirovanija]" Scientific and technical journal of information technologies, mechanics and optics [Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki], Vol. 1, pp. 81–85, 2013.

19. F. Rousseau and M. Vazirgiannis, "Main core retention on graph-of-words for single-document keyword extraction" Advances in Information Retrieval, Vienna, pp. 382–393, 2015.

20. N. Schluter, "Centrality Measures for Non-Contextual Graph-Based Unsupervised Single Document Keyword Extraction," In Proceedings of TALN 2014, Marseilles, pp. 455–460, 2014.

21. G. Tsatsaronis, I. Varlamis, and K. Norvag, "SemanticRank: Ranking Keywords and Sentences Using Semantic Graphs" Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, pp. 1074–1082, 2010.

22. T. Zesch and I. Gurevych, "Approximate Matching for Evaluating Keyphrase Extraction" Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing, Borovets, pp. 484–489, 2009.

23. X. Wan and J. Xiao, "Single Document Keyphrase Extraction Using Neighborhood Knowledge" Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, Chicago, pp. 855–860, 2008.

24. A.R. Aronson, J.G. Mork, W.G. Clifford, S.M. Humphrey, and W.J. Rogers, "The NLM Indexing Initiative's Medical Text Indexer" Studies in Health Technology and Informatics, Vol. 107, pp. 268–272, 2004.

25. C.W. Gay, M. Kayaalp, and A.R. Aronson, "Semi-automatic indexing of full text biomedical articles" AMIA Annual Symposium proceedings / AMIA Symposium. AMIA Symposium, pp. 271–275, 2005.

26. T. Nguyen and M. Kan, "Keyphrase extraction in scientific publications" Proceedings of the 10th international conference on Asian digital libraries: looking back 10 years and forging new frontiers, Hanoi, pp. 317–326, 2007.

27. O. Medelyan, I.H. Witten, and D. Milne, "Topic Indexing with Wikipedia" Proceedings of the Wikipedia and AI workshop at AAAI-08, Chicago, pp. 19–24, 2008.

28. O. Medelyan, I.H. Witten, and D. Milne, "Topic Indexing with Wikipedia (Thesis)" The University of Waikato, Hamilton, 2009.

29. O. Medelyan and I.H. Witten, “Domain Independent Automatic Keyphrase Indexing with Small Training Sets” Journal of the American Society for Information Science and Technology, Vol. 59, Iss 7, pp. 1026–1040, 2008.

30. M. Krapivin, A. Autayeu, and M. Marchese, “Large Dataset for Keyphrases Extraction” 2009. [Online]. Available at: http://eprints.biblio.unitn.it/archive/00001671/01/disi09055-krapivin-autayeu-marchese.pdf.

31. O. Medelyan, E. Frank, and I. H. Witten, “Human-competitive tagging using automatic keyphrase extraction” Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 1318–1327, 2009.

32. S. Kim, O. Medelyan, M. Kan, and T. Baldwin, “Semeval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles” Proceedings of the 5th International Workshop on Semantic Evaluation, Los Angeles, pp. 21–26, 2010.

33. L. Marujo, A. Gershman, J. Carbonell, and R. Frederking, “Supervised Topical Key Phrase Extraction of News Stories using Crowdsourcing, Light Filtering and Co-reference Normalization” In 8th International Conference on Language Resources and Evaluation (LREC 2012), Istanbul, pp. 399–403, 2012.

УДК 81'33

ON THE ROTATION OF RADICAL VOWELS IN SECONDARY IMPERATIVES

T. I. Galeev[1], Wang Yui[2]
[1]Justus Liebig University of Giessen, Giessen
[2]Kazan Federal University, Kazan
timur.galeev@slavistik.uni-giessen.de, wangyuk1917@gmail.com

The work describes the competition of grammatical synonyms – imperfective verbs with the component -iva/yva- with alternation of root vowels -o-/-a- (obsulavlivat'/obuslovlivat'). On the basis of the Google Books corpus, basic models of changes in the frequency dynamics of competing forms were described.

Keywords: variability, Google Books, language dynamics, verb paradigm.

К ВОПРОСУ О ЧЕРЕДОВАНИИ КОРНЕВОЙ ГЛАСНОЙ ВО ВТОРИЧНЫХ ИМПЕРФЕКТИВАХ

Т. И. Галеев[1], Ван Юй[2]
[1]Гиссенский университет им. Юстуса Либига, Гиссен
[2]Казанский федеральный университет, Казань
timur.galeev@slavistik.uni-giessen.de, wangyuk1917@gmail.com

В работе описывается конкуренция грамматических синонимов – глаголов несовершенного вида с компонентом -ива-/-ыва- с чередованием корневых гласных -о-/-а- (*обсулавливать*/*обусловливать*). На основе данных корпуса Google Books были описаны базовые модели изменения динамики частотности конкурирующих форм.

Ключевые слова: вариативность, Google Books, языковая динамика, глагольная парадигма.

Эволюционные изменения в грамматическом строе языка происходят в языке медленно, без резких видимых перестроек и, тем более, без разрушения его прежней внутренней структуры. Языковедов всегда привлекала данная проблематика, несмотря на сложность выявления таких изменений и необходимость обобщения огромной массы речевого материала в его противоречивой диахронической специфике.

Норма подвергается изменениям как под влиянием экстралингвистических факторов, так и в пределах самой системы при условии появления новых форм, постепенно вытесняющих старые. Как пока-

зывает опыт многолетних исследований, новые произносительные тенденции появляются поначалу в периферийной части языковой системы, в которой возникает асимметрия в результате проникновения вариативных элементов, ранее занимавших в языке второстепенное место. Поддерживаемая учеными гипотеза о волнообразном распространении инноваций в пространстве и социуме позволяет связать синхронный анализ с диахронным и установить социальную иерархию определенных языковых форм. Отношения между вариантами нормы сложны и многообразны.

Прежде всего вариант, рассматривающийся ранее как ошибка, получает статус нормативного, что существенно расширяет границы его функционирования. В течение определенного времени он употребляется наравне с прежней нормой. С другой стороны, вариант, считавшийся ранее нормативным, может впоследствии потерять этот статус. Есть большая вероятность, что новый вариант вытеснит старый и останется в системе языка как единственно правильный – сам факт его возникновения свидетельствует о стремлении системы к выравниванию и унифицированию форм в одном из ее звеньев. Придание же данному варианту статуса нормативного является подтверждением того, что его появление, обусловленное внутриязыковыми процессами нормирования, было не случайным, и необходимо на данном этапе развития социолингвистического континуума для удовлетворения изменившихся лингвистических потребностей говорящих.

Значение частотности для морфологической системы глаголов в последние годы изучается весьма активно. Так, при помощи статистических методов было доказано, что более частотные английские глаголы меньше склонны к регуляризации, чем менее частотные [1]. Аналогичное исследование, имеющее более традиционный лингвистический уклон, было выполнено и для немецкого языка [2]. Эти работы подтвердили фактическими данными интуитивно очевидное предположение о том, что более частотные слова хорошо сохраняют словоизменительный тип, а менее частотные слова склонны изменяться под воздействием аналогии.

Принимая во внимание тот факт, что русский язык – язык флективный, следует обратить особое внимание на различные морфологические процессы, сопровождающие словоизменительную парадигматику вариативных глагольных форм.

С целью выявления закономерностей эволюции вариативных форм центра глагольной парадигмы и описания эволюционной ди-

намики «конкурирующих» пар глагольных форм были получены данные о частотности 68 глаголов с вариативностью гласных *а/о* в корне (*обусловливать/обуславливать*) на материале корпусов текстов Google Books Ngram.

Для изучения эволюции вариативных форм предлагается применить квантитативный метод. На основе данных корпуса Google Books, осуществляющего поиск по книгам, изданным в основном с 1800 по 2000 гг., будут построены графики изменения частотности глаголов, имеющих избыточную парадигму. Характер корпуса определяет стилистический аспект исследования глаголов: сфера употребления глаголов – художественная, научная и научно-популярная литература.

Отдельно рассмотрим показатели частотности парадигм, которые отличаются от динамики частотности двух форм. Если ограничиваться только суммарными показателями, то можно потерять динамику и перспективу, которую в свою очередь могут отобразить именно графики.

Так, например, по таблице частотности, приведенной ниже, видно, что вариант с корневой гласной «о» доминирует в парадигмах *удваивать* (в среднем – 105 вхождений в год)/ *удвоивать* (в среднем – 413 вхождений в год), *удостаивать* (в среднем – 101 вхождение в год)/*удостоивать* (в среднем – 307 вхождений в год) и *дотрогиваться* (в среднем – 187 вхождений в год) / *дотрогиваться* (в среднем – 263 вхождения в год).

Таблица № 1. **Количественные показатели парадигм глаголов *дотрА\Огиваться* и *удвА\Оивать***

Корневая гласная	Глагольная парадигма	Среднегодовая частотность, GBN (% / 200 л.)	(шт. / год)
А	удостаивать ↑	0.0000230%	101
А	удваивать ↑	0.0000240%	105
А	дотрагиваться ↑	0.0000425%	187
О	дотрогиваться ↓	0.0000600%	263
О	удостоивать ↓	0.0000700%	307
О	удвоивать ↓	0.0000940%	413

Другой случай – примерно равное общее количество словоупотреблений за 200 лет, и, как следствие – одинаковая среднегодовая

частотность глагольных форм парадигм *присваивать* (в среднем – 505 вхождений в год) / *присвоивать* (в среднем – 461 вхождение в год) не совпадают с динамикой, отраженной в графиках, демонстрирующих смену нормы и доминирование форм с корневой «а».

Таблица № 2. **Количественные показатели парадигм глаголов *присвА\Оивать***

Корневая гласная	Глагольная парадигма	Среднегодовая частотность, GBN (% / 200 л.)	(шт. / год)
О	присвоивать ↓	0.0001050%	461
А	присваивать ↑	0.0001150%	505

Еще один случай, в котором количественные данные не отражают динамику – т.н. I этап смены нормы. Как видно из таблицы № 3, формы с корневой гласной «о» в разы превосходят свои варианты по частоте (соотношение количества форм глагола *сосредоточивать* к глаголу *сосредотачивать* – 8,25; *обусловливать* к *обуславливать* = 9,8).

Таблица № 3. **Количественные показатели парадигм глаголов *обуслА\Овливать* и *сосредА\Оточивать***

Корневая гласная	Глагольная парадигма	Среднегодовая частотность, GBN (% / 200 л.)	(шт. / год)
А	сосредотачивать ↑	0.0000230%	101
А	обуславливать ↑	0.0000880%	386
О	сосредоточивать ↑	0.0001900%	834
О	обусловливать ↓	0.0008600%	3776

Однако динамика изменения частотности демонстрирует стремительное падение частоты словоупотребления форм парадигмы с корневой гласной «о» и уверенный рост частоты словоупотреблений форм с корневой гласной «а».

Анализ результатов проведенного эксперимента показал, что «консервативность» глаголов обратно пропорциональна их частотности, а не наоборот. Наименее частотные глаголы независимо от ударения, производящей основы и словообразовательных элементов оказались более устойчивы к унификации под влиянием аналогии, а в наиболее частотных формах процесс смены нормы имел место и завершился.

Обладая информацией о схожих процессах в других глагольных парадигмах, можно предположить о возможной смене приведенных выше глагольных парадигм в ближайшем будущем и, как следствие, закрепление новой нормы в лексикографических, орфографических и орфоэпических источниках.

Если в английском и немецком языке переходу подвергаются не слишком частотные глаголы, и в первую очередь «мы имеем дело со стохастическим процессом, напоминающим радиоактивный распад» [3], при котором структура, состоящая из относительно устойчивого лексического ядра (распространенных глаголов) и более слабой маргины (редчайших глаголов) напоминает пирог, который начинает крошиться именно по краям.

Сложнее решить вопрос о структуре парадигмы и ее влиянии на сохранение/утрату нормы. Обычно процесс смены нормы внутри парадигмы происходит одинаково. Например, во всех формах вариант с «о» в корне вытесняется вариантом с «а», хотя и с различной динамикой и хронологией. Порядок этих изменений был обозначен в гипотезе Нессета–Янды [4] о прототипичности формы 3 л. ед. и мн. ч. (3Sg и 3Pl) и ее «консервативности».

Отдельно стоит упомянуть, что самой «консервативной» формой является форма страдательного причастия прошедшего времени – все обнаруженные формы в корне имели гласную «о» (*удостоенный*, но никогда – *удостаенный*). Однако, приводить данный факт в пользу прототипичности данной формы было бы рано, так как причины подобной аномалии (возможно, экстралингвистические) являются темой отдельного исследования.

Другое интересное наблюдение, периферийность форм 1 и 2 л. ед. ч.(1Sg и 2 Sg), возникло из-за стилистической несовместимости данных форм со сферой употребления этих слов (*я узакониваю, ты узакониваешь*).

Эти и другие результаты исследования глагольной вариативности доступны в электронном виде на облачном хранилище [5].

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект № 16-06-00165 А и № 17-29-09163 офи_м).

ЛИТЕРАУРА

1. Lieberman E., Michel J.–B., Jackson J., Tang T., Nowak M.A. Quantifying the evolutionary dynamics of language // Nature, 2007. Vol.

449. pp. 723–716. http://www.nature.com/nature/journal/v449/n7163/abs/nature06137.html [Электронный ресурс. Дата обращения: 01.10.2018].

2. Carroll R., Svare R., Salmons J. Quantifying the evolutionary dynamics of German verbs // Journal of Historical Linguistics 2, 2, 2012. – pp. 153–172.

3. Пиперски А.Ч. Чередование в корне как залог устойчивости: из истории сильных глаголов в немецком языке // Acta Lingüistica Petropolitana. Труды Института лингвистических исследований. Том X, часть 1. СПб.: Наука, 2014. – С. 821–831.

4. Nesset T., Janda L. Paradigm structure: Evidence from Russian suffix shift. // Cognitive Linguistics, 2010. Vol. 21(4), pp. 699–725.

5. Электронная БД проекта «Когнитивная модель словоизменительной глагольной парадигмы в русском языке: квантитативный анализ динамики частотности словоформ» НИЛ «Квантитативная лингвистика» https://kpfu.ru/bazy-dannyh-268352.html [Электронный ресурс. Дата обращения: 01.10.2018].

УДК 81-2

FEATURES AND PROBLEMS OF TRANSLATION OF THE MATHEMATICAL TERMS INTO TATAR LANGUAGE IN THE ESTABLISHMENT OF THE TAXONOMY

K. R. Galiaskarova, S. R. Mukhamedvalieva
Kazan Federal University, Kazan
Galias-alsu@yandex.ru, Sumbel@live.ru

The article explores the features of the translation of terminology into the Tatar language; translation of terms from Russian into Tatar. A number of problems have been posed related to bilingualism in the teaching of school material. Studied the structure and specificity of the Tatar language. A detailed analysis of the most significant problems encountered in the translation, with illustrative examples. The object of the study was a set of mathematical (planimetric) terms. On their basis, a taxonomy of the terms of mathematics was constructed in the section "Planimetry" in the Tatar language.

The article is descriptive and does not contain profound philological and linguistic knowledge. The practical significance of the study is that its results can help solve the problems of mathematical education in the Tatar language, preserve the language, destroy the language barrier, develop skills in the study of specialized literature in both languages. The use of the constructed ontology in the Russian and Tatar languages in the educational process of the university will improve the quality of training of the teacher of mathematics, which will contribute to improving the knowledge of mathematics of students in Tatar schools.

Keywords: planimetry, translation of the mathematical terms, Tatar language, taxonomy, ontology.

ОСОБЕННОСТИ И ПРОБЛЕМЫ ПЕРЕВОДА МАТЕМАТИЧЕСКИХ ТЕРМИНОВ НА ТАТАРСКИЙ ЯЗЫК ПРИ СОСТАВЛЕНИИ ТАКСОНОМИИ

К. Р. Галиаскарова, С. Р. Мухамедвалиева
Казанский федеральный университет, Казань
Galias-alsu@yandex.ru, Sumbel@live.ru

В настоящей статье изучены особенности перевода терминологии на татарский язык; выполнен перевод терминов с русского языка на татарский язык. Поставлен ряд проблем, связанных с двуязычием в преподавании школьного материала. Изучено строение и специфика татарского языка.

Представлен подробный разбор наиболее значимых проблем, возникших при переводе, с наглядными примерами. Объектом исследования являлось множество математических (планиметрических) терминов. На их основе была построена таксономия терминов математики по разделу «Планиметрия» на татарском языке.

Данная статья носит описательный характер и не содержит в себе глубоких филологических и лингвистических знаний. Практическая значимость проводимого исследования состоит в том, что его результаты могут способствовать решению проблем математического образования на татарском языке, сохранению языка, разрушению языкового барьера, развитию умений в изучении специализированной литературы на обоих языках. Использование построенной онтологии на русском и татарском языках в учебном процессе университета позволит повысить качество подготовки учителя математики, что будет способствовать совершенствованию знаний по математике учащихся татарских школ.

Ключевые слова: планиметрия, перевод математических терминов, татарский язык, таксономия, онтология.

Изучение специализированной литературы в любой предметной области требует знания определений терминов. В этом могут помочь различные интернет ресурсы, книжные профильные лингвистические словари. Актуальной является проблема перевода терминов определенной предметной области с русского на родной язык и обратно. Учащиеся, обучающиеся в татарских школах, где преподавание предметов первоначально ведется на татарском языке, затем для успешной сдачи единого государственного экзамена происходит переход в старших классах к предметной терминологии на русском языке испытывают некоторые проблемы с переводом и употреблением этих терминов на двух языках. Усугубляет проблему недостаточность двуязычных учебников, терминологических словарей. Рассмотрим подробнее проблемы, связанные с двуязычием в преподавании.

1. Проблема языкового барьера обусловлена тем, что многие ученики, разговаривающие с детства исключительно на татарском языке, испытывают затруднения при изучении школьного материала на русском языке.

2. Переход в старших классах, а затем в вузах к преподаванию лишь на русском языке обостряет проблему сохранения двуязычия и языковых культур в России.

3. Проблема ограниченности ресурсов по данной тематике. Недостаточность необходимой литературы на татарском языке, методик

по обучению на татарском языке, технических, профильных русско-татарских словарей и образовательных Интернет-ресурсов.

Проблема исследования связана с необходимостью предоставить возможность учащимся татарских школ самостоятельно освоить школьный материал через образовательный ресурс на татарском языке.

Для решения вышеназванных проблем в области изучения школьной математики учеными Казанского федерального университета, А.В. Кирилловичем, Е.К. Липачевым, О.А. Невзоровой, М.В. Фалилеевой, Л.Р. Шакривой, предложен подход, состоящий в создании цифровой образовательной платформы на основании онтологического подхода и семантических технологий. Для создания онтологии школьной математики выстроена таксономия математических терминов. В качестве пилотного выбран раздел школьной математики – «Планиметрия». Задача настоящего исследования – изучение особенностей и проблем перевода математических терминов на татарский язык и наполнение онтологии татарской терминологией.

Объектом данного исследования является множество математических (планиметрических) терминов. Данные термины прописаны в онтологии, отражающей формализацию предметной области «Планиметрия». В книге Б.В. Доброва [10, с. 9] под онтологией подразумевается «некоторое описание взгляда на мир применительно к конкретной области интересов. Это описание состоит из терминов и правил использования этих терминов, ограничивающих их значения в рамках конкретной области». В работе А.М. Елизарова [11, с. 224] говорится, что рассматриваемая онтология состоит из терминов (организованных в таксономию) и их связей. Составление и редактирование онтологии проводилось в специальном фреймворке – Protégé [24].

На рисунке 1 представлен фрагмент собранной таксономии на русском языке.

Необходимо перевести данные термины на татарский язык и согласовать их. Сформулируем в явном виде цели исследования:

1. Изучение соответствующей литературы об особенностях перевода математической терминологии с русского языка на татарский язык.

2. Выявление языковых особенностей на основе лингвистического анализа.

3. Перевод определенного множества терминов, соответствующих основным понятиям раздела математики – планиметрии.

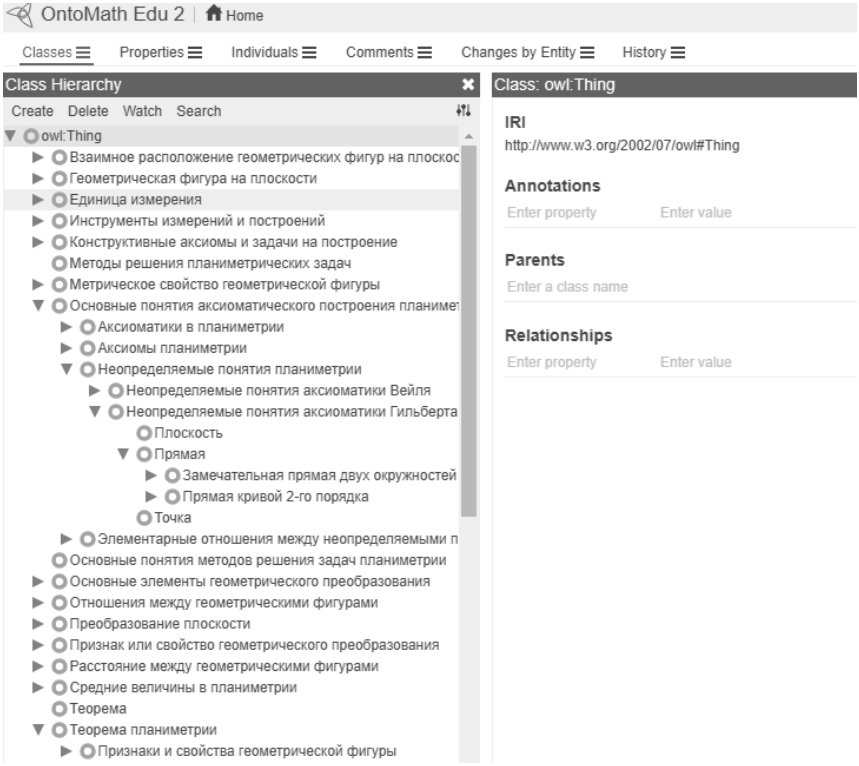


Рис. 1. Фрагмент онтологии на русском языке

4. Выделение особенностей и проблем перевода русских терминов на татарский язык.

5. Нахождение решения данных проблем.

В конечном итоге мы будем иметь фундаментальную, правильную таксономию терминов на русском и татарском языках из такого раздела математики, как планиметрия.

Татарский язык является вторым по численности носителей национальным языком России. Поэтому Русско-татарское двуязычие – это широко распространенное явление среди татар России и ряда стран СНГ [17]. Проблему обучения языкам и татарско-русское двуязычие рассматривает в своей работе М.М. Шакурова [23, с. 212].

При выявлении особенностей языка необходимо учитывать его специфику и строение в целом, знать, к какой группе относится тот или иной язык, чтобы при необходимости сопоставить или сравнить его. В книге Р.Г. Ахметьянова [3, с. 3] приводится анализ значений, этимологическая характеристика финско-тюркской группы, к которой относится татарский язык.

Исследование структуры татарского языка, его особенности, его значимость подчеркивает в своей работе Ф. С. Сафиуллина [21], говоря о том, что «На татарском языке говорят около 7 миллионов человек... Татарский язык, по сведениям ЮНЕСКО, стоит на четвертом месте в мире по своей стройности, формализованности и логичности. Знание татарского языка дает возможность общаться со всеми представителями тюркских народов. Татарский язык занимает четырнадцатое место в мире».

Для лучшего понимания способов формирования терминов на татарском языке, мы изучили работы Г.Ф. Мусиной [14, с. 41], [15, с. 108], посвященные формированию и развитию терминов в области физики, а также статью А.Ф. Галимянова [7, с. 63], в которой речь идет об образовании терминологии по информатике на татарском языке.

При группировании проблем перевода мы воспользовались критериями, приведенными в книге Л.С. Бархударова [5, с. 97], а также в статье Е.Н. Базалиной [4, с. 104].

Поскольку наша работа нацелена на расширение возможностей изучения геометрии учащихся, в процессе перевода за основу мы взяли учебник по геометрии для общеобразовательных учреждений [1], [2]. Также мы использовали русско-татарско-английский терминологический словарь по математике [20], математический русско-татарский словарь [6] и мобильное приложение «Русско-татарский и Татарско-русский словарь» [16].

Процесс перевода в книге Л.С. Бархударова [5, с. 11] определяется, как некое преобразование текста на одном языке в текст на другом языке, при этом сохраняя семантическое соответствие. Для того, чтобы представлять проблемы, возникающие при переводе терминов, рассмотрим их структуру. В работе Е.Н. Базалиной [4, с. 104] выделяются следующие виды терминов: термины-слова и термины-словосочетания. В свою очередь, согласно работе А.Г. Хайруллиной [22, с. 16], термины-слова, основываясь на морфемной структуре слова, можно разделить на: непроизводные (*нур* – «луч»), производные (*кисемтә* – «отрезок»), сложные (*күппочмак* – «многоугольник»);

термины-словосочетания на просты (*кысынкы почмак* – «острый угол») и сложные (*дүртпочмакның капма-каршы ятучы түбәләре* – «противоположные вершины четырехугольника»).

Изучив работы А.Ф. Галимянова [7], Ф.А. Ганиева [8], [9], Л.Х. Киштиковой [12], И.И. Сабитовой [18], Ф.К. Сагдеевой [19] и А.Г. Хайруллиной [22], мы выделили основные способы перевода терминов:

1. Переводческая транслитерация и транскрипция (заимствование): *квадрат* – «квадрат»;

2. Калькирование (буквальный перевод): *турыпочмаклык* – «прямоугольник»;

3. Описательный («разъяснительный») перевод: *тапкырлау* – «умножение»;

4. Приближенный перевод: *үзәкләрнең бер турыда урнашуы сызыгы* – «линия центров».

Далее определим три основные проблемы при переводе терминов на татарский язык:

1. Многозначность слов;

2. Синонимия терминов;

3. Принцип связи слов и формирования словоформ (словообразование).

В таблице 1 приведены термины на русском языке и существующий перевод данных слов на татарский язык.

Таблица 1. **Многообразие способов перевода с русского языка на татарский язык**

Термин на русском языке	Отрезок	Точка	Плоскость	Тело
Перевод слова на татарском языке	Кисәк Кыйпык Кыйпылчык Кисемтә Бер өлеш Бер ара кисәк	Бөрчек Төртке Нокта Урын	Яссылык Караш ноктасы Юнәлеш Өлкә	Җисем Тән Гәүдә Бәдән Көпшә

В таблице 2 показаны варианты перевода татарских слов на русский язык.

Для корректного выбора способа перевода необходимо обращаться непосредственно к контексту и понимать среду использования термина.

Таблица 2. **Многообразие способов перевода с татарского языка на русский язык**

Термин на татарском языке	Өчпочмак	Таралу	Кыр
Интерпри-тация слова на русском языке	Треугольник (геометрическая фигура) Треугольник (чертежный инструмент для черчения) Сажень Косынка Треугольник (традиционное национальное кушанье)	Распределение Расходиться Расселяться Разветвляться Распространяться Разваливаться Потеря душевного равновесия и самообладания Рассеяться	Грань Поле Пустошь Степь

Также рассмотрим следующую проблему – синонимии терминов. Приведем несколько примеров в таблице 3.

Таблица 3. **Синонимия терминов**

Вычитание	Линейность	Расстояние	Способ
Алу Чигерү Киметү	Сызыкча булу Сызыкчалык	Ераклык Ара	Ысул Алым Метод

Одним из основных способов словообразования в татарском языке является суффиксальный способ [8, с. 22]. Также татарский язык относят к агглютинативным языкам. Это означает, что многие слова в татарском языке формируются “приклеиванием” аффиксов. Каждый словообразовательный аффикс несет в себе только одно значение. Исходя из этого, при переводе математически терминов возникает немало проблем.

Перечислим наиболее употребляемые суффиксы, указанные в работе А.Г. Хайруллиной [22, с. 17]:

-лык/-лек: *яссылык* «плоскость», *охшашлык* «подобие», *өслек* «поверхность»;

- -ма/-мә: *орынма* «касательная», *билгеләмә* «определение»;
- -ыш/-еш: *үзгәреш* «изменение», *чишелеш* «решение»;
- -лау/-ләү: *суммалау* «складывать», *билгеләү* «определять»;
- -ча/-чә: *сызыкча* «линейно», *төшенчә* «понятие»;

• -ык/-ек, -к: *сызык* «линия», *төзек* «правильный»;
• -ым/-ем, -м: *җисем* «тело», *сызым* «чертеж»;
• -ымта/-емтә: *кисемтә* «отрезок», *төшенчә* «понятие»;
• -чы: *кисүче* «секущая», *бүлүче* «делитель».

Также в своей работе А.Г. Хайруллина [22, с. 18] выделила термины, образованные синтаксическим образом:

• Сложные слова: *ярымтуры* «полупрямая», *өчпочмак* «треугольник»;
• Парные слова: *нуль-вектор* «нуль-вектор», *радиус-вектор* «радиус-вектор»;
• Термины-словосочетания: «существительное (С) + существительное (С)», «прилагательное (П) + существительное (С)».

В работе [22, с. 19] показаны модели, с помощью которых образуются составные математические термины: С – С-ы: *почмак биссектрисасы*, С-ның – С-ы: *өчпочмакның почмагы*, С-лар – С-ы: *векторлар аермасы*.

Объектно-именные состававные математические термины: П – С: *сынык сызык* «ломаная линия», С-ле – С: *күчәрле* симметрия «осевая симметрия», С-дәш – С: *тиңдәш почмак* «соответсвенный угол».

При переводе терминов-словосочетаний, состоящих из трех и более компонент, возникли проблемы связи между компонентами. Согласно [22, с. 20], средства связи можно разделить на следующие группы:

1. Основное средство связи – аффикс принадлежности;
2. Связь компонентов осуществляется с помощью обязательного соседства словоизменительных и словообразовательных аффиксов;
3. Основное средство связи – аффиксы притяжательного падежа у зависимого компонента и принадлежности у главного.

В математических терминах-словосочетаниях так же имеются «фамильные термины»: *Пифагор теоремасы* «теорема Пифагора», *Эйлер формуласы* «формула Эйлера», *Чева теоремасы* «теорема Чевы».

Практические результаты первого этапа исследования состоят в следующем: выполнен перевод терминов с русского языка на татарский язык; все проблемы при переводе были сгруппированы по языковым признакам; была создана таксономия на татарском языке. Приведем наглядный пример построения таксономии терминов математики по разделу «Планиметрия» на татарском языке (рис. 2).

Все поставленные цели и задачи были решены.

В ходе исследования были получены следующие результаты:

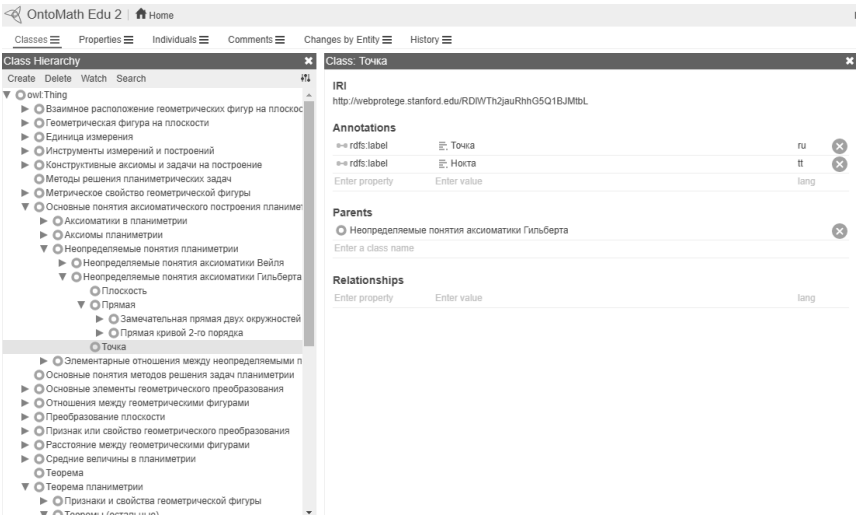


Рис. 2. Фрагмент онтологии с введенной татарской терминологией

1. Приведен подробный обзор литературы, связанной с тематикой данной работы. Стоит отметить недостаточность полезной литературы для изучения математики на татарском языке.

2. Выявлены языковые особенности татарского языка на основе лингвистического анализа.

3. Представлен подробный разбор наиболее значимых проблем, возникших при переводе геометрических терминов, с наглядными примерами.

4. Предложен перевод на татарский язык определенного множества терминов, соответствующих основным понятиям раздела математики – планиметрии.

5. Составлена таксономия на татарском языке.

6. Русскоязычная онтология дополнена терминологией на татарском языке.

Практическая значимость проводимого исследования состоит в том, что его результаты могут способствовать решению проблем математического образования на татарском языке, сохранению языка, разрушению языкового барьера, развитию умений в изучении специализированной литературы на обоих языках. Использование построенной онтологии на русском и татарском языках в учебном процессе университета позволит повысить качество подготовки

учителя математики, что будет способствовать совершенствованию знаний по математике учащихся татарских школ. Результаты исследования могут найти применение в учебном процессе, а также при подготовке школьных учебников, учебных пособий по математике.

Дальнейшее исследование будет посвящено проблемам определения терминов на татарском языке.

Благодарности. Работа выполнена при финансовой поддержке РФФИ и Правительства Республики Татарстан в рамках научного проекта № 18-47-160007.

ЛИТЕРАТУРА

1. Атанасян Л.С. Геометрия. 7–9 классы: учеб. для общеобразоват. учреждений / Л.С. Атанасян, В.Ф. Бутузов, С.Б. Кадомцев и др.; пер. с рус. З.Х. Билалова, В.З. Закиров. – Казань: Татар. кн. изд-во, 2011. – 384 с.

2. Атанасян Л.С. Геометрия. Учеб. для 10–11 кл. общеобразоват. учреждений / Л.С. Атанасян, В.Ф. Бутузов, С.Б. Кадомцев и др.; пер. с рус. З.Х. Билалова. – Казань: Магариф, 2005. – 213 с.

3. Ахметьянов Р. Г. Общая лексика материальной культуры народов Среднего Поволжья. – М.: Наука, 1988. – 220 с.

4. Базалина Е.Н. К проблеме перевода терминов научно-технических тезисов / Е.Н. Базалина. // Вестник Майкопского государственного технологического университета. – 2009. – №1. – С. 102–107.

5. Бархударов Л.С. Язык и перевод (вопросы общей и частной теории перевода) / Л.С. Бархударов. – М.: изд-во «Международные отношения». – 1975. – 240 с.

6. Галиева Л.И. Математический русско-татарский толковый словарь = Математикадан русча-татарча аңлатмалы сүзлек / Л.И. Галиева, И.Г. Галяутдинова, М.З. Хуснутдинова и др.; под общ. ред. Л.И. Галиевой, И.Г. Галяутдинова. – Казань: Татар. кн. изд-во, 2013. – 375 с.

7. Галимянов А.Ф. Система татарских терминов в компьютерных технологиях и информатике / А.Ф. Галимянов, Д.Ш. Сулейманов. // Лингвистические ресурсы и электронные корпусы. – 2013. – С. 61–69.

8. Ганиев Ф.А. Суффиксальное словообразование в современном татарском литературном языке / Ф.А. Ганиев. – Казань: Таткнигиздат, 1974. – 232 с.

9. Ганиев Ф.А. Функциональное словообразование в современном татарском литературном языке / Ф.А. Ганиев. – Казань, 2009. – 237 с.

10. Добров Б.В. Онтологии и тезаурусы: учебное пособие / Б.В. Добров, В.Д. Соловьев, В.В. Иванов, Н.В. Лукашевич. – Казань, Москва, 2006. – 156 с.

11. Елизаров А.М. Семантические технологии в математическом

образовании: онтологии и открытые связанные данные / А.М. Елизаров, А.В. Кириллович, Е.К. Липачев, О.А. Невзорова, Л.Р. Шакирова // Ученые записки. – ИСГЗ, 2018. – №1. – С. 222–227.

12. Киштикова Л.Х. Словообразовательный потенциал наречия в тюркских языках: автореф. дис. на соиск. учен. степ. канд. фил. наук. Акад. наук РТ. – Нальчик, 2004. – 22 с.

13. Математический русско-татарский толковый словарь. [Электронный ресурс] https://www.livelib.ru/book/1001414982/about-matematicheskij-russkotatarskij-tolkovyj-slovar. – (Дата обращения 11.10.2018).

14. Мусина Г. Ф. Развитие терминологии физики в татарском языке в послеоктябрьский период / Г.Ф. Мусина. // Филологические науки в России и за рубежом: материалы V Междунар. науч. конф. – СПб.: Свое издательство. – 2017. – С. 40–43.

15. Мусина Г.Ф. Термины физики татарского языка в историческом, генетическом и структурном аспекте / Г.Ф. Мусина. – Москва: Русайнс. – 2016. – С. 129.

16. Русско-татарский и Татарско-русский словарь version 2.0.2 мобильное приложение

17. Русско-татарское двуязычие. [Электронный ресурс]: Википедия. – Режим доступа: https://ru.wikipedia.org/wiki/Русско-татарское_двуязычие. – (Дата обращения: 11.10.2018).

18. Сабитова И.И. Словообразовательная характеристика лексики татарского языка: дис. на соиск. учен. степ.канд. фил. наук. Акад. наук РТ. – Казань, 1999.

19. Сагдеева Ф.К. Проблема культуры татарской речи в условиях активного двуязычия: автореф. дис. на соиск. учен. степ. канд. фил. наук. Акад. наук РТ. – Казань, 2001. – 28 с.

20. Салехова Л.Л. Математика: русско-татарско-английский терминологический словарь / Л.Л. Салехова, Н.К. Туктамышов. – Казань: Изд-во Казан. ун-та, 2014. – 120 с.

21. Сафиуллина Ф.С. Татарский язык на каждый день. [Электронный ресурс]: Татарская электронная библиотека / Ф.С. Сафиуллина. – Казань: Хәтер. – 2000. – Режим доступа: http://kitap.net.ru/safiullina. – (Дата обращения: 11.10.2018).

22. Хайруллина А.Г. Формирование и развитие математических терминов в татарском языке: автореф. дис. на соиск. учен. степ. канд. фил. наук. Акад. наук РТ. – Казань, 1996. – 22 с.

23. Шакурова М.М. Татарско-русское двуязычие: проблемы обучения языкам / М.М. Шакурова // Филологические науки. Вопросы теории и практики. – 2016. – №10-2 (64). – С. 211–214.

24. Protégé. [Электронный ресурс]. – Режим доступа: https://protege.stanford.edu/. – (Дата обращения: 11.10.2018).

УДК 811.512.145; 81'37; 81'367.625

DATABASE OF SEMANTIC CLASSES OF TATAR VERBS: METHODOLOGY AND IMPLEMENTATION ASPECTS

Alfiya Galieva[1], Madekhur Ayupov[1,2]
[1]Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic, Kazan
[2]Kazan Federal University, Kazan
amgalieva@gmail.com, madehur@mail.ru

The paper represents the conception and the main aspects of implementation of the Database of semantic classes of Tatar verbs; this new resource is developed in the Institute of Applied Semiotics of the Tatarstan Academy of Sciences. The authors outline the general principles of representing verbs in the database, main features of verb annotation, basic properties of semantic relations used and ensembles gotten. The scheme of verb classification is based on the parametric principle and includes a set of morphological, syntactic, semantic, and derivational characteristics which are relevant for Tatar grammatical and semantic systems; so that enables to distinguish homogeneous semantic subclasses within broad and heterogeneous thematic classes.

As illustration, examples of distinguished subclasses are given, description of their content, properties and quantitative characteristics are described.

Keywords: the Tatar language, verb semantics, lexicographic database, grammar and semantics.

БАЗА ДАННЫХ СЕМАНТИЧЕСКИХ КЛАССОВ ТАТАРСКИХ ГЛАГОЛОВ: МЕТОДОЛОГИЯ И АСПЕКТЫ РЕАЛИЗАЦИИ

А. М. Галиева[1], М. М. Аюпов[1,2]
[1]Институт прикладной семиотики Академии наук Республики Татарстан, Казань
[2]Казанский федеральный университет, Казань
amgalieva@gmail.com, madehur@mail.ru

Статья представляет концепцию и особенности разработки лексикографической базы данных семантических классов татарских глаголов, данный ресурс разрабатывается в Институте прикладной семиотики АН РТ. Описаны общие принципы фиксации глаголов в данном ресурсе, характер выделяемых семантических отношений и группировок, специфика использованной разметки. Разработанная авторами классификационная схема для обработки глагольной лексики основана на параметрическом принципе

и выполняется с учетом набора семантических, морфологических, деривационных и синтаксических признаков, релевантных грамматической и лексико-семантической системе татарского языка, что позволяет выделять более однородные семантические группировки внутри больших тематических классов. В качестве иллюстрации приводятся примеры выделяемых подклассов с описанием их свойств, а также количественные характеристики отдельных глагольных классов, представленных в базе.

Ключевые слова: татарский язык, семантика глагола, лексикографическая база данных, грамматика и семантика.

1. Введение

Разработка машиночитаемых словарных массивов для языков Российской Федерации является актуальной задачей, решение которой во многом определяет успешность реализации проектов по автоматическому анализу текстов, машинному переводу, распознаванию речи и т. п. Компьютерные словари глаголов являются ключевыми в процессе обработки естественного языка, нацеленного на интерпретацию данных.

В Институте прикладной семиотики Академии наук Республики Татарстан разрабатывается ряд проектов по разработке лексикографических баз татарского языка, в частности, русско-татарская лексикографическая база данных [1], а также база данных семантических классов татарских глаголов [2]. Данные проекты нацелены на представление актуального состояния татарского словарного фонда и выполняются с опорой на корпусные данные.

В статье представлен опыт фиксации в лексикографической базе данных семантических классов татарских глаголов, описана общая методология обработки лексического материала, а также архитектура и техническая реализация базы данных.

2. Вдохновившие нас проекты

Для описания семантики глагольных лексем лингвисты предлагают различные форматы представления значения и различные классификационные схемы. Перечислим лишь проекты, существенно повлиявшие на ход нашего исследования и проектирование описываемого в статье ресурса.

1. Семантическая классификация Русского национального корпуса [3] включает указание на словообразовательные показатели лексемы, лексико-грамматические классы и собственно семанти-

ческие признаки (тематические, или, как указывают разработчики, таксономические, классы), например, для глагола могут быть использованы пометы, указывающие на движение, физическое воздействие, обладание, эмоцию и т. п. [4].

2. Классификация английских глаголов (около 3100 единиц), B.Levin основана на сходстве их значений и синтаксических свойств; исследователь выделяет семантические классы, принимая в расчет широкий спектр их возможных синтаксических преобразований (например, мену диатезы), которые отражают значение глагола [5]. Работа B.Levin, основанная на идеях генеративной лингвистики, показывает, что семантические признаки глагола имеют явно выраженную корреляцию с его синтаксическими свойствами и интерпретацией его аргументов. Семантические классы, выделенные B. Levin, стали основой для многих других классификаций и лексикографических баз данных, в частности, проекта VerbNet [6].

3. VerbNet представляет собой иерархический тезаурус, не привязанный к конкретной предметной области, в котором каждый класс имеет синтаксическое описание, отражающее возможные поверхностные реализации структуры аргумента в типах конструкций, содержащих переходные и непереходные глаголы, сочетания глагола с предлогами, а также большой набор залоговых преобразований. Для каждого класса описаны тематические роли, представлены ограничения на отбор аргументов, а также синтаксические фреймы [7].

4. Разработчики электронного словаря русских глагольных конструкций FrameBank [8]. позиционируют свой проект как «создание русского фреймнет-ориентированного ресурса, спроектированного с учетом традиций отечественной лексической семантики и специфики русского языка, где информация о предложно-падежной реализации управления предикатов и поверхностно-синтаксических свойствах других конструкций имеет особую ценность». Ядро системы FrameBank составляют 2200 ключевых русских глаголов и ассоциированных с ними конструкций и корпусных примеров, FrameBank строится вокруг конструкций конкретных лексем и документирует:

– русскую лексическую систему, структуру русских лексико-семантических групп и полисемии;

– парадигматические отношения между значениями много-

значных слов (как они отражаются в системе связанных с этими значениями лексических конструкций);

– лексико-семантические ограничения на слоты конструкций;

– грамматические особенности русского языка (порядок слов, особенности использования падежей, согласования и т. п.) [9].

В нашем проекте, подобно семантической аннотации глаголов в Национальном корпусе русского языка, глаголы снабжены семантическими пометами (базовый список тэгов позаимствован именно из русского корпуса); внутри больших тематических классов выделены компактные, относительно однородные группировки, подобно классификации B.Levin и проекту VerbNet; базовые глаголы снабжены таблицами с фиксацией их синтаксической валентности и указанием тематических ролей актантов, подобно проектам VerbNet и FrameBank.

3. Общие принципы представления семантики татарских глаголов в БД

Разработанная классификационная сетка основана на параметрическом принципе и выполняется с учетом набора семантических, морфологических, деривационных и синтаксических признаков; предполагается, что данные признаки обусловлены спецификой грамматической и лексико-семантической системой татарского языка. Использованный подход является комплексным и позволяет выделять более частные семантические группировки внутри тематических классов.

В основу классификации положены следующие признаки:

– тематический (онтологический) класс, обусловленный сферой денотации глагола;

– словообразовательные модели глагольных лексем внутри тематического класса/подкласса;

– потенциал актантной деривации глаголов внутри подкласса;

– актантная структура глаголов, особенности глагольного управления и тематические роли актантов [2].

Такой интегрированный грамматико-синтаксический подход к выделению семантических классов соответствует основным тенденциям в мировой науке при описании семантики глагола и является попыткой создать своеобразный аналог проектов VerbNet и FrameBank на материале татарского языка.

При описании значения глаголов нами выделяются семантические тэги двух типов:

– тематические тэги, определяющие лексико-семантический класс глагола;

– строевые компоненты значения, не привязанные к конкретному тематическому кассу, но фиксирующие регулярные семантические признаки глаголов [10].

Нами выделяются следующие основные строевые (категориальные) пометы:

c:caus – указание на каузативы;

c:coop – указание на кооперативы (ассоциативы), выражающие мультисубъектность (совместность/взаимность) глагольного действия;

c:refl – указание на возвратность глагольного действия.

Строевые компоненты значения могут иметь или не иметь специальный морфологический маркер, например, глагол *кайтару* (*кайтар-у*) – «вернуть» является морфологическим каузативом, а глагол *ташу* – «таскать, перевозить» – лексическим (каузативное значение представлено импилицитно в самой глагольной основе). Строевые компоненты значения отчасти коррелируют с залоговыми показателями глагола, но не могут быть сведены только к ним.

При выделении тематических классов нами использован как древесный, так и фасетный принцип классификации, что позволяет при необходимости снабжать глагол одновременно несколькими тэгами.

В табл. 1 представлены примеры семантического аннотирования глаголов.

Таблица 1. **Примеры присвоения тематических и категориальных помет**

Глагол	Тематические пометы	Категориальные пометы
Хәбәрләш – «извещать друг друга»	t:speech	c:coop
серләш – «делиться секретами»	t:speech	c:coop
авырт – «болеть (ощущать боль)»	t:physiol, t:perc	-
сызла – «болеть, ныть»	t:physiol, t:perc	-
дәвала – «лечить»	t:physiol, t:impact	c:caus
илт – «относить»	t:move	c:caus

В словнике БД глаголы представлены в форме императива 2-го лица единственного числа – в грамматической форме, не осложненной аффиксами имени действия или инфинитива и представляющей собой основу глагола, которая служит базой для образования финитных и нефинитных форм. Такой способ представления является

оптимальным с учетом предполагаемого использования БД для компьютерных приложений.

Семантическая аннотация татарских глаголов (около 3200 синтетических глаголов) позволила выделить базовые тематические классы слов, которые стали объектом дальнейшей параметрической классификации для выделения внутри больших конгломератов слов относительно однородных группировок слов в учетом словообразовательной модели, актантной деривации и синтаксического поведения.

В результате проведенной классификации, как правило, в один подкласс попадают глаголы сходной семантики – синонимы, антонимы, а также часто слова, относящиеся к одному и тому же гипероним (такой гипероним может выделяться лишь в качестве концептуальной структуры, не будучи вербализованным в языке). Так, в табл. 2 представлено лексическое наполнение одного из подклассов глаголов физиологической деятельности и состояния – глаголы, обозначающие ощущения в какой-либо части тела.

Таблица 2. **Глаголы, обозначающие ощущения в какой-либо части тела**

Глаголы	Русский перевод (основное значение)	Тематические тэги
авырт	«болеть» (о боли)	t:physiol,, t:perc
сызла	«сильно болеть, ныть»	t:physiol,, t:perc
әрне	«болеть щемящей болью, ныть»	t:physiol,, t:perc
ачыт	«болеть саднящей, обжигающей болью»	t:physiol,, t:perc
кычыт	«чесаться, зудеть»	t:physiol,, t:perc
кызыш	«испытывать чувство лихорадки»	t:physiol,, t:perc
кымырҗыы	«свербеть, зудеть, першить»	t:physiol,, t:perc
чымырда	«ощущать мурашки на теле»	t:physiol,, t:perc
чемердә	«ощущать мурашки на теле»	t:physiol,, t:perc

Кроме общего семантического компонента ‘испытывать определенные ощущения в какой-либо части тела или органе’, все глаголы данного подкласса характеризуются еще рядом общих свойств:

– все глаголы являются непереходными и выражают состояние;

– все глаголы имеют каузативные дериваты типа *авырттыр* – «каузировать боль», *сызлат* – «каузировать сильную боль», *әрнет* – «каузировать щемящую боль», *кычыттыр* – «каузировать зуд» и не имеют форм взаимно-совместного и пассивного залогов;

– стандартным синтаксическим субъектом при данных глаголах являются слова, называющие части тела: *Баш авырта.* – «Голова

болит»; *Теш сызлый.* – «Зуб болит»; *Кулым кычыта.* – «У меня чешется рука».

Лексические синонимы, характеризуемые неодинаковой синтаксической валентностью, включаются нами в разные подклассы.

Таблица 3 представляет собой фрагмент БД, отражающий глагольное управление и типичное лексическое окружение для глаголов, обозначающих ощущения в какой-либо части тела, представленных в таб. 2.

Таблица 3. Фрагмент таблицы БД с валентностями и тематическими ролями актантов

	Окружение лексемы, форма	**Семантическая роль актанта**	**Возможное ограничение**
Пример	*Баш авырта шул*		
	S(NOM)	Субъект физиологического состояния	Часть человека
Пример	*Габделхәй Сабитовның йөрәге авырта иде*		
	S(NOM)	Субъект физиологического состояния	Часть человека
	S(GEN)	Субъект физиологического состояния	Живое существо
Пример	*Ул эчүләренә башы авырта торгандыр*		
	S(NOM)	Субъект физиологического состояния	Часть человека
	N(DIR)	Периферия, причина	
Пример	*Уң кулы күптәннән авырта иде инде, хәзер әйбер дә тота алмый башлаган*		
	S(NOM)	Субъект физиологического состояния	Часть человека
	ADV	Периферия, время	
Пример	Кече яшьтән үк башым авыртты, укырга сәләтем булмады		
	S(NOM)	Субъект физиологического состояния	Часть человека
	N(ABL)	Периферия, время	

В БД фиксируются как типичные актанты, так и сирконстанты, зависимые от глагольного предиката и заполняющее его активную синтаксическую валентность, а также даются примеры типичных

зависимых клауз. Таким образом, разрабатываемая БД позволяет получить данные о семантическом классе глаголов, о типах глагольного управления, семантических ролях актантов, посмотреть иллюстративный материал.

Выделение подклассов основано на совокупности семантических, синтаксических, грамматических признаков, среди которых особенности глагольного управления имеют особое значение. Лексические синонимы, характеризуемые неодинаковой синтаксической валентностью, включаются нами в разные подклассы.

4. Архитектура и техническая реализация БД

В ходе реализации проекта была разработана и реализована база данных семантических классов татарских глаголов. Логическая схема базы данных представлена на рис. 1.

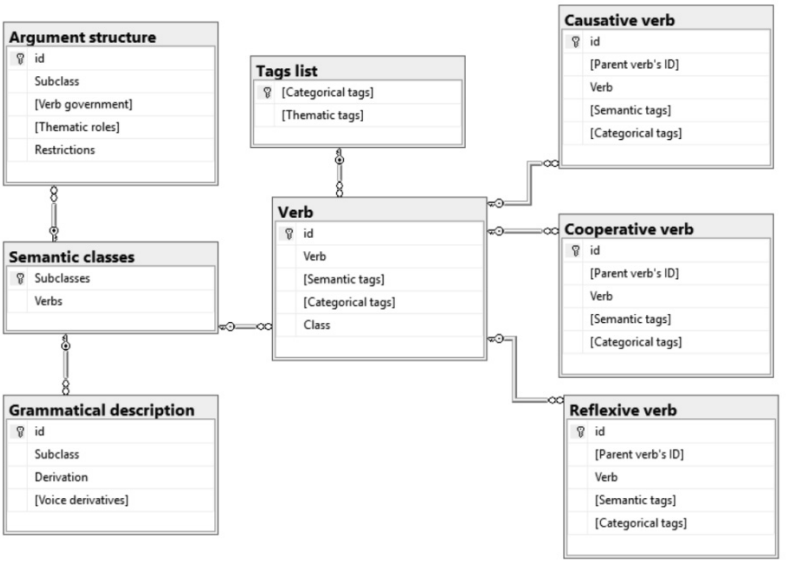


Рис. 1. Логическая структура БД

Для реализации системы использовалась система управления БД MS SQL Server. Размер БД отражает табл. 4, где представлены количественные характеристики выделяемых классов на примере ментальных глаголов (глаголов интеллектуальной деятельности), глаголов эмоций и глаголов физиологического действия и состояния.

Таблица 4. **Количественные характеристики отдельных классов в БД**

Класс	Кол-во подклассов	Кол-во базовых глаголов	Кол-во дериватов от базовых глаголов			Общее кол-во глаголов
			каузативы	кооперативы	рефлексивы	
Глаголы интеллектуальной деятельности	19	87	66	13	3	169
Глаголы эмоций	31	162	126	31	18	337
Глаголы физиологической сферы	30	203	136	4	26	369

Таблица 4 показывает, что выделенные и отраженные в БД семантические классы татарских глаголов имеют разный объем (включая разное количество залоговых дериватов) и неодинаковое количество частных семантических группировок внутри классов.

5. Заключение

Классификационная сетка, используемая при разработке и заполнении БД, основана на параметрическом принципе и выполняется с учетом набора семантических, морфологических, деривационных и синтаксических признаков, релевантных грамматической и лексико-семантической системе татарского языка. Регулярная морфологическая структура производных глаголов во многом определяет базовое значение конкретных дериватов, относимых к одному семантическому классу (подклассу).

При семантической аннотации для глагола разграничены категориальные семы (строевые компоненты значения) и таксономия (тематический рубрикатор). К категориальным нами отнесены те компоненты значения, которые могут быть выделены в словах разных тематических классов и лексико-семантических групп. Использование пучка независимых тэгов позволяет получить сложную классификацию по совокупности признаков.

Использованный при разработке БД комплексный подход позволяет выделять более частные и однородные семантические группировки внутри гетерогенных тематических классов.

Одна из перспектив расширения лингвистической информации в БД – добавление информации об образовании грамматикализованных конвербных конструкций, что имеет свою специфику по семантическим классам и подклассам.

ЛИТЕРАТУРА

1. Аюпов М.М. Русско-татарская лексикографическая база данных: реализация задачи пополнения данных // Материалы Юбилейной X Международной научно-практической конференции «Электронная Казань 2018». – Казань, 2018. – С. 70–73.

2. Galieva A., Vavilova Z., Gatiatullin A. Semantic Classification of Tatar Verbs: Selecting Relevant Parameters // Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts / Jaka Čibej, Vojko Gorjanc, Iztok Kosem and Simon Krek (Eds.) – Ljubljana: Ljubljana University Press, Faculty of Arts, 2018. – Pp. 811–818.

3. Русский национальный корпус. Электронный адрес: http://www.ruscorpora.ru/

4. Кустова Г. И., Ляшевская О. Н., Падучева Е. В., Рахилина Е. В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003–2005. Результаты и перспективы. – М., 2005. – С. 155–174.

5. Levin B. English Verb Classes and Alternations: A Preliminary Investigation. Chicago, University of Chicago, 1993. – 348 p.

6. VerbNet. Электронный адрес: http://verbs.colorado.edu/verbnet/

7. Kipper K., Korhonen A., Ryant N., Palmer M. Extending VerbNet .with Novel Verb Classes // Proceedings of the Fifth International Conference on Language Resources and Evaluation – LREC'06. May, 2006, – Genoa, 2006. – URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.380.5541&rep=rep1&type=pdf.

8. Электронный словарь русских глагольных конструкций FrameBank. Электронный адрес: http://framebank.ru

9. Кашкин Е.В, Ляшевская О.Н. Семантические роли и сеть конструкций в системе FrameBank // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной международной конференции «Диалог 2013». – Т. 1. – С. 297–311.

10. Galieva A., Nevzorova O. Semantic Annotation of Verbs for the Tatar Corpus // Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity. 6–10 September, 2016. Tbilisi, Ivane Javakhishvili Tbilisi State University, 2016. – Pp. 340–347.

УДК 811.512.145, 81'373.612

NEOLOGISMS WITH *ARA* AFFIXOID IN TATAR

Alfiya Galieva[1], Madekhur Ayupov[1,2]
[1]Institute of Applied Semiotics of the Academy of Sciences of Tatarstan Republic, Kazan
[2]Kazan Federal University, Kazan
amgalieva@gmail.com, madehur@mail.ru

Tracing new vocabulary, analyzing frequency of new lexical items, selecting appropriate ones to fix them subsequently in dictionaries is a topical task which demands modern technologies; nowadays search for new lexical material is executed on data of linguistic corpora. The paper studies neologisms in Tatar derived from nouns by means of the *ara* affixoid.

The linguistic data was retrieved from the following text collections:

– «TuganTel» Tatar National Corpus,

– the set of Tatar texts extracted from the Internet.

We performed semantic and quantitative analysis of selected new lexical items and distinguished some semantic groups depending on the meaning of the motivating nouns. We displayed how the *ara* affixiod is involved in current derivation processes in Tatar forming quite independent and productive word-formative type, which is constantly replenished with new items. We conclude that the bulk of such items appeared by component-by-component translating of corresponding Russian words with *mezh-* and *mezhdu-* prefixes, and they are mainly used in texts of socio-political and scientific domains. However calquing can not explain all cases of using items containing the ara affixoid: we found words and contexts of their using in which the analysed items do not have direct equivalents in the Russian language. We found also the derivates formed from extended noun stems containing plural affix. The linguistic data allows us conclude that calqing and original Tatar derivation mechanisms are closely interwoven and affect each other.

Keywords: word formation, the Tatar language, affixoid, calquing, new vocabulary.

НОВООБРАЗОВАНИЯ С АФФИКСОИДОМ *АРА* В ТАТАРСКОМ ЯЗЫКЕ

А. М. Галиева[1], М. М. Аюпов[1,2]
[1]Институт прикладной семиотики Академии наук Республики Татарстан, Казань
[2]Казанский федеральный университет, Казань
amgalieva@gmail.com, madehur@mail.ru

Отслеживание новой лексики, анализ частотности новообразований и отбор лексем для последующей их фиксации в словарях – одна из важных задач, для решения которых успешно применяются современные технологии, в частности, поиск материала осуществляется на данных лингвистических корпусов. В статье анализируются новообразования в татарском языке, образованные при помощи аффиксоида *ара.* Источником языкового материала послужили две коллекции текстов:

– Татарский национальный корпус «Туган тел»,

– коллекция текстов на татарском языке, извлеченных из сети Интернет.

Отобранные для анализа новообразования были подвергнуты семантическому и квантитативному анализу с выделением основных группировок по значению мотивирующего существительного. Собранный материал и его анализ позволяет сделать вывод о том, что аффиксоид *ара* участвует в актуальных процессах татарского словообразования и формирует вполне самостоятельный и весьма продуктивный словообразовательный тип, который постоянно пополняется новыми единицами. Основная часть таких единиц возникла в результате калькирования соответствующих русских слов с префиксами *меж-* и *между-*. Такие кальки используются преимущественно в книжных текстах общественно-политической и научной тематики. Однако калькирование не является единственным источником образования новых единиц с компонентом *ара*; выявлены слова и контексты, в которых искомые единицы не имеют прямых эквивалентов в русском языке. В статье также описываются слова, образованные от основы, осложненной аффиксом множественного числа. Анализ образований с аффиксоидом ара позволяет заключить, что в современном татарском словообразовании механизм калькирования и ресурсы собственно татарского словообразования и грамматики работает в органическом единстве.

Ключевые слова: словообразование, татарский язык, аффиксоид, калькирование, новая лексика.

1. Введение

В связи с необходимостью разработки специализированных лексикографических ресурсов для татарского языка в Институте прикладной семиотики АН РТ была спроектирована и создана

русско-татарская лексикографическая база данных, которая стала эффективным инструментом для научных исследований и различных лингвистических приложений [1]. База данных состоит из взаимосвязанных татарской и русской компонент, которые имеют независимую структуру, обусловленную языковой спецификой, и объединяются на уровне лексических эквивалентов. Каждая языковая компонента представляется грамматической и семантической моделью.

Русско-татарская лексикографическая база данных создана на базе русско-татарского словаря под редакцией Ф.А. Ганиева, изданного в 1997 году, и постоянно пополняется за счет других лексикографических источников и за счет отслеживания новой лексики. Основным источником для отбора новых лексем является текстовая коллекция Татарского национального корпуса «Туган тел» [2].

В статье анализируется татарские новообразования – сложносоставные слова, образованные по модели: *существительное + аффиксоид ара,* выделенные из Татарского национального корпуса «Туган тел» [3] и текстов на татарском языке из сети Интернет, даются сведения об их частотности и особенностях употребления. Перевод татарских примеров на русский язык выполнен авторами статьи.

2. Узуальные единицы с компонентом *ара*

Среди лексических единиц с атрибутивным значением (прилагательных и наречий) важное место занимают частотные в употреблении слова, образованные по модели: *существительное + словообразующий формант ара*.

Существительное *ара* на русский язык может быть переведено следующими основными способами:

– промежуток, интервал, проход, зазор;
– расстояние, дистанция (спорт.);
– пробел, просвет;
– время, отрезок (промежуток) времени;
– разница (в возрасте);
– промежуточный;
– в значении послелога *арасында* – «среди, между» [4].

Образуя сложные прилагательные, существительное *ара* теряет конкретную предметную семантику; в Татарской грамматике говорится о том, что в составе сложных прилагательных компонент *ара*

выступает в качестве послелога [5,6]. В данной работе словообразующий формант *ара* мы будем называть аффиксоидом – корневой морфемой, выступающей в функции словообразовательного аффикса.

К узуальным единицам с компонентом *ара* можно отнести следующие зафиксированные в словарях единицы (в скобках дано количество употреблений в Основном корпусе Татарского национального корпуса):

халыкара – «международный» (38647);

милләтара – «межнациональный, межэтнический» (5317);

районара – «межрайонный» (1319);

дәүләтара – «межгосударственный» (100);

шәһәрара – «междугородний» (77);

вузара – «межвузовский» (49).

Кроме того, к частотным относятся и некоторые другие единицы, образованные путем сложения основ разных частей речи и аффиксоида *ара,* например:

үзара («сам» + *ара*) – «между собой, обоюдно» (18969);

берара («один» + *ара*) – «в течение какого-то промежутка времени» (1238).

Узуальные лексические единицы с атрибутивным значением, образованные при помощи компонента *ара*, отличаются различной частотностью употребления в текстах.

Аффиксоид *ара* участвует в актуальных процессах татарского словообразования и формирует вполне самостоятельный и весьма продуктивный словообразовательный тип, который постоянно пополняется новыми единицами, что требует своего изучения.

3. Методология исследования

Источником языкового материала послужили две коллекции текстов:

– Татарский национальный корпус «Туган тел» (главным образом коллекция Основного корпуса);

– коллекция текстов на татарском языке, извлеченных из сети Интернет.

Последняя коллекция представляет собой пока неочищенные тексты, собранные для дальнейшего пополнения корпуса, ниже данная коллекция будет называться Дополнительной. Обе коллекции были подвергнуты морфологическому анализу при помощи морфо-

анализатора корпуса [7], при этом был получен список нераспознанных единиц – слов и словоформ, которые отсутствуют в словаре корпуса. Список нераспознанных слов представляет собой пеструю смесь, состоящую из различных единиц – в том числе с опечатками, со смешением кириллической и латинской графики и т. п., но вместе с тем список нераспознанных слов является надежным источником для получения перечня новообразований, функционирующих в языке, но пока отсутствующих в словаре корпуса.

Из списка нераспознанных слов нами были извлечены единицы с компонентом *ара* (первоначально более 70 единиц); для последующего анализа были отобраны только те единицы, которые не нарушают строгих правил татарского словообразования и орфографии (в частности, были удалены единицы, построенные по модели: прилагательное + *ара*, например, *этникара* – «межэтнический»; единицы, написанные через дефис, например, *этнос-ара* – «межэтнический»). В дальнейшем отобранные единицы были подвергнуты количественному и контекстному анализу с выделением основных семантических группировок (рис. 1).

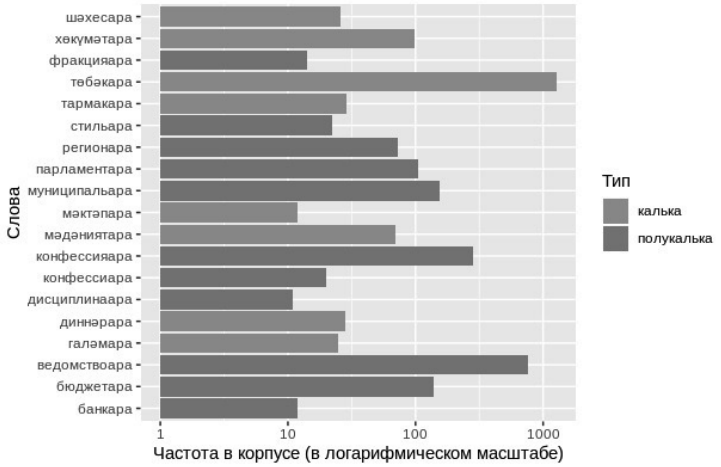


Рис. 1. Распределение слов с аффиксоидом -ара в корпусе

Сам факт, что корпусный морфоанализатор не распознает данные слова, свидетельствует о том, что они не включены в корпусный словарь. Кроме того, полученные лексические единицы про-

верялись на предмет их представления в имеющихся словарях: в Татарско-русском словаре под редакцией Ф.А. Ганиева [4] и Толковом словаре татарского языка [9].

4. Анализ языкового материала и основные результаты

Для анализа нами было отобрано 57 единиц, отсутствующих в словарях, которые были поделены на семантические группировки в зависимости от значения первого компонента.

Первый компонент – слово, обозначающее географическое понятие

К данной группе отнесены единицы, образованные от слов, которые обозначают различные территориальные образования:

төбәкара – «межрегиональный» (3742);
регионара – «межрегиональный» (72);
өлкәара – «межрегиональный, межобластной» (9);
республикаара – «межреспубликанский» (6);
поселокара – «межпоселковый» (4);
территорияара – «межтерриториальный» (1);
кантонара – «межкантонный» (2).

Слова данной группы наиболее часто используются в качестве определения при обозначении общественно-политических и экономических структур, в этом случае они являются калькой с исходных русских эквивалентов, например:

төбәкара мультимодаль логистик үзәк – «межрегиональный мультимодальный логистический центр»;

төбәкара клиника-диагностика үзәге – «межрегиональный клинико-диагностический центр»;

төбәкара һәм милли сәясәт комитеты – «комитет по региональной и национальной политике»;

төбәкара хаҗ мәркәзе – «межрегиональный паломнический центр центр (по хаджу)».

Однако синтаксическое поведение дериватов с аффиксоидом *ара* может быть не столь однородным, и не во всех случаях они могут быть сведены к соответствующим русским единицам, например:

Ел әйләнәсе эшләргә тиешле мондый станцияләр республикада гына түгел, ихтыяҗ була калса, ***өлкәара, республикаара*** *урман янгыннарын сүндерүдә дә катнашачак* («Татар-информ»). – «Такие

станции должны функционировать круглый год и, в случае необходимости, принимать участие в тушении лесных пожаров, произошедших на территории одновременно нескольких областей или республик».

Первый компонент – существительное, обозначающее понятие социокультурной сферы

К данной группе нами отнесены слова, образованные от названий важнейших социокультурных реалий:

телара – «межъязыковой» (8);

динара – «межрелигиозный» (количество единиц в корпусе невозможно определить по причине неснятой омонимии);

конфессияара – «межконфессиональный» (831);

мәдәниятара – «межкультурный» (70);

этносара – «межэтнический» (12);

культураара – «межкультурный» (1);

сыйныфара – «межклассовый» (4);

музейара – «межмузейный» (2);

цивилизацияара – «межцивилизационный» (5);

ыруара – «межродовой» (4).

В ряде контекстов данные единицы переводятся на русский язык предложно-падежными сочетаниями, например:

музейара хезмәттәшлек – «сотрудничество между музеями»;

ыруара сугышлар – «войны между родами».

Выявлено, что в татарском языке по этой модели формируются единицы, обозначающие реалии арабо-мусульманского мира, которые не имеют аналогов в русском языке:

мәзхәбара – «между мазхабами» (в атрибутивном значении) (4);

мәсхәбара – «между мазхабами» (в атрибутивном значении) (2).

Максат – ***мәзхәбара*** *низаг чыгару* (Радио “Азатлык”). – «Цель – разжигание розни между различными мазхабами».

Различное орфографическое оформление в татарском языке слова, обозначающего мазхаб – школу шариатского права в исламе, – приводит к тому, что производные слова с аффиксоидом *ара* зафиксированы в различных орфографических вариантах.

К этой же группе можно отнести единицы, мотивированные названиями структур системы образования и их частей:

мәктәпара – «межшкольный» (12);

университетара – «межуниверситетский» (2);

факультетара – «межфакультетский» (3).

Первый компонент – слово, обозначающее абстрактное понятие

Сюда нами отнесены слова с широкой сферой денотации, репрезентирующие различные общенаучные и общие понятия:

системаара – «межсистемный» (1);

кластерара – «межкластерный» (1);

компонентара – «межкомпонентный» (2);

секторара – «межсекторный» (1);

фәнара – «междисциплинарный, межнаучный» (1);

предметара – «межпредметный» (6);

бүлекара – «между разделами» (в атрибутивном значении) (2);

программаара – «межпрограммный» (1);

линияара – «межлинейный» (1).

В корпусных контекстах часто эти образования, как правило, сочетаются с существительными, обозначающими связи и отношения:

программаара функциональ элемтәләр – «межпрограммные функциональные связи»;

компонентара бәйләнеш – «связь компонентов»;

бүлекара бәйләнеш – «связь между разделами»;

предметара бәйләнешләр – «межпредметные связи».

Первый компонент – существительное, обозначающее промежуток времени

К данной группе отнесены слова, образованные от слов, обозначающих промежутки времени разной длительности:

сезонара – «межсезонный» (1);

фазаара – «межфазовый» (1 употребление в Дополнительной коллекции);

тәүлекара – «посуточный, посменно через одни сутки» (1).

В качестве производящей основы послужили как заимствованные слова (*сезон, фаза*), так и слова собственно татарского происхождения. Но между приведенными выше примерами существует важное различие. Слова *сезонара* и *фазаара* являются полукалькаами, образованными путем перевода соответствующих русских слов. Между тем со словом *тәүлекара* ситуация более интересная; рассмотрим пример употребления данного слова:

Янгын сүндерү бүлекчәсендә тәүлекара дежур тору оештырылган (Р. Галикаева). – «В пожарном отделении организовано посуточное дежурство».

Как уже отмечалось выше, одно из значений слова *ара* является значение «время, отрезок (промежуток) времени», это отражается в таких частотных оборотах татарского языка, как *көн аралаш* – «посменно через день», *шул арада* – «между тем», *кай арада* – «когда» и др. Именно это значение и стало базовым при образовании единицы *тәүлекара.*

Первый компонент – существительное, обозначающее государственные и общественные структуры и понятия, связанные с их деятельностью

партияара – «межпартийный» (2);
фиркаара – «межпартийный, межфракционный» (1);
фракцияара – «межфракционный» (14);
парламентара – «межпарламентский» (104);
ведомствоара – «межведомственный» (2197);
тармакара – «межотраслевой» (28);
хуҗалыкара – «межхозяйственный» (4);
хөкүмәтара – «межправительственный» (99);
муниципалитетара – «межмуниципальный» (4);
производствоара – «межпроизводственный» (1);
бюджетара – «межбюджетный» (134);
сайлауара – «межвыборный» (1).

Слова данной группы, образованные путем калькирования (кальки и полукальки) используются в текстах общественно-политической тематики.

Первый компонент – существительное, обозначающее единицу речевого сообщения

Единицы из данной группы получены путем калькирования соответствующих русских терминов:

текстара – «межтекстовый» (2);
сүзара – «межсловный» (3);
строфаара – «междустрофный» (?);
фразаара – «межфразовый» (2).

Они используются в текстах научной тематики, посвященных филологическому анализу:

Текстара бәйләнешләр, жанр төзелеше, конфликт төрләре, тема-мотивларны чагыштырып бәяләү компаративистик анализ юлы белән башкарыла (А. Закирзянов). – «Анализ межтекстовых связей, строение жанров, виды конфликтов, особенностей темы и мотивов осуществляется при помощи метода сравнительного анализа».

Любопытным является следующий пример, где словообразующий формант *ара* является фразовым (групповым), присоединяясь не к отдельной основе, а к сочинительной конструкции:

Шулай итеп, ***җөмлә, фраза, строфа, сүзара*** *бәйләнешләр көчәйгәннән-көчәя* (Р. Харрасова). – «Таким образом, связи **между предложениями, фразами, строфами,** словами все усиливаются».

Данная модель в татарском языке стала вполне продуктивной, что приводит к образованию единиц, не имеющих цельнооформленных аналогов в русском языке, например:

җөмләара – «между предложениями» (слово с атрибутивным значением) (2).

Дөрес, барлык строфаларда апрель тасвирлану дәвам итсә дә, ***җөмләара*** *бәйләнешләр ноктасыннан караганда, тема-рема өзеклекләре дә очрый* (Р. Харрасова). – «Хотя и во всех строфах продолжается описание апреля, с точки зрения связи между предложениями наблюдаются разрывы между темой и ремой».

Анализ новообразований с компонентом *ара* позволяет заключить, что основная часть таких единиц возникла в результате калькирования соответствующих русских лексических единиц с префиксами меж- и между-. Такие кальки и полукальки используются преимущественно в текстах общественно-политической и научной тематики. Тем не менее калькирование не является единственным источником образования единиц с формантом *ара.* Выявлены слова и контексты, в которых образования на *ара* не имеют прямых эквивалентов в русском языке, следовательно, можно говорить о том, что здесь задействованы механизмы собственно татарского словообразования и грамматики.

5. Новообразования, мотивированные формой множественного числа существительных

Татарское словообразование допускает образование новых слов от форм со словоизменительными аффиксами. Среди выявленных нами новообразований оказались слова, образованные от существительных с аффиксом множественного числа -лар/-ләр. Табл. 1 показывает распределение новообразований, мотивированные формой множественного числа существительных, в коллекции Основного и Общественно-политического подкорпусов Татарского национального корпуса.

Таблица 1. **Новообразования с формантом *ара*, мотивированные формой множественного числа существительных, в корпусных коллекциях**

Лексическая единица	Русский перевод	Количество в Основном корпусе	Количество в Общественно-политическом подкорпусе
илләрара	межгосударственный	1	0
милләтләрара	межнациональный	1	0
вузларара	межвузовский	1	2
мәдәниятләрара	межкультурный	0	6
конфессияләрара	межконфессиональный	0	5
диннәрара	межрелигиозный	27	24
телләрара	межъязыковой	0	1

Несмотря на то, что часть примеров обнаружена в единственном экземпляре, контексты с ними весьма показательны. Так, в примере ниже слово *илләрара* использовано в сочинительной конструкции с аналогичным образованием без аффикса множественного числа, и появление аффикса множественного числа в слове *илләрара* можно объяснить тем, что говорящий хочет акцентировать большое количество стран-партнеров по сотрудничеству.

Аеруча барлык илләрдә кабул ителгән кагыйдәләр уртак булса милләтара, ***илләрара*** *ширкәтләргә эшне көйләргә уңайлы була* (А. Нур). – «Особенно если во всех странах действуют одинаковые правила, легко наладить работу межнациональных, межгосударственных предприятий».

В следующем примере словообразовательный формант используется как фразовый и относится к сочинительной конструкции с двумя существительными – то есть одновременно к обоим существительным:

Балаларны бәйрәм белән ТР Мәдәният министрлыгы исеменнән әлеге ведомствоның ***төбәкләр һәм милләтләрара*** *хезмәттәшлек һәм күргәзмә эшчәнлеге бүлеге җитәкчесе Рәисә Сафиуллина тәбрикләде* («Татар-информ»). – «Детей от имени Министерства культуры РТ поздравила начальник отдела межрегионального, межнационального сотрудничества и выставочной деятельности» (Раиса Сафиуллина).

Наличие образований, возникших на базе основы, осложненной аффиксом множественного числа, можно объяснить следующим об-

разом: аффиксоид *ара* используется при образовании лексических единиц, обозначающих отношения между совокупностью мотивирующих референтов, то есть логически всегда подразумевается некая имплицитная совокупность, которая обозначается явным образом при помощи аффикса множественного числа.

6. Заключение

Пополнение словарей предполагает в качестве необходимого звена сбор и изучение новой лексики с целью отбора слов-кандидатов для включения в словари. Анализ единиц, реально употребляющихся в текстах на татарском языке, квалифицируемых автоматическими морфоанализаторами как нераспознанные, является важным шагом для обнаружения новой лексики.

Анализ татарских новообразований с компонентом *ара* свидетельствует о том, что данный компонент активно используется в качестве строительного блока при калькировании русских прилагательных с префиксами *меж-* и *между-*. Однако значение данного компонента не сводится к этому. В татарском языке появляются новые единицы, не имеющиеся в русском языке или переводимые на русский язык не только прилагательными на *меж-* и *между-*.

Материал исследования позволяет заключить, что в современном татарском словообразовании механизм калькирования и ресурсы собственно татарского словообразования и грамматики работает в органическом единстве. Продуктивность словообразовательной модели существительное + формант *ара* свидетельствует о том, что носители языка, создавая новые единицы по данной модели, не воспринимают их в качестве окказионализмов и легко продуцируют новые единицы, не зафиксированные в словарях. Употребление *ара* в качестве группового (фразового) словообразующего форманта, а также наличие слов, образованной от основы, осложненной аффиксом множественного числа, также свидетельствует о том, что компонент *ара* воспринимается носителями как стандартное татарское словообразующее средство, не ограниченное списком калькированных единиц.

ЛИТЕРАТУРА

1. Невзорова О.А., Гатиатуллин А.Р., Гильмуллин Р.А., Салимов Ф.И., Хакимов Б.Э., Аюпов М.М. Двуязычный лексикографический ресурс для задач автоматической обработки текстов: грамматическая

модель Т-компоненты // Труды Казанской школы по компьютерной и когнитивной лингвистике TEL-2012. – Казань: Изд-во «Фэн» Академия наук РТ, 2012. – С. 28–34.

2. Татарский национальный корпус «Туган тел». Электронный адрес: http://tugantel.tatar/

3. Невзорова О.А., Мухамедшин Д.Р., Билалов Р.Р. Корпус-менеджер для тюркских языков: основная функциональность // Труды международной конференции «Корпусная лингвистика – 2015». – СПб.: С.-Петербургский гос. университет, филологический факультет, 2015. – С. 344–350.

4. Татарско-русский словарь / Ред. Ф.А. Ганиев – Казань: Татарское книжное издательство, 2004. – 488 с.

5. Татарская грамматика: в 3 т. Т.1. / Гл. ред. М.З. Закиев. – Казань : Татарское книжное издательство, 1993. – 584 с.

6. Татар грамматикасы: өч томда /проект җит. М. З. Зәкиев; ф. ред. Ф. М. Хисамова.– Казан: ТӘһСИ, 2015. – 512 б.

7. Gilmullin R., Gataullin R. Morphological Analysis System of the Tatar Language. In: Nguyen N., Papadopoulos G., Jędrzejowicz P., Trawiński B., Vossen G. (eds) Computational Collective Intelligence. ICCCI 2017. Lecture Notes in Computer Science, vol 10449. Springer, Cham. – Pp. 519–528.

9. Татар теленең аңлатмалы сүзлеге / Баш ред. Ф. Ә. Ганиев. – Казан: Матбугат йорты, 2005. – 848 б.

УДК 81'32; 81.512.1; 004.9

A NEURAL NETWORK APPROACH TO MORPHOLOGICAL DISAMBIGUATION BASED ON THE LSTM ARCHITECTURE IN THE NATIONAL CORPUS OF THE TATAR LANGUAGE

R. A. Gilmullin[1], B. E. Khakimov[1,2], R. R. Gataullin[1]
[1]Institute of Applied Semiotics of the Tatarstan Academy of Sciences, Kazan, Russia
[2]Kazan Federal University, Kazan, Russia
rinatgilmullin@gmail.com, bulat.khakeem@gmail.com, ramil.gata@gmail.com

This paper presents the results of experiments on morphological disambiguation in the National corpus of the Tatar language “Tugan tel”. The experiments were conducted using the LSTM based neural network model. The tagged socio-political sub-corpus of the National corpus of the Tatar language “Tugan tel” with a volume of 2,4 million words was used as training data. Experiments have shown that LSTM models are language-independent and can be applied to the Tatar language too. The results for Tatar are on a comparable level with those for other agglutinative languages, such as Hungarian and Turkish.

Keywords: Morphological disambiguation, Tatar language, Tatar National Corpus, corpus data, morphological tagging, LSTM, neural architectures.

РАЗРЕШЕНИЕ МОРФОЛОГИЧЕСКОЙ МНОГОЗНАЧНОСТИ В КОРПУСЕ ТАТАРСКОГО ЯЗЫКА

Р. А. Гильмуллин[1], Б. Э. Хакимов[1,2], Р. Р. Гатауллин[1]
[1]Институт Прикладной семиотики Академии наук Республики Татарстан, Казань
[2]Казанский федеральный университет, Казань
rinatgilmullin@gmail.com, bulat.khakeem@gmail.com, ramil.gata@gmail.com

В данной работе представлены результаты исследований, посвященных разрешению морфологической многозначности в национальном корпусе татарского языка «Туган тел». В качестве модели для разрешения многозначности выбрана нейросетевая модель на основе архитектуры LSTM. Для обучения модели использовался размеченный общественно-политический подкорпус национального корпуса татарского языка «Туган тел» со снятой морфологической многозначностью. Эксперименты показали, что модель

на основе LSTM не зависит от языка и дает достаточно высокие результаты для татарского языка, которые сопоставимы с результатами для других агглютинативных языков, таких как венгерский и турецкий языки.

Ключевые слова: Морфологическая многозначность, татарский язык, Татарский национальный корпус, нейросетевая модель, LSTM.

Введение

Разрешение многозначности является одной из основных задач автоматической обработки естественного языка. Результаты разрешения могут использоваться для повышения точности и улучшения качества применяемых методов в таких задачах, как классификация и кластеризация текстов, машинный перевод, информационный поиск.

Сложность и особенности разрешения многозначности для каждого конкретного языка проявляются по-разному. Например, для английского языка с бедной морфологией и жестким порядком слов в предложении разрешение морфологической многозначности, как правило, сводится к задаче POS-теггинга и решается применением достаточно простых методов. Для русского языка морфологическая многозначность не столь характерна, как для английского и татарского, но, тем не менее, присуща. Дополнительную сложность добавляет свободный порядок слов в русском языке. В татарском языке, как и в других агглютинативных языках тюркской группы, морфемы являются важнейшими значащими языковыми единицами, которые несут как семантическую, так и синтаксическую информацию. При теоретически неограниченном количестве присоединяемых к основе морфем, морфологическая многозначность приобретает разнообразные формы, что значительно усложняет задачу разрешения.

К настоящему времени сформирована основная парадигма методов снятия многозначности, которая включает методы, основанные на правилах; методы машинного обучения, использующие вероятностные модели; гибридные методы [1]. Создание электронного корпуса татарского языка «Туган тел» (http://tugantel.tatar/) и общественно-политического подкорпуса со снятой вручную морфологической многозначностью дали возможность исследования данной проблемы с применением статистических методов на основе технологий машинного обучения [2, 3].

Анализ открытых программных кодов, разработанных для этой задачи в последние несколько лет, показал, что одними из эффек-

тивных являются инструментарий PurePos 2.0 [4], реализующий гибридную модель на основе скрытых марковских моделей, а также нейросетевая модель на основе рекуррентных нейросетей с долгой краткосрочной памятью LSTM [5]. Скрытая марковская модель − модель процесса, в которой процесс считается марковским, причем неизвестно, в каком состоянии находится система (состояния скрыты), но каждое состояние может с некоторой вероятностью произвести событие, которое можно наблюдать. Другими словами, изучается марковский процесс с неизвестными параметрами, и задачей является распознавание неизвестных параметров на основе наблюдаемых. Результаты по распознаванию POS-тегов татарских слов показали точность 97% [6].

Еще один подход, который достаточно успешно позволяет решать задачу разрешения морфологической многозначности, основан на рекуррентной нейронной сети с долгой краткосрочной памятью (англ., Long short-term memory, LSTM) [7, 8]. В работе [5] приводятся результаты применения данного подхода к турецкому, русскому и арабскому языкам. В работе заслуживает внимание анализ размера используемого контекста. Авторы сравнивали разные размеры и типы контекста и экспериментально выявили наиболее подходящий под каждый язык тип. Оказалось, что для турецкого языка достаточным является построение векторов на основе поверхностных форм слов без явного определения морфологических признаков, но с использованием всех слов в предложении. Тогда как для русского языка важным моментом является согласованность в роде, числе и падеже, что в свою очередь требует наличия не только поверхностной формы слова, но и морфологических признаков слов в контексте. При этом добиться лучших результатов (точность разрешения 91,13%) помогает оптимизация с помощью метода условных случайных полей (англ., Conditional Random Fields, CRF). Похожая ситуация и с арабским языком, когда поверхностных форм слов недостаточно для полного разрешения многозначности. Это можно объяснить тем, что в арабском доля многозначности больше, чем в турецком. Если, например, в турецком языке в среднем на одно слова приходится 2,81 вариантов разбора, в русском языке – 5,81, то в арабском языке – 11,31. Поэтому для правильного обучения модели требуется размеченный контекст с полностью снятой омонимией.

Далее представлены результаты применения нейросетевой модели на основе архитектуры LSTM для разрешения морфологической многозначности в корпусе татарского языка.

Модель на основе LSTM для разрешения морфологической многозначности

Для обучения модели требуются размеченные тексты со снятой многозначностью. Согласно [5], идея метода сводится к тому, что каждому разбору многозначного слова и окружающему его контексту сопоставляются вектора. В первом случае вектор строится на основе его леммы и морфологических признаков (см. рис. 1), после объединяются в матрицу (обозначена R); во втором на основе поверхностных форм окружающих слов (обозначена вектором h) (см. рис. 2); дополнительно вектор можно расширить и за счет морфологических признаков. При этом контекст не ограничивается несколькими словами в непосредственной близости и может достигать размеров всего предложения.

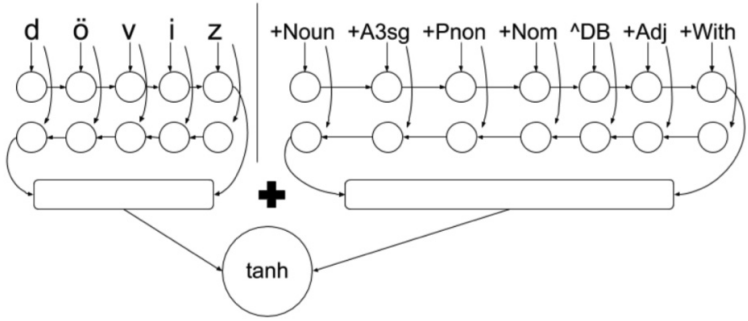


Рис. 1. Нейросетевая архитектура LSTM для получения векторного представления морфологического разбора

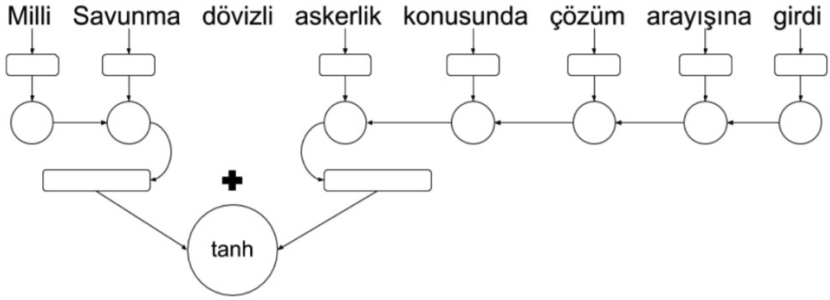


Рис. 2. Нейросетевая архитектура LSTM для получения вектора контекста

Далее путем применения функцию softmax на произведении матрицы R на вектор h, строится распределение вероятностей каждой из альтернатив многозначного слова в данном конкретном контексте, на основе которого морфологическая многозначность разрешается в пользу наиболее вероятной альтернативы:

$$p(y_t = a \mid x) = softmax(R_{x_t} \times h_t)$$

Подготовка данных

Для проведения экспериментов с обучением моделей необходимо создание корпуса со снятой морфологической многозначностью. В качестве данных для обучения использовался морфологически размеченный общественно-политический подкорпус национального корпуса татарского языка «Туган тел» со снятой вручную морфологической многозначностью. Статистика подкорпуса приведена в таблице 1. Ручное снятие морфологической многозначности в общественно-политическом подкорпусе выполнялось экспертами-лингвистами с помощью веб-инструментария для снятия морфологической многозначности в текстовом корпусе татарского языка [9]. По результатам проведенной работы было подготовлено 56524 предложения со снятой морфологической многозначностью.

Таблица 1. **Статистика обучающей и тестовой выборки по общественно-политическому корпусу**

	Обучающая выборка	Тестовая выборка
Количество контекстов (предложений)	54580	944
Количество токенов (включая пунктуацию)	600.480	11.655
Количество многозначных разборов	125480 (21%)	2527 (21%)
Количество уникальных словоформ	29953	2788
Количество уникальных лемм	7117	1226
Количество уникальных морфологических форм	1898	346

Результаты экспериментов

Как видно из таблицы 1, размеченная выборка данных была разделена на выборку для обучения и тестовую выборку. Модели на

LSTM обучались только на обучающей выборке, тестовая выборка использовалась только для тестирования. Каждая модель обучалась на одной и той же обучающей выборке и проходила валидацию на одной и той же тестовой выборке. В таблице 2 и 3 приведены оценка точности по нескольким показателям: распознавания леммы, аффиксальной цепочки и разрешения многозначности.

Таблица 2. **Показатели точности распознавания лемм и аффиксальной цепочки**

Показатели	LSTM NN
Точность распознавания леммы	11299 / 11655 = 96,94%
Точность распознавания аффиксальной цепочки	11127 / 11655 = 95,46%

Таблица 3. **Количество вариантов морфологического разбора и точность разрешения моделей**

Количество вариантов	LSTM NN
n=2	1545 / 1826 = 84,61 %
n=3	268 / 424 = 63,21 %
n=4	141 / 192 = 73,44 %
n=5	7 / 9 = 77,78 %
n=6	37 / 72 = 51,39 %
n=7	0 / 2 = 0,00 %
n=8	0 / 1 = 0,00 %
Общее	1999 / 2527 = 79,10%

Заключение

В данной работе представлены результаты работ по разрешению морфологической многозначности в татарском языке с использованием нейросетевой модели на основе архитектуры LSTM. Учитывая ограниченный набор корпусных данных для обучения, результаты экспериментов показали достаточно хороший уровень точности для разрешения морфологической многозначности: 84,36% и 79,10% соответственно. По мнению авторов, более низкие показатели точности нейросетевой модели прежде всего связаны с объемом обучаю-

щих данных. Тем не менее результаты LSTM практически близки к результатам других методов разрешения, а в некоторых случаях и превосходят их, например, для турецкого языка, как показано в таблице 5, авторы [5] добились результата в 96,41% точности разрешения.

Вместе с тем, полученные результаты могут быть эффективно использованы в создании «золотого» подкорпуса со снятой морфологической многозначностью, значительно сократив объем многовариантных разборов, требующих ручного снятия морфологической многозначности.

ЛИТЕРАТУРА

1. Гатауллин, Р.Р. Аналитический обзор методов разрешения морфологической многозначности. / Р.Р. Гатауллин // Российский научный электронный журнал (Электронные библиотеки). Том 19, № 2 (2016). – С. 98–114.

2. Gataullin R., Khakimov B., Suleymanov D., Gilmullin R. (2017) Context-Based Rules for Grammatical Disambiguation in the Tatar Language. // N.T. Nguen et al. (Eds). ICCCI 2017, Part II, LNAI 10449. – 2017. pp. 529–537.

3. Хакимов, Б.Э. Разрешение грамматической многозначности в корпусе татарского языка / Б.Э. Хакимов, Р.А. Гильмуллин, Р.Р. Гатауллин // Учен. зап. Казан. ун-та. Сер. Гуманит. науки. – 2014. – Т. 156, кн. 5. – С. 236–244.

4. Orosz, G. and Novák, A. 2013. PurePos 2.0: a hybrid tool for morphological disambiguation. Proceedings of Recent Advances in Natural Language Processing, pages 539–545, Hissar, Bulgaria, 7–13 September 2013. Online version: http://aclweb.org/anthology//R/R13/R13-1071.pdf.

5. Qinlan Shen, Daniel Clothiaux, Emily Tagtow, Patrick Littell, Chris Dyer. 2016. The Role of Context in Neural Morphological Disambiguation. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 181–191, Osaka, Japan, December 11–17, 2016. http://aclweb.org/anthology/C16-1018.

6. Гильмуллин, Р.А. Разрешение морфологической многозначности текстов на татарском языке на основе инструментария PurePos. / Р.А. Гильмуллин, Р.Р. Гатауллин // V МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ ПО КОМПЬЮТЕРНОЙ ОБРАБОТКЕ ТЮРКСКИХ ЯЗЫКОВ «TURKLANG 2017». – Труды конференции. В 2-х томах. Т – Казань: Издательство Академии наук Республики Татарстан, – С. 30–37.

7. V. Berment, “Me´thodes pour informatiser des langues et des groups

de langues peu dote´es", Ph.D. Thesis, J. Fourier University, Grenoble I, 2004.

8. S. Krauwer, "The basic language resource kit (BLARK) as the first milestone for the language resources roadmap", In Proc. of International Workshop Speech and Computer SPEECOM, Moscow, Russia, 2003, P. 8–15.

9. Гатауллин, Р.Р. Веб-инструментарий для снятия морфологической многозначности в текстовом корпусе татарского языка / Р.Р. Гатауллин // Сохранение и развитие родных языков в условиях многонационального государства: проблемы и перспективы: материалы V Международной научно-практической конференции (Казань, 19–22 ноября 2014 г.). – Казань: Отечество, – С. 71–73.

УДК 81'32

AUTHOR'S SPEECH IN THE NOVEL 'CRIME AND PUNISHMENT' BY FYODOR DOSTOEVSKY (DYNAMIC FREQUENCY DICTIONARY ANALYSIS)

A. A. Glezina
National Research University «Higher School of Economics», Saint-Petersburg
aaglezina@edu.hse.ru

The article presents the result of the analysis of the author's speech and author's remarks in the novel 'Crime and Punishment'. The analysis was carried out by compiling frequency and dynamic frequency dictionaries. The purpose of the analysis was to identify stylistic and lexical patterns in the author's speech.

Keywords: Dostoevsky, frequency dictionary, author's speech, author's remarks.

АВТОРСКАЯ РЕЧЬ В РОМАНЕ Ф.М. ДОСТОЕВСКОГО «ПРЕСТУПЛЕНИЕ И НАКАЗАНИЕ» (АНАЛИЗ ДИНАМИЧЕСКОГО ЧАСТОТНОГО СЛОВАРЯ)

А. А. Глезина
Национальный Исследовательский Университет «Высшая школа экономики», Санкт-Петербург
aaglezina@edu.hse.ru

В статье представлен результат анализа авторской речи и авторских ремарок в романе «Преступление и наказание». Использовался метод составления частотных и динамических частотных словарей. Целью анализа было выявление стилистических и лексических паттернов в авторской речи.

Ключевые слова: Достоевский, частотный словарь, авторская речь, авторские ремарки.

Язык, несмотря на свою непрерывную изменчивость посредством внешних влияний и внутренних преобразований, на сегодняшний день остается основным средством общения. Лингвистические инновации в нем приживаются достаточно медленно, – именно поэтому человек имеет возможность прочитать и понять написанное спустя не один век после создания текста. Особое внимание при рас-

смотрении данной проблемы следует уделить художественной литературе, так как этот вид искусства непременно должен включать в себя ярко-выраженный индивидуальный язык автора произведения. Проще говоря, язык литературного произведения представляет собой «индивидуально-языковое творчество» [Ружицкий 2015: 24]. Это понятие тесно связано с термином «языковая личность», который появлялся в трудах Виноградова, но, в отличие от «образа автора», не получило там большого развития. Тем не менее, если совместить эти два термина, можно сказать, что «языковая личность автора реализуется через совокупность созданных им текстов, их лексикон, особенности синтаксиса, композиционную структуру и т. д., через различные проявления в них образа автора и созданные художественные образы, которые, в свою очередь, также можно рассматривать как языковую личность» [Там же: 25]. В основном языковую личность принято рассматривать на базе употребляемых ею языковых средств, то есть авторского литературного стиля.

В данной работе рассматриваются отличительные черты одного художественного произведения – романа Ф.М. Достоевского «Преступление и наказание» – с использованием стилеметрических и лексикографических методов. Согласно определению, «Стилеметрия – прикладная филологическая дисциплина, занимающаяся измерением стилевых характеристик с целью систематизации (атрибуции, таксономии, периодизации, датировки и т. п.) текстов и их частей» [Мартыненко, 1988]. Лексикографией же называют «прикладную лингвистическую дисциплину, занятую составлением словарей» [Шайкевич 2005: 192] и исследованием предпочтений и интенсивности в использовании слов в речи.

В данной работе важным было найти не столько особенности языка, сколько особенности авторской речи. В этом случае будет полезен метод составления частотных словарей. Частотный словарь позволяет увидеть набор наиболее употребительных лемм, необходимых для понимая основных положений и концепций определенного подраздела информации, для описания которого используется индивидуальный язык. Алексеев утверждает, что «первая тысяча самых употребительных слов может обеспечить покрытие 80% всех словоупотреблений текста, а первые две тысячи слов – до 90%» [Алексеев 2001: 10]. В данном случае частотные словари будут составлены на основе романа Достоевского. Нельзя игнорировать тот факт, что миру уже известен один словарь языка Достоевского, созданный под руководством Шайкевича (2005) и максимально

полный. Однако существует способ расширить данную область исследований посредством составления динамического частотного словаря для начала на материале одного романа. Подобный подход до сих пор не был освящен не только среди исследователей Достоевского и других русских классиков.

Выбор в качестве объекта исследования творчества Ф. М. Достоевского обусловлен следующими факторами. Как пишет Г.Я. Мартыненко: «Значительность писателя сильно коррелирует с самобытностью, оригинальностью индивидуальной манеры письма» [Мартыненко 1988: 106]. Этот тезис, спроецированный на личность Достоевского, находит отражение в работе И.В. Ружицкого: «Именно Достоевский, благодаря своей ярко выраженной нестандартности, экспериментированию с литературным языком и функциональными стилями, практически полному растворению в своих персонажах, создает у читателя впечатление абсолютного исчезновения авторской точки зрения, и поэтому тексты Достоевского являются наиболее подходящим исследовательским полем для решения задачи реконструкции образа автора...» [Ружицкий 2015: 20]. По причине имплицитности образа автора другие герои (включая второстепенных) становятся активными выразителями или оппонентами авторской позиции. Именно поэтому тематические сферы и жанрово-стилистические особенности творчества Достоевского вызывают неподдельный интерес.

Исследовать, как Достоевский осуществляет переплетение образов автора и героев на уровне текста романа, будет удобнее, выделив перед составлением частотных словарей три текстовых микрожанра (с опорой на вышеупомянутый «Статистический словарь языка Достоевского» (2005) В.Я. Шайкевича):

– авторская речь (поделена по частям романа);

– авторские ремарки при репликах Раскольникова (поделены по частям романа);

– авторские ремарки при репликах героев (поделены по героям).

Авторскую речь в данной работе рассматривается в динамике: по частям текста, соответствующим частям романа. Голос автора часто сливается с голосом Раскольникова (отсюда частое появление несобственнопрямой речи), а значит, образ автора так же переживает некую эволюцию, почему его и следует детально изучить в динамике.

Таким образом, в статье будут рассмотрены частотные словари, составленные согласно списку микрожанров, указанных выше. На

основе выявленных паттернов употребляемости каких-либо лемм будет возможность воссоздать образ автора. Также нашей целью будет выявление высокочастотных лемм, которые бы с неочевидной до этого точки зрения характеризовали героев романа через речь автора и его ремарки. Рассматривая элементы, встречающиеся нам в таблицах частотности, обратим внимание на конкорданс и биограммы, в состав которых они входят. В роли варьирующего признака выступят ранги, а в роли статистических весов – абсолютные и относительные (ipm) частоты, соответствующие им.

1. Авторская речь

Появление термина «языковая личность», как отмечает Ружицкий, имеет прямое отношение к личности В. В. Виноградова, а именно – к поставленным им задачам по изучению литературного языка и языка художественной литературы и необходимости введения в связи с этим категории «образа автора» [Ружицкий 2015: 20]. В этом разделе речь пойдет именно об этом образе, а потому будет не лишним дать пояснение, почему нижеизложенная интерпретация некоторых словоупотреблений тесно связана с главным героем романа, Родионом Раскольниковым, и (в меньшей степени) с другими персонажами. Прежде всего, по знаменитой концепции Бахтина, «Преступление и наказание» – это полифонический роман [Бахтин 1979: 59]. Применяя это понятие к произведению, мы априори признаем, что «голос автора допустим лишь на равных правах с голосами героев» [Габдуллина 2007: 93]. Кроме того, несмотря на то, что «Преступление и наказание» – это единственное крупное произведение Достоевского, написанное от третьего лица, изначальный замысел автора был другим: первоначальная идея, по словам писателя, – «это психологический отчет одного преступления. Действие современное, в нынешнем году: “Я под судом и все расскажу. Я все запишу. Я для себя запишу, но пусть прочтут и другие. Это исповедь. Ничего не утаю”» [Батурина 2006: 37]. Это объясняет тесную связь внутренних миров автора и его героев. Первый постоянно присутствует в их жизни не только сторонним наблюдателем, но и непосредственным участником и сопереживающим. Не стоит забывать, что все поднятые в романе Достоевским проблемы волновали и его самого. Это подтверждает следующее высказывание: «Достоевский был художником-лириком, который в особенности пишет о себе, для себя и от себя. Все его повести и романы – одна огненная река

его собственных переживаний. Это сплошь признание сокровенного своей души. Это первый и основной момент в его творчестве. Второй – постоянное стремление заразить, убедить, потрясти читателя и исповедать перед ним свою веру» [Габдуллина 2007: 93]. Любая оценка, которая встречается в романе, является не только авторской, но и проходит через призму сознания героя, о чем говорит еще и обилие конструкций с использованием несобственно-прямой речи: «Наконец, пришло ему в голову, что не лучше ли будет пойти куда-нибудь на Неву?» [Достоевский 2014: 67]. Перейдем непосредственно к частотному анализу авторской речи в романе, потому как, делая автора повествователя ключевой фигурой произведения, объединяющей все идеи, смыслы и переживания его действующих лиц, Достоевский не лишает его права на собственный язык, который делает всю систему речевых структур единой и создает стилевую целостность текста и его восприятия. К анализу этого языка мы и приступаем на следующем этапе.

Начнем с рассмотрения части частотного словаря первой части романа в жанре авторской речи, попутно проводя сопоставление с другими частями для прослеживания наличия динамики определенных лемм. Здесь и далее в работе были взяты первые по частоте 100 слов из всего списка и отобраны, на наш взгляд, наиболее значимые, то есть смыслообразующие слова; опускались в основной массе предлоги, частицы, местоимения, союзные слова и другие леммы, служащие более для синтаксической и лексической связности текста, нежели для смыслообразования. Изначально самым употребляемым глаголом является глагол «быть» (ранг 4, частота 301), что можно будет наблюдать и в остальных частях микрожанра авторской речи. В данном случае это является не только типичным маркером языка авторской речи, повествования, но и, если посмотреть на конкретные случаи словосочетаний, показывает предрасположенность Достоевского описывать внутренние состояния героев чаще, чем сухие факты: «был задавлен бедностью», «был очень доволен», «взгляд был неопасен», «был голоден», «было душно». Кроме того, в первой части романа эта лемма часто употребляется для описания различных пространств, что создает гнетущую атмосферу тесноты и погружает в нее читателя: «была крошечная кухня», «было всего только два стула». Наконец, нередко в тексте появляются сочетания прямо написанные в форме условного наклонения или же создающие такой эффект при восприятии: «должен был бы», «надо было», «хотел было» – такие конструкции изначально формируют представление

о том, что главный герой делает что-то не так, делает неверный выбор. Он осознает это внутри и уже как будто сожалеет, но пока мы видим это только через подсказки в речи автора.

Следующей наиболее символичной стоит назвать лемму «один» (ранг 21, частота 66). Кирпотин пишет, что Достоевский «показывает Раскольникова подчеркнуто одиноким» [Кирпотин 1970: 7]. Непосредственно в тексте не так очевидна семантика одиночества, но из-за этого скрытый эмоциональный компонент действует еще эффективнее. Прежде всего, Раскольников сам отгораживается от других людей, они не представляют для него никакой ценности. С помощью подстановки слова «один» перед наименованием любой персоны автор этого человека просто-напросто обезличивает: «один пьяный», «один семейный немец», «один господин», «один знакомый» и т. д. Это можно соотнести с философией Раскольникова о людях как «тварях дрожащих», которых он, человек исключительный, имеет право судить. Тем не менее, автор в самом начале имплицитно указывает внимательному читателю на ошибочность этой теории: Раскольников назван «одним молодым человеком», а значит он сам является одним из многих обычных людей и имеет с ними равные права. Этот случай употребления обнажает, прежде всего, внутреннее ощущение покинутости и разрозненности общества. Есть, конечно, и словосочетания, указывающие непосредственно на физическое одиночество героя: «он один в комнате», «один он не мог», «оставшись один», «один только Раскольников». Также нередки использования в паре с предметами быта и окружения: «один огарок», «один угол», «выбил один кирпич»; в револьвере у Свидригайлова остается «один капсюль», многое в романе происходит «в один миг». Все это указывает не только на постоянное состояние одиночества, но и на то, что в жизни героев Достоевского достаточно одного верного или неверного поступка, чтобы полностью поменять ход развития событий. Если проследить изменение частоты появления леммы «один» по всем частям романа, можно увидеть (см. рис. 1), что заметное снижение частоты приходится на третью и четвертую части и эпилог (7), в котором уровень ipm практически сравнивается с изначальным. Такие колебания могут быть логично объяснены через призму развития событий в сюжете произведения: в начале романа герой вынашивает свою аморальную идею и ни с кем не может поделиться внутренними переживаниями, полностью уходит в себя. В третьей и четвертой части Раскольников, напротив, проводит много времени с матерью и сестрой, которых он дей-

ствительно любит и присутствие которых возвращает его в семью, круг людей, которым он не безразличен. Также в жизни Родиона Романовича появляется любопытный ему Свидригайлов, человек, наверное, наиболее интересный с точки зрения психологизма, и, хоть и неприятный, но в чем-то близкий по духу Раскольникову: («Мы одного поля ягоды. <…> Вот, может, сойдемся поближе» [Достоевский 2014: 216]). Кроме того, в четвертой части главный герой находит поддержку в лице Сони Мармеладовой, и после знаменитого чтения библии и признания в убийстве между ними случается духовное сближение. В шестой же части (вновь взлет показателя, но на этот раз ipm выше, чем в начале повествования) совершается, во-первых, расставание Раскольникова с близкими людьми, во-вторых, самоубийство Свидригайлова (который, в отличии от протагониста, не выдержал тяжести своих грехов) и, наконец, признание Раскольникова в убийстве с последующим его нежеланием быть сопровождаемым Соней на каторгу. В эпилоге (здесь и далее на рисунках будет обозначаться как 7) же, как становится ясно из повествования, происходит духовное перерождение Раскольникова, Соня осознает истинность любви молодого человека к ней («Она поняла, и для нее уже не было сомнения, что он любит, бесконечно любит ее» [Достоевский 2014: 307]), и автор дает нам надежду на возможность лучшего будущего. Это отражается и на частотном показателе рассматриваемой леммы: он почти сравнивается с изначальным, но все же незначительно превышает его.

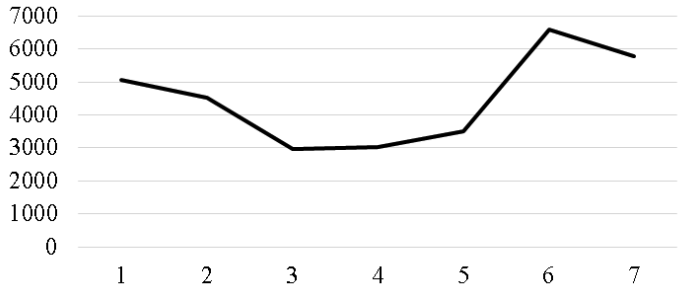


Рис. 1. Динамика ipm леммы «один» в микрожанре авторской речи

Схожую с предыдущей тенденцию распределения частот имеет лемма «вдруг». На ее особенный статус у Достоевского, в отличии от леммы «один», исследователи уже обращали внимание (в частно-

сти, одним из первых был Бахтин, написавший об этом в «Проблеме поэтики Достоевского» (1979)). Касаткина характеризует употребление этого слова в романах русского классика следующим образом: «... переход совершается вдруг. Это вдруг, эту моментальность перехода от одного к другому, от отсутствия к присутствию, от ада к раю, от падения к восстанию Достоевский подчеркивает просто навязчиво <…> В этом же вдруг заключается любимая мысль Достоевского о возможности мгновенного преображения, в любой момент, преображения, зависящего от каждого из нас.» [Касаткина 2008: 30] Несомненно, несмотря на то, что путь к преображению для Раскольникова был далеко не мгновенным, в данной лексеме в рамках «Преступления и наказания» заключена мысль о том, что все действительно происходит неожиданно, и, как бы герою ни хотелось взять контроль над происходящим, для него все не перестанет случаться «вдруг». Это вновь отсылает нас к главной идее романа: контроль человеческой жизни подвластен исключительно Богу.

Как можно видеть на графике (см. рис. 2), чем ближе Раскольников к смирению, тем меньше ситуаций в его жизни сопровождается наречием «вдруг». Это снижение последовательно. Небольшое увеличение показателя относительной частоты присутствует в четвертой части, но, приняв во внимание процентное соотношение с полным объемом текста, этот рост можно считать незначительным или же связать с неожиданным приездом семьи героя в Петербург.

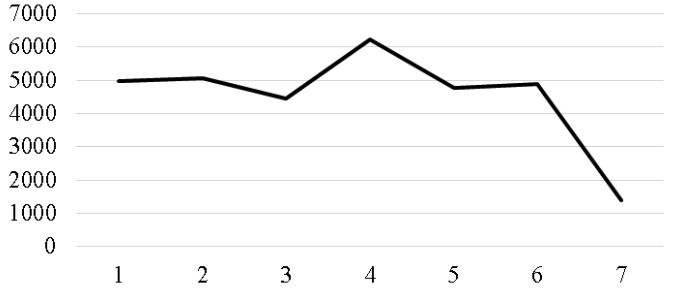


Рис. 2. Динамика ipm леммы «вдруг» в микрожанре авторской речи

Подобный анализ в дальнейшем возможно провести и для других смыслоформирующих лемм, что дает обширный материал для будущих исследований. Даже на рассмотренной паре примеров становит-

ся очевидно, что данная тема и направление исследований актуально и перспективно. (В частности, после просмотра частотных словарей первого микрожанра интерес также вызвали слова: *мочь, будто, лицо, рука, становиться, минута* и др.) Такого же подхода решено придерживаться и далее при рассмотрении второго микрожанра.

2. Авторские ремарки

Для усиления психологизма и достижения максимального обнажения внутренних состояний своих персонажей Ф.М. Достоевским в романе были также проработаны и авторские ремарки, сопровождающие прямую и внутреннюю речь героев. Ремарки в «Преступлении и наказании» не только передают интонационную составляющую реплики говорящего, но и иллюстрируют эмоциональное и даже физическое состояние героя на момент произнесения им фразы. Как отмечает в своей работе Л. Пашкявичене, «определенная чуткость – в разной степени – к звуковой стороне речи своих героев характерна для литературы конца XIX–XX вв.: в большей степени это проявляется в творчестве Ф. М. Достоевского, великого знатока души человека, который отличается обостренным вниманием к интонации своих героев. Важную роль в тексте романа «Преступление и наказание» как средство отражения интонации играют мимика, жесты, физиологические реакции персонажа, поскольку описание невербальной речи подчеркивает особенности голоса, тона, а также в некоторых случаях заменяет их, позволяя тем самым угадать характер звучания» [Пашкявичене 2009: 42].

Если говорить конкретно о классе «глаголов речи» (глаголами речи называются все глаголы, которые вводят в тексте прямую речь, называют акт говорения [Маркевич 2016: 2]), Пашкявичене было установлено (и подтвердилось при анализе частотных словарей, составленных для данной работы), что наиболее высокой частотой употребления характеризуются такие леммы как «сказать», «говорить», «спросить», «ответить», «прибавить», «возопить». Кроме того, в романе очень часто используются глаголы говорения, которые отражают особенности артикуляции и звучания по разным акустическим признакам: «пробормотать», «прошептать».

Все же Достоевский не был бы Достоевским, если бы не привнес и в этот компонент своего произведения «стилистическую инновацию»: даже в сочетании с такими «бесцветными» словами как «сказать» или «ответить» писатель намеренно использует ка-

кой-либо эмоционально окрашенное, экспрессивное определение (напр. «сказала Дуня, вспыхнув) [Хан-Пира 1999: 18]. Новаторством называют, кроме того, сочетание глаголов речи и глаголов движения, что создает «языковую метафору в авторском метанимическом употреблении» [Маркевич 2016: 2]. Для Достоевского очень характерно использование так называемых «потенциальных глаголов речи», которые семантически не обозначают процесс речи, будучи взятыми вне контекста, но в контексте художественного произведения выполняют функцию ввода прямой речи [Там же].

Учитывая все вышесказанное, можно сделать предположение, что авторские ремарки, сопровождающие речь каждого, даже второстепенного, из персонажей исследуемого романа, отражают их личности и характерные черты, связанные с их эмоциональным фоном, психотипом или профессиональной деятельностью. Иначе говоря, ремарки являются частью языковой личности героя, но оказываются вынесены Достоевским в пласт авторской речи. Выглядит подобное взаимопроникновение микрожанров абсолютно гармонично, так как ранее в работе уже было установлено, что автор имплицитно проникает в жизнь всех своих персонажей.

Наибольший корпус ремарок представлен у Родиона Раскольникова, так как в его словаре присутствуют и леммы, служащие для введения реплик его внутренней речи, тоже. Поэтому возьмем именно главного героя для сопоставления его частотного словаря в микрожанре авторской ремарки с частотными словарями всех остальных персонажей, чей объем ремарок в тексте оказался достаточным для составления показательного ранжированного списка. Рассмотрим сразу все леммы одного микрожанра для попытки выявления сходств. Для этого были попарно взяты частотный словарь ремарок Раскольникова и словари каждого из персонажей, затем были отобраны и посчитаны совпадающие в обоих списках леммы. Наконец, вычислена процентная составляющая лексикона каждого героя от ремарок Родиона Раскольникова. Результаты можно увидеть на графике (см. рис. 3).

При рассмотрении шкал схожести заметна однозначная тенденция большого количества совпадений в авторском лексиконе, сопровождающем речь Раскольникова и Пульхерии Александровны. Такой высокий показатель (62%) может служить вероятным доказательством того, что выводы о связи сына и матери, сделанные в других работах, являются не случайными. Накамура пишет, что, читая «Преступление и наказание» с точки зрения отношений матери и

Рис. 3. Доля схожести словарей Раскольникова и других героев в микрожанре «ремарки»

сына, мы можем увидеть роман о захватывающей дух любви между ними: «Они не говорят друг другу того, что хотят в действительности сказать, за словами друг друга они чувствуют то, что не высказано – то, о чем они не хотят говорить» [Накамура 2011: 176]. Вычисленное соотношение математически подтверждает, что между Раскольниковым и его матерью есть особая связь, они находятся на одной душевной волне. Ремарки, будучи вынесенными за вербализованный акт речи, как раз обнажают те чувства двух героев, которые они не решаются или не хотят высказать друг другу вслух. Кроме того, читатель по ходу повествования наверняка замечает сходства образов Дуни Раскольниковой, сестры главного героя, и Сони Мармеладовой: обе девушки готовы пожертвовать собой, своей честью, чтобы помочь близким. Проще говоря, обе героини иллюстрируют смиренную жертву жестокой гнетущей реальности своего времени. Показатели схожести их ремарок с Раскольниковым сопоставимы (21% и 28% соответственно), а значит, лексиконы для введения речи двух героинь практически идентичны, что демонстрирует идентичность их характеров также и на материале их языковых личностей.

Следующим этапом сопоставления ремарок стал подсчет количества попарно совпадающих лемм, начиная с верхнего ранга словаря главного героя и заканчивая первой по рангу леммой из словаря второго героя, которая не совпала с лексиконом из списка Раскольникова. Результат можно видеть на рисунке 4:

Рис. 4. Количество у героев лемм, присутствующих с тем же рангом в словаре ремарок Раскольникова

Наибольшее количество совпадений с ремарками Раскольникова отмечается у Свидригайлова и Порфирия Петровича (35 и 41 соответственно). В случае с первым из них, как уже было упомянуто ранее, главный герой сам признает это сходство. «Мне все кажется, что в вас есть что-то к моему подходящее» – произносит во время одного из диалогов Свидригайлов [Достоевский 2014: 187]. Второй же, вероятно, будучи следователем по делу Раскольникова, проводит с ним три психологических поединка, говоря, что тот «психологически не убежит». Логично предположить, что именно из-за психологического воздействия на преступника Порфирий Петрович установил с ним контакт. Из-за этого и ремарки, сопровождающие их реплики, имеют столько соприкосновений. Также данный график точнее первого иллюстрирует сходство словаря ремарок Дуни и Сони: здесь это практически одно и то же число (23 и 24 соответственно).

Такие высокие показатели для речи Свидригайлова и Порфирия Петровича, несомненно, отличаются от тех, что были зарегистрированы в случае Пульхерии Александровны (цифры при последнем подсчете относительно низкие). Объяснение этому может быть предложено следующее: при первом анализе исследовались полные словари ремарок вне зависимости от их ранга, т. е. частоты

использования. Совпадения матери и Раскольникова были найдены даже на низких рангах частотности, что свидетельствует об их *общей* эмоциональной схожести. Учитывая то, что диалогов между матерью и сыном происходило не много, большое количество совпадений свидетельствует о *перманентной* связи. Анализируя же совпадения лемм с высокими уровнями частотности, мы выявили аналогии ремарок Раскольникова с теми персонажами, с которыми он не имел долгосрочных эмоциональных связей и вынужден был устанавливать их прямо во время диалогов, которые, стоит отметить, в случаях с обоими героями были весьма объемными. Таким образом, здесь рассматриваются более *спонтанные* эмоциональные проявления Раскольникова и его партнеров, происходившие исключительно в *контексте* конкретных эпизодов. Наконец, обратимся к таблице 5 для анализа и интерпретации первых несовпадающих с частотным словарем Раскольникова ремарок каждого из героев:

Таблица 1. **Первые несовпадающие леммы**

Герой	Первая несовпадающая лемма
Катерина Ивановна	*кричать*
Лужин	*несколько*
Пульхерия Александровна	*бедный*
Настасья	*смех*
Порфирий Петрович	*весело*
Разумихин	*зареветь*
Свидригайлов	*захохотать*
Соня	*вырываться*
Дуня	*Дунечка*
Лебезятников	*подтверждать*
Мармеладов	*беспокойство*
Зосимов	*больной*
Илья Петрович	*бумага (6), поручик(15)*

Таким образом, посредством авторских ремарок персонажи романа получили следующие авторские характеристики:

1. Катерина Ивановна: «кричать». Среди лемм, имеющих отношение к репликам Раскольникова, встречаются глаголы «закричать» и «вскричать» – глаголы совершенного вида, указывающие на завершенность действия. Для Катерины Ивановны автор намеренно использует несовершенный вид глагола, чтобы показать, что женщина кричит постоянно. Это иллюстрирует ее неуравновешенное психическое состояние.

2. Лужин: «несколько». Ключевой чертой характера Лужин является лицемерие. Он постоянно уклоняется от прямых ответов на неудобные вопросы, пытаясь смягчить этим словом ситуацию: по словам матери Раскольникова, он «*несколько* как бы тщеславен и очень любит, чтоб его слушали» [Достоевский 2014: 102]. Также стоит вспомнить его теорию о кафтане, чтобы объяснить высокую частоту употребления слова со значением «в определенной мере».

3. Пульхерия Александровна: «бедный». Это авторское определение точно описывает душевное состояние матери главного героя, которая каждый раз невероятно волнуется за состояние «Роди»: она непрерывно мучается, жалея своих детей и пытаясь защитить их, забывая при этом о себе.

4. Настасья: «смех». Служанка в доме, где снимает комнату Раскольников, отличается жизнелюбием и энергичностью, несмотря на свое, такое же, как у окружающих, бедное положение. Т.А. Касаткина даже соотносит ее образ с былинной героиней-богатыркой [Касаткина 1996: 211].

5. Порфирий Петрович: «весело». Эта лемма отражает неколебимую натуру следователя, одним из психологических приемов которого становится сохранение непринужденности и веселого расположения духа при обсуждении с Раскольниковым убийства старухи. Кроме того, при первом упоминании герой предстает перед читателем приятным и гостеприимным человеком.

6. Разумихин: «заревeть». Приятель главного героя, являясь его антиподом, отличается взрывным характером и решительностью действий. Данная лемма наиболее ярко подчеркивает контраст между друзьями: Раскольников чаще «бормочет».

7. Свидригайлов: «захохотать». Эта лемма имеет очевидную стилистическую окраску и в данном случае усиливает зловещий и пугающий образ аморального персонажа.

8. Соня: «вырываться». Героиня отличается невероятной робостью и каждая реплика дается ей с большим трудом. Тем не менее,

если Соня говорит, то настолько искренне и чисто, что слова «вырываются» из нее сбивчиво, одновременно с мыслями.

9. Дуня: «Дунечка». В уменьшительно-ласкательной форме Авдотью Романовну называют только автор и ее мать. Первый, таким образом, компенсирует нежность, которую не решается проявить Раскольников к сестре, но которую на самом деле испытывает.

10. Лебезятников: «подтверждать». Такая лемма, часто встречающаяся в ремарках, заставляет этот образ остаться незаметными для читателя, так как герой только соглашается со всеми вокруг, не выражая собственного мнения. Тем сильнее оказывается эффект неожиданности, произведенный речью Лебезятникова в эпизоде жестокого выселения семьи Мармеладовых из их квартиры.

11. Мармеладов: «беспокойство». Через это слово автор подчеркивает жалкое положение и бессилие героя. Отец семейства без перерыва ощущает беспокойство и свою вину, но не предпринимает никаких действий.

12. Зосимов: «больной». В большинстве случаев автор указывает, что врач обращается к Раскольникову именно как к «больному». Зосимов был одним из тех людей, кто представлял сторону верящих в умственное помешательство главного героя, как единственную причину его плохого самочувствия.

13. Илья Петрович: «бумага», «поручик». Решено было выделить две леммы (с рангом 6 и 15 соответственно), так как они одинаково ярко иллюстрируют узкость кругозора и жизненных приоритетов героя: он зациклен на бумах и дорожит своей должностью, постоянно угождая начальству – на этом построено все его существование.

Таким образом, в данном разделе удалось доказать, что ремарки как отдельный микрожанр художественного произведения являются еще одним средством раскрытия языковой личности героев через обозначение автором их эмоциональных состояний. Метод анализа частотных словарей показал себя полезным для сопоставления сразу нескольких списков с последующим получением релевантных для интерпретации результатов. Наконец, стало ясно, что уже одна лемма в списке ремарок, относящихся к определенному герою, способна стать материалом для расшифровки целого образа и индивидуального отличительного характера. Этими словами, что не маловажно для инновационного стиля Достоевского, явились не только глаголы речи, но и другие части речи, содержащие эмоциональный компонент.

Подводя итоги анализу, проведенному над текстом романа Ф.М. Достоевского «Преступление и наказание», можно сделать несколько выводов:

– деление текста авторской речи на речевые микрожанры дает хорошую базу для выявления стилистических и лексических паттернов не только в образе автора, но и образах главного и второстепенных персонажей;

– несмотря на вышеупомянутое разделение, частотные словари героев образуют систему, каждый из элементов которой способен влиять на другой: были выявлены параллели между словоупотреблениями в речи Родиона Раскольникова и его матери, Пульхерии Александровны, между главным героем романа и другими женскими образами: Соней и Дуней и, наконец, между ним же и его идейными двойником и противником – Свидригайловым и Порфирием Петровичем;

– метод составления частотных словарей является эффективным инструментом для выявления характерных черт индивидуального авторского языка, что, в свою очередь, при работе с Достоевским, после становится предметом философских размышлений и анализа эмоциональных и психологических связей внутри романа;

– динамический частотный словарь, не будучи на данный момент активно используемым в квантитативных исследованиях методом, имеет право на статус перспективного инструмента для дальнейших исследований в этой области, поскольку по ходу работы с его помощью удалось установить некоторые объясняемые закономерности изменчивости частот в романе на примере отдельно взятых лемм.

Использованные методы должны быть развиты в последующих исследованиях в области лексикографии и смежных с ней наук, потому как тексты авторов, обладающие глубоком авторским психологизмом, не ограничиваются Достоевским и все еще требуют подробной «расшифровки». Что касается романа «Преступление и наказание», – в нем остались незатронутые микрожанры, которые тоже могут быть логично выделены: например, сны Раскольникова, которым посвящено множество научных статей. Тем не менее, метод динамического частотного словаря к ним применен еще не был. Таким образом, данная работа является своеобразной иллюстрацией того, насколько много подходов может быть применено к художественному тексту для нахождения в нем до сих пор не обнаруженных черт.

ЛИТЕРАТУРА

1. Алексеев, П. М. Частотные словари: учеб. пособие – СПб, СПбГУ, 2001. С. 6–10.

2. Батурина Е. Н. Особенности повествовательной формы в романе Достоевского «Преступление и наказание» // Историческая поэтика жанра. Биробиджан: Изд-во ДВГСГА, 2006. С. 37.

3. Бахтин М.М. Проблемы поэтики Достоевского. М., 1979. С. 59, 204.

4. Габдуллина В. И. Проблема авторского дискурса в художественной системе Ф. М. Достоевского // Вестн. НГУ. История, филология. Т. 6. Вып. 2. Новосибирск, 2007. С. 93.

5. Достоевский Ф.М. Преступление и наказание. М.: Эксмо, 2014. С. 67–307.

6. Касаткина Т.А. Авторская позиция в произведениях Достоевского «Вопросы литературы», №1, 2008. С. 3, 30.

7. Касаткина Т.А Характерология Достоевского. Типология эмоционально-ценностных ориентаций. – М., Наследие, 1996. С. 211.

8. Кирпотин В.Я. Разочарование и крушение Родиона Раскольникова: Москва.: Советский писатель, 1970. С. 3–7.

9. Маркевич Ю.В., Середа П.В. Глаголы речи и их семантическая классификация научные труды, КУБГТУ, №2, 2016. С. 2.

10. Мартыненко Г.Я. Основы стилеметрии – Л.: Изд-во Ленингр. ун-та, 1988. С. 106.

11. Накамура К. Словарь персонажей произведений Ф.М. Достоевского. – СПб.: Гиперион, 2011. С. 176–181.

12. Пашкявичене Л. Эмоциональная интонация в художественном тексте (на материале романа Ф.М. Достоевского «Преступление и наказание»): магистерская работа, Вильнюс, 2009. С. 42.

13. Ружицкий И.В. Языковая личность Ф.М. Достоевского: лексикографическое представление: дис. … д-ра филол. наук: 10.02.19. Москва, 2015. С. 19–24.

14. Хан-Пира Э. О творческом характере авторских ремарок Ф. М. Достоевского, «Русская речь» сер. Язык художественной литературы. Точка зрения, №1, 1999. С. 18.

15. Шайкевич В.Я. Статистический словарь языка Достоевского. – М.: Изд-во: Языки славянской культуры, 2005. С. 23.

УДК 811.161+811.512.111]:130.3

THE TOP-LEVEL ONTOLOGY FOR THE AUTOMATIC PROCESSING OF TEXT (ON THE MATERIAL OF LANGUAGES OF DIFFERENT STRUCTURES)

A. R. Gubanov, V. P. Zheltov, A. M. Ivanov, G. F. Gubanova, E. A. Kojemiakov
Chuvash state University, Cheboksary
alexgubm@gmail.com, chnk@mail.ru, amivano@rambler.ru, rggalina@gmail.com, ekozxemyakova@yandex.ru

The problem of constructing top-level ontologies for automatic text processing based on semantic analysis of natural language texts is discussed. The process of building ontology consists of several consecutive steps: a) marking of the corpus syntactic tags; b) definition characteristic, i.e. the most frequently occurring in the case of linguistic structures and the formation on their basis of linguistical studies of the patterns; C) search of text fragments, respective data patterns. The proposed approach of constructing a top-level ontology combines the speed of statistical methods and the accuracy of the linguistic approach from the standpoint of reference literals, syntaxes (or linguistic patterns). The considered models provide users (experts, linguists) with materials for automatic text processing, and knowledge engineers with tools for designing automatic text processing systems and conceptual data representation schemes.

Keywords: ontology, ontology, top-level, automatic text processing, expert, knowledge engineer, artificial intelligence, multi-structured languages, Russian language, Turkic languages, Chuvash language, semantic text analysis, natural language, linguistic patterns, explanatory texts, complex sentences structures, supporting literals, syntaxeme, conceptual schema.

ОНТОЛОГИЯ ВЕРХНЕГО УРОВНЯ ДЛЯ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ТЕКСТА (НА МАТЕРИАЛЕ РАЗНОСТРУКТУРНЫХ ЯЗЫКОВ)

А. Р. Губанов, В. П. Желтов, А. М. Иванова, Г. Ф. Губанова, Е. А. Кожемякова
Чувашский государственный университет, Чебоксары
alexgubm@gmail.com, chnk@mail/ru, amivano@rambler.ru, rggalina@gmail.com, ekozxemyakova@yandex.ru

Обсуждается проблема построения онтологий верхнего уровня для автоматической обработки текста на основе семантического анализа текстов на

естественном языке. Процесс построения онтологии состоит из нескольких последовательных этапов: а) разметка корпуса синтаксическими тегами; б) определение характерных, т. е. наиболее часто встречающихся в корпусе лингвистических структур и формирование на их основе лингвистических шаблонов; в) поиск фрагментов текста, соответствующих данным шаблонам. Предлагаемый подход построения онтологии верхнего уровня сочетает в себе быстроту статистических методов и точность лингвистического подхода с позиций опорных литералов, синтаксем (или лингвистических шаблонов). Рассмотренные модели предоставляют пользователям (экспертам, лингвистам) материалы для автоматической обработки текста, а инженерам знаний – инструменты для проектирования систем автоматической обработки текста и концептуальные схемы представления данных.

Ключевые слова: онтология, онтология верхнего уровня, автоматическая обработка текста, эксперт, инженер знаний, искусственный интеллект, разноструктурные языки, русский язык, тюркские языки, чувашский язык, семантический анализ текстов, естественный язык, лингвистические шаблоны, каузальные тексты, полипредикативные конструкции, опорные литералы, синтаксема, концептуальные схемы.

В языке наработано достаточно много способов выражения общезначимых логических утверждений на основе причинно-следственных отношений. Интерес к этому направлению исследования обусловлен бурным развитием искусственного интеллекта. Эта тема становится особенно актуальной также с появлением экспертных систем, в которых центральное место занимает вопрос автоматической обработки текста. Для того, чтобы сделать автоматическую обработку текстов более качественной и надежной, необходимо использовать, как известно, знания и о языке, и об окружающем мире. Знания о мире могут быть представлены с помощью онтологий – систем понятий, для которых описаны отношения, в частности, отношения обусловленности, выступающие как основные текстообразующие единицы, отражающие предмет реального или идеального мира и хранящиеся в национальной памяти носителей национальных (в данном случае русского и чувашского) языков как вербальный субстрат [3,4,6,7,8,9,10]. Онтолингвистические системы ориентированы на решение сложных задач обработки текстов, требующих семантических знаний [11].

Представленный в работе подход к моделированию процессов автоматического анализа текста опирается на знания о причинно-следственных отношениях анализируемого текста, явно формализованные в виде онтологии, что позволяет применять методы локаль-

ного семантического и синтаксического анализов, не требуя наличия полного корректного синтаксического разбора и грамматически правильно построенного текста. Использование информационного контента онтологии при идентификации и сравнении объектов, найденных в тексте, позволяет использовать неявные знания, т. е. информацию, не содержащуюся в тексте.

Известны онтологии двух типов – онтологии верхнего уровня и предметные онтологии. Онтологии верхнего уровня (Дж. Совы, СУМО, Микрокосмос и др.) содержат наиболее общие и самые абстрактные фундаментальные концепты, как сущности, время, пространство, причинность и т. д. Их назначение – содействовать улучшению автоматической обработки естественного языка, извлечения и поиска информации. Чтобы применить онтологию верхнего уровня – обусловленность для автоматической обработки текстов – необходимо сопоставить понятие онтологии с набором языковых выражений отношений обусловленности (слов и словосочетаний), которыми понятия могут выражаться в тексте. Одним из эффективных соответствующих методов является использование лингвистических шаблонов [1], представляющих собой характерные выражения (словосочетания и обороты), конструкции из определенных элементов языка (такие шаблоны позволяют построить семантическую модель, соответствующую тексту, к которому они применяются). Исследования в этой области показали, что при использовании шаблонов на большом корпусе текстов одной тематики можно построить «достаточно адекватную» таксономию понятий соответствующей предметной области, элементами шаблонов для наиболее точного описания могут быть литералы, т. е. конкретные лексемы; определенные части речи; грамматические конструкции; условия, уточняющие грамматические характеристики рассмотренных элементов.

Одной из наиболее актуальных проблем автоматического анализа текста является особенность подхода к анализу опорных лексем, синтаксем – лингвистического базиса системы поверхностно-синтаксического анализа, минимальных семантико-синтаксических единиц русского языка, выступающих одновременно и как носитель элементарного смысла, и как конструктивный компонент с функциональностью. «Синтаксический строй речи (текста), – отмечает Золотова Г. А., – организуется регулярными комбинациями «элементарных» единиц-синтаксем, из которых строятся все другие более сложные (речевые или текстовые) конструкции [6]. Синтаксема

обладает способностью реализовываться в коммуникативных единицах, и функциональные свойства синтаксем имеют транзитивный характер.

Изучая способы формального описания синтаксической структуры предложения (такое описание необходимо для решения задач, связанных с автоматической переработкой естественно-языковой информации), подробно излагая «традиционные» способы, основанные на использовании деревьев синтаксического подчинения и систем составляющих, другой известный специалист по теоретической и прикладной лингвистике Гладкий А.В. вводит новый способ описания синтаксических структур естественного языка в автоматизированных системах общения, основанный на использовании систем синтаксических групп, являющихся одновременным обобщением систем составляющих и деревьев синтаксического подчинения [2]. Этот способ, по мнению автора, позволяет добиться большей гибкости и естественности описания.

Рассматривая синтаксемы в тексте как динамическую систему, мы имеем возможность описать весь процесс построения онтологии концепта обусловленности как целенаправленную операционную деятельность, организованную для решения задач содержательного наполнения элементов онтологии. Предлагаемый подход построения онтологии верхнего уровня сочетает в себе быстроту статистических методов и точность лингвистического подхода с позиций опорных литералов, синтаксем (или лингвистических шаблонов).

Процесс построения онтологии состоит из нескольких последовательных этапов: а) разметка корпуса синтаксическими тегами; б) определение характерных, т. е. наиболее часто встречающихся в корпусе лингвистических структур и формирование на их основе лингвистических шаблонов; в) поиск фрагментов текста, соответствующих данным шаблонам.

Рассматриваемая лингвистическая модель знаний (концептуальная схема-модель) включает три компонента: словарь литераловлексем концепта обусловленности, который задает лексическую модель подъязыка, и модели фактов-синтаксем, связывающие семантико-синтаксические модели, описывающие структуру выражений, принятых в данной области для описания информации, с формальным представлением этой информации, определяемым рассматриваемой онтологией и жанровая модель текста формирует жанровую структуру рассматриваемого текстового источника, сужая при этом область поиска определённой информации.

Рассмотрим концептуальную схему литералов-лексем. Следует отметить, что процедура сопоставления понятий онтологий и языковых выражений может быть осуществлена различными способами. Одним из способов является способ сбалансирования обеих частей: система понятий и система лексических значений разделены в разумном соотношении в создаваемом ресурсе и описаны их взаимосвязи, т. е. онтологии концепта обусловленности должны иметь свое непосредственное отражение в языковых значениях. Каждое понятие имеет набор синонимов, а их особенности описываются в дополнительных, внутрипонятийных структурах. В рамках предлагаемого подхода в качестве лексической модели выступают таксоны, отнесённые к определённому семантическому классу (иерархия классов позволяет отнести тот или иной таксон к её определённому уровню, более общему или конкретному с наследованием свойств общего класса (эти единицы также могут иметь необходимые данные для описания валентностной структуры предикатных слов. Концептуальное моделирование отношений обусловленности представлено в системе Java, которая поддерживает два основных способа моделирования онтологий посредством редактора KOMMA-Frames и KOMMA-OWL (Рис 1.):

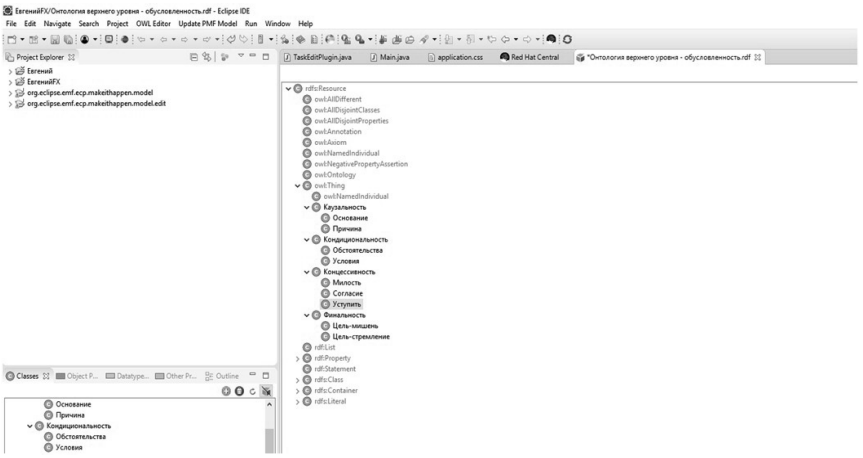


Рис 1. Концептуальное моделирование отношений обусловленности

Рассматриваемые подклассы, как видно из рис.1, могут быть представлены фреймами, в частности, подкласс «каузальность»

имеет два фрейма: 1) таксоны «причина явления»; 2) таксоны «основание», которые в свою очередь составляют несколько семантических сетей и структурируются поверхностно-синтаксическими фреймами. Каждый концепт снабжается языковыми выражениями, значения которых соответствуют определённому понятию. И такие выражения являются между собой онтологическими синонимами, в частности, следующие концепты имеют такие синонимы:

1) каузальность – причинно-следственная связь, обуславливать, способствовать, первопричина, тайна (скрытая причина), беспричинный, беспричинность, мотив, основание, повод для действий. предлог для действий, обстоятельство, фактор, мотивированный, обоснованный, мотивировать, обусловливать (обусловить), обусловливающий (в чув.яз. – *сăлтав çыхăнăвĕ, сăлтавлăх, малтанхи сăлтав, вăрттăн сăлтав, сăлтавсăр* и др.;

2) финальность – предмет стремления, задача, самоцель, достичь, добиться, ставить перед собой цель, целесообразный, целеустремленность, физический объект, мишень для стрельбы, навести на цель, попадать в цель (в чув.яз. – *тĕллев, задача, тĕллев лартмалла, тĕллеве пурнăçа кĕрт* и др.);

3) кондициональность – обстоятельство, фактор, гарантия осуществления, необходимое условие, предпосылка для реализации, предрасположенность, обстоятельство, положение дел, ситуационный, ситуация, безвыходное положение, вопрос, проблема, кризис (в чув.яз. – *услови, фактор* и др.); 4) концессивность – согласие, согласиться, дать согласие, компромисс, уступить, поступиться принципами, снисхождение, снисходительный, поддаваться. податливый, дрогнуть, упустить [отдать] инициативу, отступать (перед трудностями), отступиться, отступительный, отступной, сдаться, одолжение (делать), подачка. Снисхождение (в чув.яз. – *килĕшӳ, килĕш* и др.).

Для различных приложений автоматической обработки текста применяются группировки отношений между концептами (такие отношения, отношения онтологической зависимости чаще всего применяются в онтологиях верхнего уровня):

I. Отношения «вверх – вниз»:

1) каузальность: а) причина (явления): причинно-следственная связь – вина (причина неблагоприятного – первопричина (основная причина) – стимул – толчок – тайна (скрытая причина); б) основание: обстоятельство – фактор – мотив – повод для действий – предлог для действий – обоснованный – обусловливать (обусловить);

2) финальность: а) цель (предмет стремления): задача – самоцель и др.; б) цель (для попадания) – физический объект – мишень для стрельбы;

3) кондициональность: а) условие (исполнения): фактор – гарантия осуществления – необходимое условие – предпосылка для реализации, развития – склонность, предрасположенность; б) обстоятельства: обстановка – положение дел – ситуация – адские условия – безвыходное положение – вопрос – проблема – кризис;

4) концессивность: а) согласие: согласиться – дать согласие – компромисс – уступить – поступиться принципами – снисхождение; б) уступить подешевле – давать скидку, дать скидку, уступать, уступать в цене, уступить, уступить в цене, уступить дешевле, уступить подешевле; в) уступать в борьбе: поддаваться – податливый – дрогнуть (неприятель дрогнул) – упустить [отдать] инициативу – отступать (перед трудностями) – сдаться (уступить); в) милость: одолжение (делать) – подачка – снисхождение;

II. Ассоциативные отношения:

1) каузальность: беспричинный, неубедительный, мотивированный, обоснованный, мотивировать (в чув. яз. – *сăлтавсăр, ĕненмелле мар* и др.);

2) финальность: ставить перед собой цель, целесообразный, целеустремленность, навести на цель, попадать в цель (в чув. яз. – *тĕллев ларт, кирлĕ, тĕле лек*);

3) концессивность: скидка.

Введение вышеназванных концептуальных путей используется в процедурах автоматического разрешения лексической неоднозначности, вывода рубрик по тексту и т. д.

Следующим этапом построения онтологии выступает процесс определения характерных, т. е. наиболее часто встречающихся в корпусе лингвистических структур и формирование на их основе лингвистических шаблонов.

Лексико-синтаксическими шаблонами, маркирующими, в частности, каузальную ситуативную информацию, выступают:

1. Таксисные маркеры: а) гипотактические маркеры-релятивы *потому что (мĕншĕн тесен)*; *и оттого (çавăнпа та),* вследствие того, по этой причине, благодаря этому (*çавна пула та*); б) паратактические – скрепы-релятивы *и, а, но: та (те), тата, анчах, çапах, е, те-те* (особенности конструкций: тема-рематическая связь между конъюнктами (отсутствуют тема-рематические пересечения; семантический тип предиката);

2. Маркеры полипредикативных конструкций (ППК):

1) полипредикативные конструкции синтетического типа (в роли зависимого предиката выступает причастие, которое способно принимать различные падежные аффиксы): модели ППК синтетического типа с управляемыми ПЕ: а) Pr зависим. N исходн.; б) Pr зависим. N каузалис; в) Pr зависим. N творит.; модели ППК синтетического типа с неуправляемыми ПЕ: а) Pr зависим на *-са/-се*: б) Pr зависим. на *-сан/-сен*: моносубъектная реализация (действие, выраженное зависимым предикатом, относится к тому же субъекту, что и финитное); разносубъектная реализация (в чувашском языке, как и в других тюркских языках, каузальная ситуативная информация оформляется полипредикативными конструкциями, где в зависимой предикативной части в роли предиката (Pr зависим.) могут выступать инфинитные формы глагола);

2) гипотактические конструкции аналитико-синтетического (гибридного) типа – причастно-послеложные конструкции, где причастие прошедшего времени на *-нӑ/-нӗ* сочетается с послелогами и служебными именами (в аналитико-синтетических ППК способ связи получает формальное выражение в аналитических показателях (послелогах и частицах): модели с неуправляемым зависимым предикатом: а) Pr завис.T_v *-нӑ/-нӗ* + послелог *пирки* – моносубъектные/ разносубъектные конструкции): б) модель с Pr завис. T_v -нӑ/-нӗ + *енне* (разносубъектные); в) Pr завис. T_v *-нӑ/-нӗ* +_май (майӗпе) – моносубъектные; разносубъектные; модели с управляемыми причастно-послеложными формами Pr: а) Pr завис. T_v -нӑ/-нӗ *пула* – моносубъектные; разносубъектные; б) Pr завис. T_v -нӑ/-нӗ кура;

3) нетаксисные маркеры (предикативный план представлен имплицитно: событийное содержание причинного смысла предикативно не оформлено, модальные и временные характеристики свернуты): а) субстантивные синтаксемы-каузативы: активные (центральные) каузативы *из+N, от+N, с+N, из-за+N, по+N, за+N; по причине+N, вследствие+N, в результате+N, в силу+N* (по своей этимологической определенности четко указывают на причинный смысл); б) адвербальные компоненты *спьяна, сдуру* и т. д.

Аналогичную функцию в чувашском языке выполняют:

1) субстантивные маркеры синтаксемы: а) аффиксальные синтаксемы: а) каузема N1; б) каузема N6 – -шӑн/-шӗн (причинно-целевой падеж (специальный падеж для выражения причинного смысла)); в) каузема N7 -па: б) послеложные синтаксемы с причинными послелогами пула, пирки;

2) адвербальные комопненты – застывшие формы синтаксем: а) творительного падежа местоимений, принявших аффикс принадлежности 3-го лица *çавăнпа*: б) причинно-целевого падежа местоимений: *«мĕншĕн», «çавăншăн»*: в) исходного падежа местоимений: *мĕнрен, çакăнтан*; г) основного падежа + послелог *пирки*: *мĕн пирки*; д) дат.-винительного падежа + послелог *пула (кура)*: *çавна пула, та(те), тата, анчах, çапах, е, те-те и др.*

Маркеры в инкаузальных (концессивных) конструктах зависят от модификации доминирующей доминантной семы: 1) без модификации доминантной семы: в русских центральным маркером выступают: таксисные релятивы *хотя (хоть), несмотря на то что,* а в чувашском языке эквивалентными маркерами являются: а) релятивловная скрепа *пулсан* + частица *та* (компонент *пулсан*, стоящий после придаточного предложения, предполагает наличие предложения, выражающего следствие, и частица *та* лишь предупреждает, что это следствие носит адверсативный характер); б) концессивная скрепа *пулин* (совпадает по звучанию с формой 3-го лица ед. ч. положительного аспекта уступительного наклонения глагола *пул*); в) частицы *ĕнтĕ, та*; г) аффикс *ех;* синтаксемы – зависимые пропозитивные компоненты: 1) синтаксемы *несмотря на+N, невзирая на+N, вопреки+N, против+N, наперекор+N* подчеркивают противоречие, противоположность события-уступки и события-следствия; 2) синтаксемы *независимо от+N, помимо+N.,* а в чувашском языке в качестве ситуативных типологических форм выступают маркер-зависимый пропозитивный компонент, формируемый при помощи глагольного компонента *пăх* («смотреть») и дат.п.; 2) с модификацией доминантной семы: а) *согласие* – маркером модифицированной доминантной семы выступает *пусть* (в чув. языке данный оттенок согласия выражается риторической формулой *ан тив*); б) допущение – форма маркера релятив *хотя/если* +частицей *и/ бы*, а в чувашском языке маркеры – *пулин те, пулсан та*.

Рассматриваемая ситуация может быть репрезентирована также маркером – зависимым таксисом – специальной конструкцией: деепричастным оборотом или одиночным деепричастием (в сопоставляемых языках мы наблюдаем аналогичные маркеры).

Подчеркнем, что часто с маркерами концепта обусловленности в моделях выступают так называемые типизированные лексические маркеры, взаимодействующих с грамматической организацией той или иной конструкции. Например, в концессивных подконцептах частотны следующие компоненты: а) типизированные лексические

элементы *напрасно, тщетно, вотще*, которые, выступая в роли фразеологических средств связи, в то же время сохраняют свою лексическую полнозначность, однако употребление их в функции, близкой к союзной, сопровождается сужением семантики, лексической их специализацией; б) соотносительные адвербальные компоненты: *еще – уже (уже – еще), все еще;* в) устойчивые предложно-падежные местоименные сочетания: *между тем, в то же время, вместе с тем, при всем том*; г) адвербиальные сочетания темпорального характера (*всегда – на этот раз; за полчаса до того, как*) – чувашские лексические элементы аналогичны русским.

Реальные трудности могут возникать в автоматической обработке текста в случаях, когда лексико-синтаксические шаблоны вступают в синонимические отношения. Например, внутри кондициональных нетаксисных структур можно выделить две вариативные синонимические модели: 1) синтаксемы *в случае/ при/ в/ с +N – при условии N* (называющие явления, присутствие которых обусловливает или могло бы обусловить какое-то действие или признак. Ср.: *в случае – при случае необходимости... – при необходимости ... – кирлĕ вăхăтра... – кирлĕ чухне... при условии заключения договора..., – в случае заключения выгодной сделки..., – при выгодных условиях – килĕшу тăвас пулсан..., тупăшлă ĕç пирки калаçса татăлас пулсан, тупăшлă условисенче);* 2) синтаксемы *без+N - при отсутствии/в отсутствии+N* (называющие явления, отсутствие которых вызывает или может вызвать какое-то действие или явление). Еще пример: синонимия синтаксемного шаблона *без+N2* с таксисным маркером *если* и т. д.

Для успешного извлечения информации из текста система должна располагать некоторой текстовой информацией, связанной с системной моделью каузальных (аргументативных) отношений, которая дает в распоряжение исследователя не только представление об уровне логической культуры автора текста.

Предлагаемые решения по автоматической обработке текста также основываются на понятии жанра как совокупности логико-композиционных, формально-лингвистических и лексико-грамматических компонентов (определённая жанровая лексика, определённые формальные сегменты: логико-композиционная структура текста определяется на основе лексикона жанровых маркеров и шаблонов, выделяющих содержательные блоки. При выборе классифицирующего признака обусловленности необходимо учитывать специфику текстов с отношениями обусловленности (компонент-

ный состав структуры, тип каузальной реляции при корне дерева, их позиция в текстовом фрагменте – обязательность их признака и достаточная степень их обобщенности, ибо текстообразующие средства обусловленности неоднородны по своему составу. Например, в интродуктивной позиции («задание темы») текст, выражающий каузальные отношения, реализует свою проспективную синсематическую направленность, обусловливает установление катафорических смысловых связей в тексте, тем самым обеспечивает каузальную смысловую и структурную связность и целостность каузальных текстов: распространяет семантическое влияние на весь дискурс, и самое главное – определяет тональность всего дальнейшего повествования. *Ср.: И потому* в мелькнувшем образе Корделии, в огне страсти Обломова отразилось только одно мгновение, одно эфемерное дыхание любви, одно ее утро, один прихотливый узор. А завтра, завтра блеснет уже другое, может быть, такое же прекрасное, но все-таки другое (Гончаров). – *Çавăнпа та* ĕнтĕ Обломов пуçĕнче пĕр самантлăха çеç Корделин сăнарĕ, унăн нумайлăха пыман юратăвĕ, унпа пĕрле пулнă пĕр ир мĕлтлетсе илчĕ. Ыран, ыран вара урăххи пулать, тен, çакăн пек хитриех, анчах та урăххи. Занимая же медиальную позицию в текстовом фрагменте и объединяя контактно и дистантно расположенные предложения, каузальные текстовые фрагменты способствуют реализации преемственности смысла между предложениями текстового комплекса, т. е. выступают в текстообразующей роли. Данные текстовые комплексы с каузальными отношениями обладают как проспективной, так и ретроспективной семантической направленностью: Более ничто не напоминало старику барского широкого и покойного быта в глуши деревни. Старые господа умерли, фамильные портреты остались дома и, чай, валяются где-нибудь на чердаке; предания о старинном быте и важности фамилии все глохнут или живут только в памяти немногих, оставшихся в деревне же стариков. *Поэтому* для Захара дорог был серый сюртук: в нем да еще в кое-каких признаках, сохранившихся в лице и манерах барина, напоминавших его родителей, и в его капризах, на которые хотя он и ворчал, и про себя и вслух, но которые между тем уважал внутренне, как проявление барской воли, господского права, видел он слабые намеки на отжившие величия (Гончаров). – Тĕттĕм ялти улпутăн пуян та канлĕ пурнăçĕ çинчен старике урăх пĕр япала та аса илтермест. Ват улпутсем вилнĕ, вĕсен сăн ӳкерчĕкĕсем килте юлнă та халĕ таçта мачча çинче йăваланса выртаççĕ пулĕ; ĕлĕкхи пурнăçпа чаплă хушаматсем çинчен калакан

халапсем манăçа тухсах пыраççĕ, е ялта пурăнакан ватăсен асĕнче çеç упранаçĕ. *Çавăнпа та* Захаршăн кăвак сюртук хаклă: çав сюртук тата улпутăн сăнĕнчи, хăтланкаларăшĕнчи хăйĕн ашшĕ-амăшне аса илтерекен хăш-пĕр паллăсем Захара иртнĕ чаплă пурнăç мĕлки çинчен калаççĕ; улпут курнăçланнăшăн Захар кăмăлсăр мăкăртататать *пулин те*, çав хушăрах çав курнăçланăва улпут ирĕкĕ вырăнне хурса хисеплет, ĕлĕкхи пурнăçăн мăнаçлăхне систернине курать. А занимая финальную позицию, по линии тематического развития и возвращая к содержательно значимой информации пропозитивной части, обусловливают поступление обобщающе-итоговой информации в содержании текста. Соответствующие комплексы имеют ретроспективную синсемантическую направленность: Он довольно остер: эпиграммы его часто забавны, но никогда не бывают метки и злы: он никого не убьет одним словом; он не знает людей и их слабых струн, *потому что* занимался целую жизнь одним собою. Его цель – сделаться героем романа. То так часто старался уверить других в том, что он существо, не созданное для мира, обреченное каким-то тайным страданиям, что он сам почти в этом уверился. *Оттого-то* он так гордо носит свою толстую солдатскую шинель (Лермонтов). – Вăл çивĕч чĕлхеллĕ çын: унăн эпиграммисем час-часах кулăшла, анчах нихăçан та кирлĕ çĕре лекеймеççĕ, усал та мар вĕсем: вăл никама та пĕр сăмахпа çапса хуçаймасть; çынсене те, вĕсен айван енĕсене те пĕлеймест вăл, *мĕншĕн тесен* хăй ĕмĕрĕнче хăй çинчен кăна шухăшлать. Унăн пĕтĕм ĕмĕчĕ те роман геройĕ пуласси. Çынсене вăл вара час-часах ĕнентерме тăрăшатчĕ: эпĕ тĕнчере пурăнма çуралман, эпĕ темĕскерле, никам чухласа илейми асаппа пурăнма кăна çуралнă, тетчĕ; ку чăнах та çакăн пек иккенне хăй те ĕненес патнех çитетчĕ вара. *Çавăнпа ĕнтĕ* вăл хăйĕн хулăн салтак шинельне çав тери мăнаçланса тăхăнса çӳретчĕ те.

Следует отметить также, что взятый в качестве примера каузальный дискурс в большинстве случаев представляет макроструктуру мозаичного вида. Речевая форма или текстовый модуль “рассуждение” в причинной текстовой мозаике выступает в виде такого доминирующего модуля, как “размышление”, хотя и встречаются и другие виды рассуждения. Приведем примеры:

а) рассуждение-доказательство: Литературу у нас преподавал Лихо – очень глупый человек, которого вся школа называла Лихосел. Он всегда ходил в кубанской шапке, и мы рисовали эту шапку на доске, и в ней проекцией – ослиные уши. Лихо меня не любил (почему?), и вот по каким причинам. Во-первых, он однажды

диктовал что-то и сказал: “Обстрактно”. Я поправил его, мы заспорили, и я предложил запросить Академию Наук. Он обиделся. Во-вторых, большинство ребят составляли свои сочинения из книг и статей – прочтет критику и спишет. А я так не любил. Я сперва писал сочинение, а потом читал критику. Это-то не нравилось Лихо! Он надписывал: “Претензия на оригинальничанье. Слабо!” Он, разумеется, хотел сказать – на оригинальность. Кто не станет претендовать на оригинальничанье? Словом, я боялся, что по литературе у меня в году будет “плохо” (Каверин). – Литературăпа пире çав тери айван çын – Лихо хушаматлăскер – вĕрентетчĕ. Пĕтĕм шкулĕпех ăна Лихосел *тесе* витлетчĕç. Яланах вăл кубанка тăхăнса çӳретчĕ. Пĕррехинче эпир унăн кубанкине доска çине ӳкертĕмĕр. Çĕлĕкĕ çумне ашак хăлхисем те туса лартрăмăр. Лихо мана тĕрлĕ сăлтавсене *пула* юратмастчĕ. Пĕррехинче вăл диктант çыртарнă чух: ”Обстрактно”, – терĕ. Эпĕ унăн йăнăшне тӳрлетсе каларăм, тавлашса кайрăмăр хайхи. Эпĕ ăслăлăхсен Академийĕнчен ыйтса пĕлме сĕнтĕм. Вăл çилленсе кайрĕ. Ачасем ытларахăшĕ тата темшĕн хăйсен сочинений ĕсене кĕнекесемпе пĕр-пĕр статьясем çинчен çыра-çыра илетчĕç. Эпĕ ун пек хăтланма юратмастăм. Эпĕ малтан сочинени çыраттăм: кайран тин кĕнеке вулаттăм: Çакна юратмастчĕ те ĕнтĕ Лихо. Вăл вара: “Çынсенчен уйрăмрах çырма тăрăшатăрха!“ – тетчĕ. Пĕр сăмахпа каласан: литературăпа çулталăкшăн “начар” паллă пуласран хăраттăмччĕ эпĕ;

б) рассуждение-размышление: Она вздрогнула и посмотрела на меня с изумлением. Потом она покраснела и обняла меня. Она меня обняла, и мы поцеловались с закрытыми глазами – по крайней мере я, но кажется, и она тоже, *потому что* потом мы одновременно открыли глаза. Мы целовались в сквере на Триумфальной, в середине Москвы, в этом сквере, где нас могли видеть три школы – наша, 143-я и 28-я. Но это был горький поцелуй. Это был прощальный поцелуй. Хотя, расставаясь, мы условились о новой встрече, я чувствовал, что этот поцелуй – прощальный. Вот почему, когда Катя ушла, я остался в сквере и долго еще бродил по дорожкам в тоске, садился на эту скамейку, уходил и опять возвращался. Я снял кепку – у меня горела голова, и сердце ныло. Я не мог уйти (Каверин). – Вăл шартах сиксе тĕлĕннĕ пек пулчĕ, ман çине пăхса илчĕ. Унтан вăл хĕрелсе кайрĕ, мана ыталаса илчĕ, эпир вара куçа хупсах чуптурăмăр. Эпĕ куçа хупнăччĕ-ха, вăл та хупнăччĕ пулас, *мĕншĕн тесен* эпир иксĕмĕр те куçсене харăсах уçрăмăр. Эпир Триумфальнăй скверĕнче, Мускав варринче чуптурăмăр. Кунта пире виçĕ шкул кантăкĕнчен курма

пултарнă – хамăрăн, 143-мĕш тата 28-мĕш шкулсенчен. Анчах та ку чуптăву юлашки пулчĕ, пĕр-пĕринчен уйрăлмалла чуптуни пулчĕ, уйрăлнă чух эпир татах тĕл пулма сăмах патăмăр *пулин те*, эпĕ ăна хам юлашки хут чуптунине чухласа илтĕм.Çавăн *пиркиех* ĕнтĕ эпĕ, Катя кайсан та, унтах тарса юлтăм, чылайччен çул тăрăх тунсăхласа каллĕ-маллĕ уткаласа çӳрерĕм, иксĕмĕр ларнă сак çине пыра-пыра лартăм, унтан пăрăнса кайрăм, каллех çав вырăнах таврăнтăм. Эпĕ кепкăна хыврăм – манăн пуçăм вĕриччĕ, чĕрем çурăлас пек ыратрĕ, эпĕ кайма пултараймарăм.

Как видно из сопоставляемых текстов, в рассуждении-объяснении не доказывается истинность или ложность тезиса, а только раскрывается содержание тезиса;

в) рассуждение-объяснение: Узнаю ли я когда-нибудь, что случилось с этим человеком, как будто поручившим мне рассказать историю его жизни, его смерти? Оставил ли он корабль, *чтобы* изучить открытую им землю, или погиб от голода вместе со своими людьми, и шхуна, замерзшая во льдах у берегов Ямала, годами шла путем Нансена к Гренландии с мертвой командой? Или в холодную, бурную ночь, когда не видно ни звезд, ни луны, ни северного сияния, она была раздавлена льдами, и с грохотом полетели вниз мачты, стеньги и реи, ломая все на палубе и убивая людей, в предсмертных судорогах затрещал корпус, и через два часа пурга уже занесла снегом место катастрофы? Или еще живут где-нибудь на безлюдном полярном острове люди со “Св. Марии”, которые могли бы рассказать о судьбе корабля, о судьбе капитана? Ведь прожили же несколько лет на необитаемом уголке Шпицбергена шесть русских матросов, били медведей и тюленей, питались их мясом, одевались в их шкуры, устилали шкурами пол своего шалаша, который они сделали изо льда и снега! Да нет, куда там! Минуло двадцать лет, как была высказана “детская” безрассудная мысль покинуть корабль и идти на Землю Марии (Каверин). – Çак çынпа, мана хăйĕн пурнăçĕн историйĕ çинчен каласа пама хушса хăварнă пек çынпа мĕн пулни çинчен хăçан та пулин пĕлейĕп-ши эпĕ? Хăй уçнă çĕре тĕпчесе пĕлес *тесе* карапне пăрахса хăварнă-ши вăл е хăйĕн çыннисемпе пĕрле выçă вилнĕ-ши? Е Ямал çыранĕсем патĕнче, пăр хушшинче шăнса ларнă шхуна, хăйĕн вилсе пĕтнĕ командипе пĕрле, Нансен çулĕпе Гренланди еннелле темиçе çул хушши кайрĕ-ши? Е пĕр-пĕр сивĕ, çил-тăвăллă каç, çăлтăр та, уйăх та курăнман каç вăл пăр хушшине пулса ванса, саланса пĕтрĕ-ши? Мачтăсем, рейсем, стеньгăсем палуба çинчи япаласене ватса, çĕмĕрсе, çынсене вĕлерсе ишĕле-ишĕле анчĕç

пулĕ? Унăн корпусĕ вилес умĕн чĕтренсе шатăртатса кайрĕ пуль? Тепĕр икĕ сехетрен çил-тăман вара катастрофа пулнă вырăна юрпа хупласа хучĕ пулĕ? Тен, çурçĕрте пĕр-пĕр çынсăр-мĕнсĕр утрав çинче "Святая Мария" çыннисем пурăнаççĕ пулĕ? Вĕсем карап пурнăçĕпе капитан пурнăçĕ çинчен те каласа парĕччĕç, Шпицбергенĕн çын пырса кĕмен кĕтесĕнче 6 вырăс матросĕ темиçе çул хушши пурăннăçке-ха. Тюленьсем тытса, тюлень ашĕ çисе, çийĕсене тюлень тирĕ тăхăнса, хăйсен юрпа пăртан тунă шалаш урайне те унăн тирне сарса пурăннă-çке-ха? Çук, ăçтан ун пек пултăр-ха! Карапа пăрахса Мария çĕрĕ çине каяс *mece* "ачалла", "айванла" шухăшланăранпа çирĕм çул та иртсе кайнă.

Рассуждение-размышление включает в себя объяснение и доказательство, в котором необходимо привести примеры, указать причинно-следственные отношения.

Менее частотными являются монологи-повествования, которые призваны передать развитие действия и следуют друг за другом во времени и пространстве: Семья распадалась на глазах у Пантелея Прокофьевича. Они со старухой оставались вдвоем. Неожиданно и быстро были нарушены родственные связи, утрачена теплота взаимоотношений, в разговорах все чаще проскальзывали нотки раздражительности и отчуждения. За общий стол садились не так, как прежде – единой и дружной семьей, а как случайно собравшиеся вместе люди. Война была всему этому причиной. Пантелей Прокофьевич это отлично понимал (Шолохов). – Çемье Пантелей Прокофьевич куçĕ умĕнчех арканса пычĕ. Вĕсем ĕнтĕ карчăкĕпе иккĕшех тăрса юлчĕç. Тăванла çыхăнусем кĕтмен çĕртен хăвăрт саланчĕç, пĕр-пĕринпе шăкăл-шăкăл килĕштерсе пурăнасси пĕтрĕ. Каласу хушшинче час-часах тарăхса кайни, пĕр-пĕринчен ютшăнни сисĕнсе тăма пуçларĕ... Пĕр сĕтел хушшине апатланма ĕлĕкхи пек пĕрешкел шухăш-кăмăллă та туслă килйыш майлă мар, ăнсăртран пĕрле пуçтарăннă çынсем пек кĕре-кĕре ларчĕç. Çакăн пĕтĕм сăлтавĕ – вăрçă. Ăна Пантелей Прокофьевич питĕ аван ăнланать.

Средства межфразовой связи в каузальных, финальных, кондициональных, концессивных текстовых комплексах неоднородны по своему составу, далеко не одинаково выполняют эту функцию и могут по-разному комбинироваться в целях создания связности текста. Текстовые комплексы со значением обусловленности выступают как единицы гиперсинтаксиса. Их текстуальными средствами являются как лексические, так и лексико-синтаксические актуализаторы межфразовой причинно-следственной связи. Для практического исполь-

зования лексико-синтаксических шаблонов нами была разработана на языке Java библиотека PattemLib, а в дальнейшем соответствующая библиотека может быть востребована для первичного извлечения информации русско-чувашским лингвистическим процессором, чтобы сузить круг задач, требующих специфического предметно-ориентированного решения при автоматической обработке текста (в рамках этого подхода задача обработки текста ограничена распознаванием множества классов базовых (ключевых) понятий и игнорированием всякой другой информации. Рассмотренные модели предоставляют пользователям, как экспертам, так и лингвистам, материалы для автоматической обработки текста, а инженерам знаний инструменты для проектирования систем автоматической обработки текста и концептуальные схемы представления данных.

ЛИТЕРАТУРА

1. Большакова Е.И., Васильева Н.Э., Морозов С.С. Лексико-синтаксические шаблоны для автоматического анализа научно-технических текстов // Десятая Национальная конференция по искусственному интеллекту с международным участием КИИ-2006. Труды конференции в 3-х томах. – М.: Физматлит, 2006. – Т. 2. – С. 506–524.

2. Гладкий А.В Синтаксические структуры естественного языка в автоматизированных системах общения. – М.: Наука. Главная редакция физико-математической литературы. 1985. – 144 е. – (Серия «Проблемы искусственного интеллекта»).

3. Губанов А.Р., Губанова Г.Ф, Свеклова О.В. Тезаурус чувашского языка (чăваш пĕлĕвĕн мулĕ) как языковая система знаний / А. Р. Губанов, Г. Ф. Губанова, О. В. Свеклова // Вестник Чувашского университета. Гуманитарные науки. – 2017. – № 2. – С. 190–194. 95.

4. Губанов А.Р., Исаев Ю.Н., Свеклова О.В., Губанова Г.Ф. Обусловленность как лингвистическая онтология верхнего уровня / // Вестник Чувашского университета. Гуманитарные науки. – 2017. – № 4. – С. 271–278.

5. Горшков С. И. Введение в онтологическое моделирование. – М., 2016.

6. Золотова Г.А. Синтаксический словарь. Репертуар элементарных единиц русского синтаксиса. – М. 2011.

7. Кибрик А.Е. Очерки по общим и прикладным вопросам языкознания. – М.: УРСС, 2001. – С. 123–124.

8. Козеренко Е.Б. Функциональная семантика в компьютерных решениях // Труды Международного семинара «Диалог'2002» по

компьютерной лингвистике и интеллектуальным технологиям. г. Протвино Московской обл. – 2002. – Т. 1. – С. 218–226.

9. Козеренко Е.Б. Унифицированные категориально-функциональные представления для синтаксической разметки полнотекстового документа//Системы и средства информатики. – М.: Наука, 2002. – Вып. 11.

10. Kozerenko E.B. Portable Language Egineering Solutions for Multilingual Processors // Proceedings of the International Conference on Artificial Intelligence IC-AI'02// CSREA Press, 2002, pp. 447–453.

11. Невзорова О.А. Онтолингвистические системы: технологии взаимодействия с прикладной онтологией // Ученые записки КГУ. – Том 149. Серия Физико-математические науки. – Книга 2. – С. 105–115.

УДК 81’32

FEATURES OF TRANSLATION THE TERMS FROM ONTOLOGY OF PLANIMETRY IN ENGLISH

Anastasia Eduardovna Dyupina
Kazan Federal University, Kazan
anastasiya.dupina@yandex.ru

The article covers the issues of translating the terms of the school Planimetry course into English. Supporting studies are the researches by Grinev S.V., Vinogradov V.V. and other linguists in the field of translation. The main difficulty in translating planimetry terms is the need to know not only Russian and English, but also mathematics. The aim of the study is to identify suitable ways to translate terms based on dictionaries, textbooks on geometry and Internet resources. As a result of the work done, a list of terms needing particularly careful translation was identified, and an algorithm for translating complex terminological structures was compiled. The data obtained may be useful to mathematicians engaged in writing scientific articles in English, as well as specialists in other areas for expanding horizons.

Keywords: terms, translation, Planimetry.

ОСОБЕННОСТИ ПЕРЕВОДА ТЕРМИНОВ ОНТОЛОГИИ ПЛАНИМЕТРИИ НА АНГЛИЙСКИЙ ЯЗЫК

Анастасия Эдуардовна Дюпина
Казанский федеральный университет, Казань
anastasiya.dupina@yandex.ru

Статья освещает вопросы перевода терминов школьного курса планиметрии на английский язык. Опорными являются исследования Гринева С.В., Виноградова В.В и других лингвистов в области перевода. Основная трудность перевода терминов планиметрии заключается в необходимости знания не только русского и английского языков, но и математики. Целью исследования является выявление подходящих способов перевода терминов с опорой на словари, учебники по геометрии и Интернет-ресурсы. В результате проделанной работы был выявлен список терминов, нуждающихся в особо тщательном переводе, составлен алгоритм перевода сложных терминологических структур. Полученные данные могут быть полезны математикам, занимающимся написанием научных статей на английском языке, а также специалистам других областей для расширения кругозора.

Ключевые слова: термины, перевод, планиметрия.

Введение

Расширение международной коммуникации невозможно без знания языков. Роль языка в науке очень велика: благодаря ему выражаются понятия и теоретические положения. В современной лингвистике язык науки определяется как научно-технический стиль речи. Люди, занимающиеся наукой, излагают свои мысли и достижения в виде научных трудов. Для большего распространения результатов исследования приходится прибегать к общению и написанию текстов на международном языке науки – английском, для этого требуется выполнить качественный перевод.

Наибольшую трудность при переводе текстов технической направленности представляет перевод терминов. Это связано с тем, что технические тексты исключают возможность двойственности смысла и искажения используемых понятий. Без правильного перевода терминов научный текст может потерять свое предназначение.

В процессе работы над проектом «Разработка семантических программных инструментов для повышения качества математического образования в Республике Татарстан» (номер проекта – 18-47-160007) осуществляется создание новой математической онтологии OntoMathEdu для системного описания математических знаний школьного курса математики. Наполнение понятиями школьного курса планиметрии осуществляется на трех языках: английском, русском и татарском.

Данное исследование является актуальным, поскольку в настоящее время отсутствуют подобного рода проекты, содержащие полный перечень понятий школьного курса планиметрии на многоязычной основе. В рамках настоящей статьи описываются особенности перевода терминов и понятий онтологии планиметрии с русского языка на английский.

Обзор литературы

На протяжении долгого времени вопросы перевода технических терминов находятся в центре внимания многих ученых, среди которых можно выделить следующих лингвистов: Виноградов С.Н., Винокур Г.О., Гринев С.В., Реформатский А.А. и др. Несмотря на это, до сих пор существует ряд нерешенных проблем: проблема многозначности термина, в частности, актуальна для нашего исследования.

Гринев С.В. дает следующее определение термина: «Номинативная специальная лексическая единица (слово или словосочетание) специального языка, принимаемая для точного наименования специальных понятий» [5]. По мнению В.Н. Шевчука, даже в пределах конкретной терминологии технический термин может быть многозначен, из чего делает вывод, что «однозначность – не свойство термина, а требование, к нему предъявляемое» [6].

Для нашего исследования очень важно найти пути перевода терминов и терминологических конструкций, не потеряв их первоначальное значение. Перед тем, как перейти непосредственно к трудностям, возникшим при переводе терминов онтологии, рассмотрим сам процесс перевода слов наиболее часто встречающимися способами. Все примеры терминов и понятия, использованные в дальнейшем, взяты из разрабатываемой онтологии OntoMathEdu.

В процессе перевода происходит сближение двух лингвистических систем: язык оригинала и язык, на который осуществляется перевод. При этом первый из них остается устойчивым и не подвергается изменениям, второй же адаптируется под оригинал. В нашем случае в роли языка оригинала выступает русский язык, языка перевода – английский.

Случай, когда перевод термина осуществляется легко и к оригиналу можно подобрать эквивалент на языке перевода, основан на параллельности категорий (явление структурного параллелизма), либо на параллельности понятий (явление металингвистического параллелизма): треугольник – triangle, уравнение – equation. Бывают случаи, когда подобрать эквивалент не представляется возможным, и в результате в языке перевода образуется пробел, устранить который помогает парафраз: ломаная – polygonal line.

В теории перевода выделяются два основных пути: *прямой* и *косвенный* [4]. К прямому переводу относятся заимствование, калькирование, дословный перевод, к косвенному – транспозиция, модуляция, эквиваленция и описательный перевод.

Заимствование, как самый простой способ перевода, подразумевает передачу слова с помощью фонем: метр – meter, линия – line, вектор – vector. Однако среди заимствований часто можно встретить так называемых «ложных друзей переводчика»: задача – problem (проблема – trouble).

Калькированием осуществляется перевод по морфемам слова: методологический – methodology, неколлинеарные – non-collinear.

При *дословном переводе* необходимо обращать внимание на соблюдение правил языка перевода, а именно на согласование частей речи: теорема Пифагора – Pythagorean theorem, пучок окружностей – bundle of a circles.

Метод *транспозиции*, как способ косвенного перевода, заключается в замене одной части речи (язык оригинала) другой (язык перевода) без искажения смысла исходной фразы: прямая – line (имя существительное, являющееся субстантивированным прилагательным в русском языке, перешло в имя прилагательное в английском языке), дифференциальное уравнение – differential equation, дифференциал – differential (имя прилагательное перешло в имя существительное). Методом транспозиции следует пользоваться в том случае, если конечный оборот лучше передает смысл исходной фразы или позволяет скорректировать стилистические нюансы.

Модуляция представляет собой изменение посыла, достигаемое варьированием точки зрения. Этот способ удобен, например, в том случае, когда при дословном переводе получается высказывание грамматически правильное, но не соответствующее стилю перевода или языка в целом. При модуляции происходит замена слов или словосочетаний такими конструкциями, которые можно получить логическим путем из посыла. При дословном переводе фразы «задачи на построение циркулем и линейкой» получится следующее: «task of building a compass and ruler». С точки зрения английского языка данная конструкция выглядит «нелепо» и вовсе не существует в курсе планиметрии, гораздо более приемлемо будет выглядеть такой вариант перевода: «constructing exersises using compass and straightedge».

Эквиваленция является достаточно трудным для понимания способом перевода, поэтому для лучшего восприятия этого способа перевода разумнее сразу привести пример. Когда человек болеет за свою любимую спортивную команду, он активно выражает свои эмоции, при этом россиянин будет использовать слова «ура!», «эх!» и др., англичанин же выкрикивает «cheers!», «oh!». Данный пример хорошо иллюстрирует тот факт, что большая часть эквиваленций, которые мы часто используем в повседневной жизни, представляют собой устойчивые сочетания, входящие раздел идиоматической фразеологии. Необходимо понимать различие между потенциально достижимой эквивалентностью и эквивалентностью переводческую. В первом случае достигается максимальное сходство содержания текстов обоих языков, во втором случае осуществляется

реальная смысловая близость, достигаемая переводящим. Идеалом переводческой эквиваленции выступает максимальная возможность сохранения содержания оригинала при переводе. Под эквивалентностью перевода первого вида понимается сохранение части оригинального текста, составляющей предназначение коммуникации, представляющую собой наиболее информативную и важную часть содержания. Примером эквиваленции в рамках нашего исследования служит следующий: средняя линия – midsegment. Осуществляя перевод дословно, получим middle line, такой термин не существует в языке перевода в рамках изучаемой темы.

Метод *описательного* перевода используется в случае, когда в языке перевода отсутствует ситуация, описываемая языком оригинала, и ее необходимо заменить аналогичной. Иными словами, перевод осуществляется путем описания слова (словосочетания) или его объяснением: середина отрезка – bisecting point of a segment. Хорошо иллюстрирует данный способ перевод с английского на русский язык: incenter – центр вписанной окружности, circumcenter – центр описанной окружности.

Наибольшую трудность представляет перевод сложных терминологических структур, которых в нашей онтологии достаточно много. В математической литературе часто встречаются термины, состоящие из нескольких слов, которые нуждаются в правильном переводе. Примеры таких структур уже приводились выше. Для понимания техники перевода произведем пошаговый перевод одного из подклассов онтологии «Характеристическое свойство четырехугольника».

Шаг 1. Выявление ключевого слова – КС (ядра), левого (ЛО) и правого (ПО) определений: *свойство* – ядро; *характеристическое* – левое определение, *четырехугольника* – правое определение.

Шаг 2. Перевод КС, как основополагающего элемента структуры: *свойство – property*.

Шаг 3. Перевод основного элемента группы (КС) совместно с левым определением (ЛО). Такое определение является наиболее близким для КС. В этом случае задается вопрос: *какое свойство*? В результате получаем перевод *характеристическое свойство – characteristic property*.

Шаг 4. Осуществление дальнейшего перевода совместно с правым определением (ПО). Ставится вопрос: *характеристическое свойство чего*? Важным моментом является согласование частей речи (существительного, выраженного подлежащим, и су-

ществительного-дополнения для данного примера): *характеристическое свойство четырехугольника – characteristic property of quadrilateral*. Для данного термина перевод закончен, однако для более распространенных терминологических конструкций будет проделано большее количество шагов.

На основании проделанных выше шагов можно сделать вывод о том, что перевод сложных терминологических структур осуществляется от ключевого слова к последнему из определений. Как правило, последнее определяющее слово придает термину более узкое значение, конкретизирует его.

Рассмотрим перевод класса онтологии «Взаимное расположение геометрических фигур на плоскости»:

1. *расположение* – КС

взаимное – ЛО, *геометрических фигур* – $ПО_1$, *на плоскости* – $ПО_2$

2. *расположение – arrangement* (КС)

3. расположение какое?

взаимное расположение – mutual arrangement

4. взаимное расположение чего?

взаимное расположение геометрических фигур – mutual arrangement of geometric figures

5. взаимное расположение геометрических фигур на чем?

взаимное расположение геометрических фигур на плоскости – mutual arrangement of geometric figures on a plane

Можно заметить, что при переводе конструкции был пропущен шаг выделения ядра и определения фрагмента «геометрических фигур», который уже является очевидным на основании разобранного выше примера. Также отличительной особенностью левого и правого определений можно назвать их принадлежность к части речи: левое определение чаще всего отвечает на вопрос *какой?* и бывает представлено именем прилагательным, причастием, именем числительным или существительным без предлога, правое определение может отвечать на большее количество вопросов (*чего? на чем?*) и выражается существительным или существительным с предлогом в рамках нашей терминологии (существуют и другие представления).

Материалы и методы

В процессе перевода терминов онтологии приведенных выше способов оказалось недостаточно: работа усложнилась наличием нескольких вариантов перевода одного и того же термина, либо его

отсутствием в словаре. В процессе перевода терминов онтологии в первую очередь использовались специальные словари математических терминов [1], [7], [11], зарубежные учебники по геометрии [8], [13]. В случае возникновения пробела при переводе использовался онлайн-переводчик [10], осуществлялся поиск на различных ресурсах сети Интернет [12] при помощи парафраза (объяснения термина на английском языке). Выбор словарей [7], [11] обусловлен тем, что они являются наиболее информативными среди современных электронных словарей, содержат большое количество математических терминов, просты в использовании, что немаловажно.

Результаты исследования

Из определения термина, приведенного в начале статьи, следует, что термин должен быть однозначен в рамках определенной терминологии, вследствие чего возникает вопрос, какой из предложенных вариантов перевода термина наиболее адекватен и информативен. Поскольку нашей целью является создание онтологии, которой в дальнейшем смогут пользоваться школьники, студенты и преподаватели, мы должны быть уверены, что термин русского языка будет правильно переведен на английский язык.

Приведем примеры терминов, содержащихся в онтологии планиметрии и имеющих несколько вариантов перевода на английский язык (в выбранных источниках):

Высота – altitude, height

Катет – leg of right triangle, catheter, leg, side

Кривая второго порядка – point conic, quadratic curve, second-order curve, curve of the second order

Линия центров – line of centers, centre line, centerline

Ломаная – kinked curve, polygon, polygonal curve, polygonal line, broken line, polygonal chain

Прямая – line, straight line

Отрезок касательной – length of tangent, tangent segment

Середина отрезка – bisecting point of a segment, midpoint

Это далеко не весь список, но можно увидеть, что проблема поиска наиболее подходящего эквивалента имеет место для нашего исследования.

Кроме того, при переводе возникла проблема выбора единственного и множественного числа некоторых терминов. С одной стороны, разумно употреблять все термины в выстраиваемой онтологии

в единственном числе, однако, с другой стороны, в курсе планиметрии встречаются парные понятия (например, смежные углы, вертикальные углы), которые вступают в определенные отношения зависимости одного объекта от другого. Другими словами, один из смежных углов не может существовать без другого, смежного с ним, аналогичные рассуждения имеют место и в случае с вертикальными углами. Однако выбранные источники предлагают варианты как множественного, так и единственного числа:

Вертикальный угол – altangle, vertical angle [7], [11]

Вертикальные углы – vertical angles [1], [9], [13]

Смежный угол – adjacent angle, adjoining angle [1], [7], [11]

Смежные углы – adjacent angles [8], [9], [13]

Подавляющее большинство подобных примеров содержатся в классе онтологии «Взаимное расположение геометрических фигур на плоскости», например, касающиеся окружности – touching circles, параллельные прямые – parallel lines и др.

В курсе планиметрии существуют такие понятия, как *замечательные точки* и *замечательные прямые*. Перевод их на английский язык оказался весьма затруднителен, поскольку термины отсутствуют в словарях. Онлайн-переводчик предлагает дословный перевод, который не соответствует реальным объектам. В результате поиска информации по данной теме в учебниках [8], [13] и в сети Интернет [12] были найдены следующие варианты перевода:

Замечательные точки – triangle centers, triangle concurrency points

Замечательные прямые – special lines

Можно заметить, что одно и то же слово «замечательные» имеет различные варианты перевода, который зависит от определяемого слова. Кроме того, само слово отдельно от данного контекста имеет совершенно другой вариант перевода – remarkable.

Выводы

В настоящее время известно много способов перевода, однако не все они применимы к терминам языка науки. Научные тексты лишены эмоциональной экспрессии и имеют нейтральный стилистический окрас. Термины, употребляемые в таких текстах, должны однозначно определять рассматриваемые объекты. Однако полученные результаты еще раз подтверждают предположение Шевчука В.Н. о возможной многозначности термина в определенной терминологии.

Проблема выбора правильного варианта перевода остается актуальной и зачастую требует проведение дополнительного исследования. При переводе текста недостаточно хорошо знать язык оригинала и язык перевода: необходимо учитывать также культуру и реалии второго языка.

При переводе терминов выстраиваемой онтологии планиметрии осуществлялась опора на несколько источников, проводился глубокий анализ, изучались темы школьного курса планиметрии по зарубежным учебникам. Полученные расхождения в переводе свидетельствуют о необходимости такого подхода и исключают использование только одного источника. При заполнении онтологии терминами, имеющими несколько вариантов перевода, предпочтение будет отдаваться тем вариантам, которые найдены в учебниках, которыми в настоящее время пользуются в школах и вузах в англоговорящих странах.

Данная работа носит прикладной характер, поскольку будет полезна при дальнейшем наполнении онтологии терминами из других разделов математики. Исследование вопроса перевода терминов планиметрии не является законченным и будет продолжено в дальнейшем.

Благодарности. Работа выполнена при финансовой поддержке РФФИ и Правительства Республики Татарстан в рамках научного проекта № 18-47-160007.

ЛИТЕРАТУРА

1. Англо-русский словарь математических терминов. / Под ред. П.С. Александрова. – 2-е, исправл. и дополн. изд. – М.: Мир, 1994. – 416 с.

2. Бархударов Л.С. Язык и перевод: Вопросы общей и частной теории перевода. / Л. С. Бархударов. – М.: Международные отношения, 2005. – 240 с.

3. Борисова Л.И. Лексико-стилистические трансформации в англо-русских научно-технических переводах. / Л.И. Борисова. – М.: ВЦП, 2003. – 168 с.

4. Вине Ж. –П. Технические способы перевода // Вопросы теории перевода в зарубежной лингвистике / Ж.-П. Вине, Ж. Дарбельне. – М., 1978. http://www.philology.ru/linguistics1/vinay-darbelnet-78.htm

5. Гринев-Гриневич С. В. Терминоведение учеб. пособие для студ. высш. учеб. заведений / С. В. Гринев-Гриневич. – М.: Издательский центр «Академия», 1993, 2008. – 304 с.

6. Шевчук В.Н. Производные военные термины в английском языке: [Аффиксальное словопроизводство]. Москва: Воениздат, 1983. – 231 с.

7. Электронный словарь «Мультитран». [Электронный ресурс] / Режим доступа: https://www.multitran.ru/

8. Daniel C. Alexander, Geralyn M. Koeberlein. Elementary Geometry for College Students, 6th Edition. 2015. 628 p.

9. Douglas Downing, Ph.D. Dictionary of mathematics terms. Third Edition. 2009. 441 p.

10. Google Переводчик https://translate.google.com/?hl=ru

11. Lingvo 12.0: Большой англо-русско-английский общелексический словарь. Электронная версия, 2008.

12. Math Open Reference. [Электронный ресурс] / Режим доступа: https://www.mathopenref.com/index.html

13. Roland E. Larson, Laurie Boswell, Lee Stiff. Heath Geometry an Integrared Approach. Teacher's Edition. 1998. 876 p.

УДК 81-25

PRAGMATIC MARKERS IN THE CORPUS “ONE DAY OF SPEECH”: APPROACHES TO THE ANNOTATION

K. D. Zaides, T. I. Popova, N. V. Bogdanova-Beglarian
Saint Petersburg State University, Saint Petersburg
kristina.zaides@student.spbu.ru, tipopova13@gmail.com,
n.bogdanova@spbu.ru

The article describes the scheme of the annotation of pragmatic markers in the corpus of Russian everyday speech “One Day of Speech”. Pragmatic markers are defined as special units in the speech that have only pragmatic function without any (or with ‘bleached’) lexical meaning. The annotation of pragmatic markers is usually performed manually due to the existing ambiguity of markers in different contexts. The typology of pragmatic markers includes different groups marked with special annotation tags. The annotation process was split into two stages since several issues of tagging of PMs arose. The main problems, which occurred during the annotation process, and the possible ways of their solution are also discussed in the research. The paper propose the improved methods of problem solving during the annotation of pragmatic markers applied to the corpus of oral speech, which can be useful for the linguistic annotation of any other levels of oral speech.

Keywords: pragmatic markers, spoken speech, corpus of everyday speech, corpus linguistics, corpus annotation.

АННОТИРОВАНИЕ ПРАГМАТИЧЕСКИХ МАРКЕРОВ В КОРПУСЕ «ОДИН РЕЧЕВОЙ ДЕНЬ»: ВОЗМОЖНЫЕ ПОДХОДЫ

К. Д. Зайдес, Т. И. Попова, Н. В. Богданова-Бегларян
Санкт-Петербургский государственный университет, Санкт-Петербург
kristina.zaides@student.spbu.ru, tipopova13@gmail.com,
n.bogdanova@spbu.ru

В статье описываются основные принципы аннотации прагматических маркеров в корпусе русской повседневной речи «Один речевой день». Прагматические маркеры определяются как условно-речевые единицы, не имеющие лексического и/или грамматического значения, выполняющие в речи ряд определенных прагматических функций. Аннотация прагматических маркеров в корпусе была реализована впервые и сделана вручную

ввиду отсутствия инструментов автоматического аннотирования и наличия в речи полифункциональных маркеров, а также нерешенной проблемы снятия омонимии маркеров и полнозначных выражений. Перед аннотацией была разработана типология прагматических маркеров, которая легла в основу системы используемых тегов. Аннотация маркеров проходила в два этапа, что позволило решить возникшие в ходе этого процесса проблемы. Описание решения проблем аннотации составляет важную часть данного исследования; описываемые принципы аннотирования материала могут быть полезны корпусным лингвистам при решении других проблем аннотирования устной речи.

Ключевые слова: прагматический маркер, разговорная речь, корпус повседневной речи, корпусная лингвистика, аннотация корпуса.

Прагматический маркер (ПМ) – относительно новый термин в современной лингвистике, введенный в научный обиход Н.В. Богдановой-Бегларян и обозначающий те единицы устной речи, которые утрачивают свое непосредственное лексическое и/или грамматическое значение и в процессе повторяющегося речевого употребления начинают выполнять только определенные прагматические функции [1]. Отличия прагматических маркеров от дискурсивных (ДМ), под которыми в зарубежной практике понимается очень широкий класс дискурсивных функциональных единиц, сводятся к следующему [2, 3]:

ПМ употребляются говорящим бессознательно, на уровне речевого автоматизма; ДМ вводятся в текст осознанно, прежде всего с целью его структурирования;

ПМ не имеют (или имеют ослабленное, почти исчезнувшее) лексическое и/или грамматическое значение; ДМ являются полнозначными единицами устного дискурса;

ПМ употребляются только (или по большей части) в спонтанной речи; ДМ встречаются как в письменном тексте, так и в спонтанной речи;

ПМ демонстрируют отношение говорящего к самому процессу порождения речи, вербализуя все затруднения и колебания говорящего и являясь зачастую метакоммуникативными единицами; ДМ передают лишь отношение говорящего к тому, о чем он сообщает;

ПМ не включены в словари в их функциональном разнообразии; ДМ являются частью традиционной лексикографии, будучи лексемами, с одной стороны, а также рассматриваются в дискурсивных исследованиях как операторы структурирования высказываний, с другой стороны.

Типология прагматических маркеров включает несколько групп единиц: маркеры границы высказывания, маркеры-аппроксиматоры, маркеры-заместители, маркеры-ксенопоказатели, метакоммуникативные, дейктические, поисковые, рефлексивные, самокорректирующие, ритмообразующие и хезитационные маркеры [1, 3, 4], например:

• *ну в общем дефект кишки / когда (э) на ней такой отросточек / как бывает* ***вот*** *(...) (э-э) в венах / как аппендикс /* ***вот такой вот*** *какой-то* ***там*** [И130];

• *Наташа% / вы уже отпустили этого / () Алексея%(:) / Максима% /* ***и всего прочего*** *?* [И19];

• ***я говорю*** *я тогда в девяти три... там к девяти пятнадцати приду / пока* ***то сё*** *...* [СИ124];

• *ну и Вадик% приезжает / *П и они ему* ***говорят слушай*** *чувак мы тебе всё отремонтировали / *П только мы тебе* ***короче*** *(...) (э-э) в бак (...) вместо(:) (э) дизеля девяносто восьмой залили* [И72];

• *ну Андрей% / тогда* ***вы смотрите*** */* ***значит*** *я до девяти буду (...) ну (э) телефон выключу / и отвечать не буду* [И123];

• ***ну там*** *(...) сильно дешевле не было / потому что я () здесь* ***как бы*** */ они всё равно ехали* [И103].

Материалом для аннотации послужили 12 файлов из корпуса «Один речевой день» (ОРД), крупнейшего ресурса для изучения устной спонтанной диалогической речи, сформированного по методике 24-часовой записи [5, 6, 7]. Аннотирование выполнялось в два этапа четырьмя аннотаторами, независимо друг от друга, в программе ELAN. Для аннотирования было создано 4 дополнительных уровня, содержащие сам маркер, его функциональный тип, атрибуцию говорящему и дополнительные комментарии.

Для аннотирования прагматических маркеров на первом этапе была разработана специальная инструкция, включающая список тегов (аббревиатура из первых букв названия типа ПМ), список интонационных и иных помет. Функции маркеров описывались в направлении от главной к побочным, в алфавитном порядке. Возможность выделять новые маркеры и новые функции также предоставлялась экспертам. В помощь аннотаторам была создана таблица предварительных функций (главной и дополнительных) наиболее частотных ПМ с контекстами из корпуса ОРД.

После первого этапа аннотирования выяснилось, что согласованность аннотаторов, вычисленная с помощью коэффициента каппа Козна, довольно низка [8]. Было решено изменить инструкцию та-

ким образом, чтобы повысить согласованность разметчиков. Помимо этого, были обнаружены некоторые общие проблемы разметки прагматических маркеров.

Первой проблемой стало само членение спонтанной речи на синтагмы, которое не всегда можно выполнить однозначно, ср.:

• *я сейчас позвоню Марине% / и выясню // дело в том что / к вам собиралась Марина% ехать Жданова% // не не не не не не // *В Марина% Глухарева% // *Н* ***вот*** */ *П и (:) (э-э)* ***вот*** */ я выясню / поедет она сегодня или завтра к вам* [И19].

В примерах такого типа маркер *вот* описывался непоследовательно: как маркер старта, как навигационный ПМ или как маркер финала. Ясно, что основной функцией такого маркера является указание на границу между высказываниями, структурирование речи. Вследствие этого стартовые, направляющие и финальные маркеры были объединены в одну группу маркеров границы. Это позволило решить проблему присваивания тегов для маркеров, стоящих в сложных, однозначно не квалифицируемых, синтаксических позициях в устной речи.

Во-вторых, большую трудность представляет живой процесс прагматикализации, в рамках которого можно выделять стадии, которые проходит ПМ, – от полнозначной конструкции, имеющей парадигму форм, через процесс грамматикализации и прагматикализации, к лексикализованному выражению, употребляющемуся как речевой автоматизм, ср.:

• *ну понятно дело / ну *пта / а(:) да тебе вообще / даже законные выходные могут не дать / да ?* ***я думаю*** *// *П у меня там поднакопилось этих самых / неиспользованного отпуска / да / поэтому я и использую* [И110].

Думается, что выявление этапов прагматикализации с относительно четким набором критериев позволит улучшить процесс прагматического аннотирования и повысить согласованность разметки.

Третьей крупной задачей в ходе прагматического аннотирования стало определение главной и побочных функций конкретных ПМ. Так, в следующем примере:

• *ну там в основном советскую читал / знаешь литературу // нашу* ***там*** */ а(:) ! вперед к коммунизму !* [И15]

невозможно однозначно сказать, является ли главной функцией ПМ *там* хезитативная или аппроксимативная. На втором этапе аннотирования было отменено деление функций на главные и дополнительные, что позволило решить проблему иерархии функций. В

дальнейшем планируется определить точные характеристики превалирования функций маркеров в разных типах контекстов.

Последняя проблема, возникшая в ходе прагматической аннотации, – решение вопроса о том, является ли выражение одним неоднословным маркером или цепочкой из нескольких ПМ:

• *вчера мы с на... с Надей% выходим с работы // *П она меня просит / у вас есть там телефон (э-э) Глухаревой% ? я говорю да // *П ну и* ***значит там*** *(...) нахожу / диктую ей* [И19].

Маркирование границы, передача хезитации и размытие значения являются общими для двух единиц – *значит* и *там* – функциями. На наш взгляд, решение вопроса о составе маркера должно опираться на формулу: «одна интонационная единица + одна функция = один маркер»; в противном случае, перед нами два разных ПМ.

Улучшенная с учетом рассмотрения данных проблем инструкция для разметчиков (однобуквенные теги, группировка ПМ, критерий определения состава ПМ и т. д.) позволила увеличить показатель согласованности аннотаторов и повысить точность и скорость прагматической разметки. Планируется дальнейшее расширение аннотируемого материала и развитие фундаментальных принципов разметки, а также решение частных задач.

Благодарности. Работа выполнена при финансовой поддержке РНФ (проект № 18-18-00242 «Система прагматических маркеров русской повседневной речи»).

ЛИТЕРАТУРА

1. Богданова-Бегларян Н. В. Прагматемы в устной повседневной речи: определение понятия и общая типология // Вестник Пермского университета. Российская и зарубежная филология. 2014. № 3(27). С. 7–20.

2. Bogdanova-Beglarian N. V., Filyasova Yu. A. Discourse vs. pragmatic markers: a contrastive terminological study. 5th International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2018, SGEM2018 Vienna ART Conference Proceedings, 19–21 March, 2018, vol. 5, 2018, pp. 123–130.

3. Богданова-Бегларян Н. В. О возможных коммуникативных помехах в межкультурной устной коммуникации // Мир русского слова. 2018. № 3 (в печати).

4. Bogdanova-Beglarian N., Sherstinova T., Blinova O., Martynenko G., Baeva E. Towards a description of pragmatic markers in Russian everyday speech. LNAI, vol. 11096: Speech and Computer. 20th International

Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings. Springer Publishing Company, 2018, pp. 42–48.

5. Богданова-Бегларян Н. В., Асиновский А. С., Блинова О. В., Маркасова Е. В., Рыко А. И., Шерстинова Т. Ю. Звуковой корпус русского языка: новая методология анализа устной речи // Язык и метод: Русский язык в лингвистических исследованиях XXI века. Вып. 2 / Ред. Д. Шумска, К. Озга. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego, 2015. С. 357–372.

6. Bogdanova-Beglarian N., Sherstinova T., Blinova O., Martynenko G. Linguistic features and sociolinguistic variability in everyday spoken Russian. SPECOM 2017, LNAI, vol. 10458, 2017, pp. 503–511.

7. Русский язык повседневного общения: особенности функционирования в разных социальных группах. Коллективная монография / Ред. Н. В. Богданова-Бегларян. СПб.: ЛАЙКА, 2016.

8. Bogdanova-Beglarian N., Blinova O., Sherstinova T., Martynenko G., Zaides K. Pragmatic markers in Russian spoken speech: an experience of systematization and annotation for the improvement of NLP tasks. Proceedings of the 23rd Conference of Open Innovations Association FRUCT. Bologna, Italy, 13–16 November 2018 (в печати).

УДК 81’25

MAIN DIFFICULTIES IN THE PROCESS OF CREATING MACHINE TRANSLATION SYSTEMS

E. V. Zamiraylova
e.zamiraylova@gmail.com

The article describes the main difficulties that arise in the process of creating machine translation systems on the example of Russian and English languages.

Keywords: machine translation, machine translation system.

ОСНОВНЫЕ ТРУДНОСТИ ПРИ СОЗДАНИИ СИСТЕМ МАШИННОГО ПЕРЕВОДА

Е. В. Замирайлова
Санкт-Петербургский государственный университет, Санкт-Петербург
e.zamiraylova@gmail.com

В статье описываются основные трудности, которые возникают при создании систем машинного перевода на примере русского и английского языков.

Ключевые слова: машинный перевод, система машинного перевода.

Чтобы воспроизвести коммуникативное воздействие и добиться того, чтобы результат перевода имел такое же впечатление, какое имеет изначальный документ, следует соблюдать нормы и условия, предъявляемые к переводу. Разработка систем машинного перевода выдвинула перед переводчиками и лингвистами ряд трудностей, относящихся как к самому переводу, так и к анализу документа: неоднозначности, грамматические и лексические языковые различия, устойчивые фразеологические конструкции.

В зависимости от стиля и области, к которой относится текст, слово часто имеет несколько значений. Например, в английском языке слово *well* может переводиться как *колодец*, а может – как *скважина*, если это текст по нефтегазовой тематике. Или еще один пример: англ. глагол *to get* имеет большое количество значений: *дать, взять, сделать* и т. д. Только полностью понимая описываемую ситуацию, можно правильно интерпретировать смысл. Для систем автоматического перевода это становится проблемой из-за отсутствия однознач-

ной определенности. К лексической неоднозначности относится и омонимия, например, *лук* (оружие) и *лук* (овощ) пишутся одинаково, но имеют разные значения. Фраза *купи лук* допускает двойную интерпретацию, поэтому автоматически, без уточнения, ее перевести сложно. В то же время человек, услышавший эту фразу, скорее выберет вариант покупки овоща, если до этого никак не упоминалось оружие или реконструкция. Из этого следует, что отсутствие знаний о картине мира мешает машине сделать правильный перевод. Выходом из данной ситуации являются статистические данные о частоте использования той или иной лексемы в языке.

Если же в предложении можно выделить несколько разных структур, то это свидетельствует о структурной неоднозначности. Например, *Миша встретил ее в двери*. При переводе могут возникнуть две ситуации. Первый вариант может звучать, что *Миша встретил ее в двери*, где дверь – место. А во втором варианте, что встретил ее в двери, по аналогии – *Миша встретил ее в пальто*. Машинный перевод может трактовать ошибочно. Одна и та же фраза может быть интерпретирована по-разному даже при однозначности всех ее словоформ, потому что ситуация, когда у предложения есть несколько значений, встречается довольно часто. Затрудняет перевод и то, что количество слов с неоднозначной трактовкой может увеличиваться. Например, если в предложении только два слова, и у каждого из них есть по паре значений, то возможно перевести как минимум четырьмя способами. Логично что, может потребоваться рассмотрение всех возможных случаев перевода, чтобы выбрать верный вариант. Сложные случаи неоднозначности могут быть связаны с конверсией, когда получается незавершенный текст, недосказанный. Например, в русском языке наиболее частотным является переход из прилагательного в существительное, например, «мороженое».

Грамматические различия в языках (оригинал-перевод) объясняют основные характерные расхождения в переводческих правилах для данных языков. Потребность фиксирования этих особенностей была замечена при составлении правил перевода с английского языка на русский язык, так как в английском не существуют некоторые формы русского языка, поэтому намного меньше представлено согласование как грамматическая связь между частями речи в предложении. Например, одно из главных отличий то, что порядок слов в английском языке определяет часть речи, например, существительные переходят в глаголы. Например*, I milk a cat. – Я пою молоком кота* (авторский перевод). Или морфологическая неоднозначность

при совпадении некоторых форм, например, числительное «три и «три», форма повелительного наклонения единственного числа глагола тереть. Все перечисленное выше осложняет распознавание текста системам машинного перевода.

Также сюда относится грамматическая омонимия, когда совпадают формы одного слова. Например, совпадение форм падежей в русском языке (*живая речь, слушаю речь* – совпадение именительного и родительного падежа). Омографы, например, *плачУ* или *плАчу*. В отдельной фразе *я плачу* машине сложно интерпретировать верно. Такие параметры как интонация и ударение отсутствуют в написанных текстах. Одна из главных проблем синтаксического анализа связана с неоднозначностью языковых единиц. Например, слово *журнал*, может иметь другое значение в особых условиях, например, *журналу нельзя верить*. Журнал сравнивается с человеком, и это нехарактерно для данного слова. В машину сложно «заложить» атипичные ситуации, которые легко распознаются человеком. Помимо этого, большое количество комбинаций могут создать такие стилистические фигуры как грамматические эллипсы – пропуски словоформ. Например, *Татьяна – в лес. Медведь – за ней*. Чтобы обнаружить эллипс, где он действительно есть, программе нужно предположить подобные пропуски везде и сделать большое количество вычислений.

Сложности для машинного перевода вызывают фразеологические выражения и идиомы. Например, в произведении «Трое в лодке, не считая собаки» Джерома К. Джерома, фразеологизм «Dog's day of summer», который означает *самые жарки дни лета*. Машина может ошибочно перевести *собачьи дни лета*, что может помешать адекватно воспринять текст. В русском языке *собачьи дни лета* будут иметь отрицательную коннотацию, никак не связанную с теплой погодой. В одном издательстве переводчики опустили дословный перевод и написали *жаркие дни* (обобщив), но можно использовать развернутый описательный перевод, например, *самые жаркие дни, когда на восходе солнца видно собачью звезду, Сириус* (авторский перевод).

Идиомы можно сравнить с языковыми пробелами. Их следует выявлять в самом начале переводческого процесса, чтобы миновать их отсутствие, системой обрабатывать как одну единицу. Сложность появляется в вопросе соотнесения идиомы к слову, по которому будет осуществляться ее поиск. Проблема состоит в том, что они могут быть внесены в нескольких вариантах и часто алогично

основным грамматическим правилам. И еще трудности могут возникнуть, если идиома или фразеологизм относится только к какому-то региону и мало известна, и если человек может подобрать по контексту приблизительное значение, обобщив, то машинный перевод будет буквальным и непонятным.

Изучив сложности машинного перевода, можно сделать следующие выводы. На текущий момент остаются нерешенными многие основные переводческие проблемы на разных уровнях. Как видно из выше приведенного текста, системы машинного перевода пока не могут целиком и полностью обработать все смысловые значения исходного текста. Практические проблемы, появляющиеся при выполнении машинного перевода разнообразные и сложные. Тем ни менее все проблемы должны лишь стимулировать интерес к автоматизации переводческих задач. С каждым годом технологические возможности совершенствуются и постепенно улучшается качество машинного перевода. Главная цель – разработать компьютерное оборудование, перевод которого выдавал бы достаточно эквивалентный результат, которому редактирование потребуется в самой минимальной степени. А пока большинству текстов, переведенных автоматическим способом, еще нужны экспертные правки и корректировки. Самым приоритетным направлением развития систем машинного перевода является создание компьютерных систем, полностью моделирующих языковую картину мира, усовершенствование грамматического и синтаксического анализов.

ЛИТЕРАТУРА

1. Комиссаров В.Н. Теория перевода (лингвистические аспекты) М.: Высшая школа, 1990.

2. Маслов Ю. С. Введение в языкознание. М., 2005.

3. Миньяр-Белоручев Р.К. Общая теория перевода и устный перевод. М., 1980.

4. Николаев И.С., Митренина О.В., Ландо Т.М.(ред.) Прикладная и компьютерная лингвистика. М.,2017.

5. Соссюр Ф. де. Курс общей лингвистики Е.,1999.

Электронные ресурсы:

www.rsl.ru – Российская государственная библиотека

www.elibrary.ru – научная электронная библиотека eLIBRARY

www.pu.ru/library – научная библиотека СПбГУ

https://www.multitran.ru/ – Электронный словарь Мультитран

https://news.nationalgeographic.com – Электронный журнал National Geographic

УДК 81.33

TRANSFORMATION OF THE NARRATIVE UNDER THE INFLUENCE OF ALCOHOL

Y. S. Kolesnikova
National Research University Higher School of Economics, Moscow
2080508@mail.ru

The article studies the transformation of the narrative under the influence of alcohol based on the Corpus of Russian Oral Speech.

Keywords: corpus, oral speech, narrative, tale.

ТРАНСФОРМАЦИЯ НАРРАТИВА ПОД ВЛИЯНИЕМ АЛКОГОЛЯ

Я. С. Колесникова
Национальный исследовательский университет «Высшая школа экономики», Москва
2080508@mail.ru

В статье проводится исследование трансформации нарратива под влиянием алкоголя на базе созданного корпуса русской устной речи.

Ключевые слова: корпус, устная речь, нарратив, сказка.

Речь – одна из важнейших составляющих жизни человека, она тесно связана с процессом коммуникации. Существует множество факторов, способных деструктивно влиять на речевые процессы и речевой акт, одним из таких факторов является алкоголь. Алкогольная зависимость и ее воздействие на организм широко исследуются за рубежом [9, 10, 12], однако в России экспериментов, связанных с трансформацией речи под влиянием алкоголя сравнительно мало.

Данная работа посвящена изучению спонтанной речи, представленной в виде нарратива, и пересказа. В качестве объекта исследования выступает собранный корпус русской устной речи, в который вошли рассказы группы людей о жизненных ситуациях (негативного и позитивного опыта) и сказки, основанные на визуальном стимуле – книге с картинками Мерсера Майера «Frog, where are you?»[1]. В рамках эксперимента проверяются гипотезы, основанные на стереотипных ожиданиях и общих представлениях в обществе о людях, находящихся в нетрезвом состоянии: принято считать, что

[1] Mayer M. et al. Frog, where are you? – New York: Dial Press, 1969.

речь человека в состоянии интоксикации характеризуется потерей беглости, медлительностью, обрывистостью, большим количеством ошибок.

В эксперименте приняли участие пять человек, четверо студентов и один взрослый. Каждый из них рассказал две истории о личном опыте и сказку, находясь в нормальном и состоянии алкогольного опьянения. Таким образом, корпус насчитывает тридцать записей, двадцать нарративов и десять сказок. Перед началом эксперимента информантам предлагалось заполнить документ о согласии на обработку персональных данных и ознакомиться с правилами проведения эксперимента. Для сбора данных использовался диктофон, общение с информантами велось тет-а-тет в непринужденной обстановке у них дома. Участники были случайно разделены на две группы: группу А (Аня и Миша) и группу Б (Алиса, Сюзанна, Римма), где первая изначально записывалась в состоянии алкогольного опьянения, а через две недели в нормальном состоянии, а вторая в первый раз была записана в нормальном состоянии, а последующая запись проводилась, когда они находились в состоянии интоксикации. Степень опьянения у всех информантов была легкая, они принимали бокал красного полусухого вина. Эксперимент делился на два этапа: первый этап заключался в сборе данных, вторым этапом стала обработка записей и их транскрибирование. Для анализа данных использовались различные стратегии: сопоставление количества филлеров, вариантов исправления собственных ошибок, наблюдение за образностью и ассоциативностью речи, темпом речи в двух состояниях.

Нарратив

В спонтанной речи возникает потребность в выборе подходящих слов для выражения мысли, соблюдении логики повествования, однако совершение ошибок практически неизбежно, в результате чего говорящему приходится исправлять себя, в его речи появляется самокоррекция [13]. На основе корпуса устной речи можно составить классификацию стратегий, к которым участники прибегают для исправления ошибок в речи: замена целого слова (A), замена слова в процессе произнесения (B), повтор целого слова (C), повтор первой части слова (D), слияние слова с последующим высказыванием (E), исправление грамматической ошибки (F), сложность в выборе правильного слова с заполнением паузы филлером (G) и новый старт (H). Результаты анализа речи представлены в таблице 1.

Таблица 1. **Исправления в нарративах**

Состояние алкогольного опьянения		Нормальное состояние	
Позитивный опыт	Негативный опыт	Позитивный опыт	Негативный опыт
это-это (C)	неужели-неужели (C)	я чувство-я понимала (A)	ни-не калеченная (F)
в-вот (D)	чт-тогда (B)	че-школа четырнадцать (E)	мы-мы (C)
предст-продолжаю (B)	кор-этот-ээ-оранжевый (G)	уч-на учебу (D)	у-уже (D)
ты-ты (C)	разби-разбираться (D)	м-немецкого (B)	никаких-все магазины закрыты-никаких (H)
повли-поменяли (B)	с-с (D)		в-в-вообще (D)
пр-тоже (B)	она-они (F)		в-и пока (B)
м-мое (D)	меня-меня (C)		то-то (C)
находятся в-на (B)	я-я (C)		
дыа-на данный (E)	все-все (C)		
реально-действительно (A)	п-закричать (B)		
	на-сзади нас (E)		
	оказалось только – оказалось только (C)		
	прово-проводятся (D)		
	и-и (C)		
	девушка-женщина-девушка (H)		
	м-на (B)		
	ждал-пережидал (G)		
	куда-куда он (C)		
	что-с-э-что с ним (G)		
	по-по темной (C)		
	без-без задней (C)		
	я живу-живу (C)		
	в-вообще (D)		

Используя корпус, мы можем посчитать популярность стратегий самокоррекции, их частоту применения в речи информантов (рис. 1).

На графике видно, что информанты более склонны повторять слово целиком, то есть пользоваться стратегией C, заменять слова в процессе произнесения (B) и повторять первую часть слова (D). Ошибок, нарушающих беглость речи, в речи участников, находящихся в состоянии алкогольного опьянения, действительно больше, чем в нормальном состоянии.

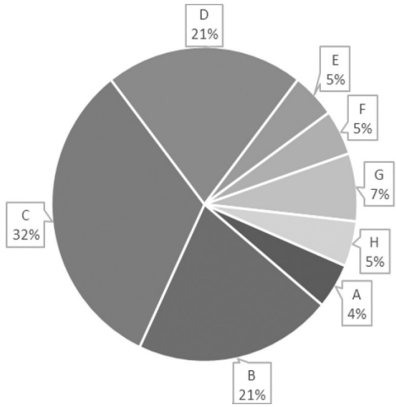


Рис. 1

Филлеры также часто используются в спонтанной речи, как и исправления. Они позволяют сократить пространство между словами, когда говорящий определяется, что сказать дальше [7]. Под филлерами понимаются слова, которые не имеют (или утратили) семантическую ценность, например, слово «короче» не говорит о том, что следующее высказывание будет коротко передавать всю суть, а обычно вводит новую мысль, помогая упорядочить повествование, или выступает в качестве «слова-паразита». В категорию филлеров включают «хм», «ам», «эм», «э», «а», которые сигнализируют о модификационных исправлениях и выражают хезитацию говорящего [11], слова типа «вообще», «в общем», «короче», «собственно», «получается», «во всяком случае», которые предполагают введение новой мысли, выражающие степень уверенности филлеры «конечно», «наверно»/«наверное», «казалось бы», филлеры «вот», «ну», «а», которые дают время на формулирование следующей мысли. Несмотря на полезность филлеров для говорящего, они ухудшают восприятие его речи слушающими [8].

Исходя из стереотипных ожиданий, можно предположить, что филлеров в рассказах будет много, поскольку людям в состоянии

алкогольного опьянения труднее выражать мысли. Однако результаты показывают обратное: в среднем все участники эксперимента чаще используют филлеры, находясь в нормальном состоянии (таблица 2). Все это может свидетельствовать в пользу предположения, что при приеме алкоголя концентрация внимания ухудшается [2] и человек хуже контролирует речевой поток, отвлекаясь на посторонние мысли и сиюминутные желания. Количество используемых филлеров снижается, поскольку говорящий меньше следит за грамматической составляющей языка и связностью идей, в большей степени опирается на передачу смысла сообщения.

Таблица 2. Филлеры в нарративах

Информант	Состояние алкогольного опьянения		Нормальное состояние	
	Позитивный опыт	Негативный опыт	Позитивный опыт	Негативный опыт
Сюзанна	*ну* x2	*ну*	*ну*	*ну* x6
	э	*вот*	*вот* x2	*а*
			э-м x2	*слава богу*
Результат	3 филлера на 115 слов Итог: 2,6%	2 филлера на 163 слова Итог: 1,23%	5 филлеров на 106 слов Итог: 4,72%	9 филлеров на 133 слова Итог: 6,77%
Алиса	*хм*	*ну* x8	*well*	*ммм* x2
	вот x3	*пожалуй*	*вот*	*в общем* x2
	наверное	*наверное*	*получается*	*вот* x6
	ну x3	*в общем* x2	*ам*	*ам* x2
	не знаю	*конечно*	*не знаю*	*эм* x2
	как бы x2	*не знаю*	*ну*	*всякое такое*
	там	*вот* x3	*собственно*	*получается*
		там x3	*да*	*и прочее*
		э	*и прочее* x2	что-то
		да		то есть
		как бы		ну
		получается x2		вроде как
Результат	12 филлеров на 300 слов Итог: 4%	26 филлеров на 392 слова Итог: 6,63%	10 филлеров на 165 слов Итог: 6%	22 филлера на 142 слова Итог: 9%

Римма	*ну* х5	*наверное* х3	*м*	*ладно*
	наверное	*вот* х3	*да* х2	*да*
	наверно	*ну* х2	*конечно*	*типа того*
	вот х3	*при чем* х2	*эм*	*как бы*
	честно говоря	*вообще*	*типа*	
		конечно	*наверное*	
		прям	*все дела*	
			вот	
Результат	11 филлеров на 110 слов Итог: 10%	13 филлеров на 129 слова Итог: 10%	9 филлеров на 55 слов Итог: 16,4%	4 филлера на 63 слова Итог: 6,35%
Миша	*в общем*	*в общем* х6	*опять же*	*в общем* х3
	э	*э* х7	*э* х6	*э* х7
	м х3	*м* х3	*м* х4	*м* х2
	ну х2	*ну* х5	*ну* х2	*ну* х2
	да х2	*видимо*	*а*	*а*
	то есть	*то есть* х2	*то есть* х5	*и* х3
	собственно	*собственно* х3		*собственно* х4
	конечно	*конечно*		*и так далее*
	во всяком случае	*казалось бы*		*там*
		как бы		*типа*
		получается		*короче*
		вот х6		*вот* х3
		наверное		
		наверно		
Результат	13 филлеров на 265 слов Итог: 5%	39 филлеров на 668 слов Итог: 6%	19 филлеров на 234 слова Итог: 8,12%	31 филлер на 331 слово Итог: 9,34%
Аня	*в общем* х2	*м* х6	*м* х2	*так*
	агр	*в общем*	*вот*	*э*
	вот		*мн* х2	*вот* х3
	ну			*ну*
	и			*ау*
Результат	6 филлеров на 82 слова Итог: 7,3%	7 филлеров на 87 слов Итог: 8%	5 филлеров на 41 слово Итог: 12%	7 филлеров на 55 слов Итог: 13%

Участники, находящиеся в состоянии алкогольного опьянения, также более склонны к ассоциативному мышлению, что выражается в добавлении нерелевантной для данного нарратива информации. Подобные добавления рассеивают внимание слушателей, отвлекают от основной мысли повествования. Идеи сменяют друг друга, один воображаемый образ заменяется другим по принципу ассоциации (таблица 3), из-за чего повествование становится прерывистым, тяжело воспринимаемым, возрастает темп речи, рассказ становится ориентированным на описание, нежели повествование.

Таблица 3. **Дополнительная информация**

Алиса	И вот… я с тех пор*, просто капец, меня в машине укачивает,* я дорог дико боюсь
Миша	В общем, это было в классе третьем-четвертом, не помню, мы с… моим другом на тот момент… Глебом – *очень хорошо с ним общались* – м, катались на санках.
Миша	Вверх по горке, то есть, там дорога и справа тротуар, но тоже в гору – *гора градусов, наверное, ну, под 35*, вот, вся улица так идет, Цандара, на которой я живу.

Сказка

Для того, чтобы участвовать в языковой коммуникации, человеку необходимо владеть речевыми жанрами, принятыми в обществе, знать их особенности, чтобы выстроить свое высказывание в их рамках [1]. Например, чтобы пожаловаться, нужно знать, как жалуются в этом языке и в этом обществе, какую эмоцию должен испытывать говорящий, какой тон и выражения использовать, то есть, восхваление предмета уже не будет являться жалобой. Сказка как устный жанр также имеет определенную структуру: в ней должна присутствовать фигуры сказителя и аудитории, определенные формулы и средства, такие как устойчивые выражения типа «гуси-лебеди», инверсия, лимитация времени для обозначения начала и конца истории [3].

Первое, что нарушает канон жанра сказки, это появление большого количества филлеров в речи. И участники эксперимента нарушают повествование филлерами, находясь в состоянии алкогольно-

го опьянения, гораздо чаще, чем в нормальном состоянии (процент филлеров в речи увеличивается). Например, Сюзанна, у которой есть ребенок, достаточно опытный сказитель: ее речь обладает соответствующей интонацией, будто она рассказывает сказку своему ребенку, отсутствуют долгие паузы и хезитации, которые бы были заполнены филлерами. Однако в состоянии алкогольного опьянения навык сказительства ухудшается, мысль часто теряется, и Сюзанна заполняет разрывы протяжными «эм», «а», «ну», «вот» и другими филлерами (17 филлеров на 416 слов, 4% филлеров).

Таблица 4. **Филлеры Сюзанны**

Состояние алкогольного опьянения	Нормальное состояние
вот x4	(филлеры отсутствуют)
так	
там	
в общем x2	
ну x4	
блин x2	
а	
эм	
опять-таки	

Второе, что недопустимо для детской сказки, это использование матерных конструкций, а также сниженной лексики, которые нарушают рамки жанра. Таблица 5 иллюстрирует примеры бранной и сниженной лексики, которую употребили участники эксперимента[1]. Поиск альтернативных им выражений, вероятно, потребовало бы больших усилий, времени подумать, а ненормативная лексика и филлеры облегчали участникам задачу, сокращая разрывы между словами, упрощая речь, но нарушая установленные жанром детской сказки рамки.

[1] Классификация слов была составлена с помощью данных из словарей русского языка; Ефремова Т. Ф. Современный толковый словарь русского языка: В 3 т. – М.: АСТ, Астрель, Харвест, 2006 и Шагалова Е. Н. Самый новейший толковый словарь русского языка XXI века: около 1500 слов. – М.: АСТ, Астрель, 2012.

Таблица 5. **Бранная и сниженная лексика**

Состояние алкогольного опьянения	Нормальное состояние
че x10 (разговорное)	*эпик* (сленговое)
орешь x3 (разговорное)	
придурок (бранное)	
блин x4 (эвфемизм)	
попутал (разговорное)	
отплевываясь (разговорное)	
пацан x7 (разговорное)	
башкой (разговорное)	
дурацкую x2 (разговорное)	
ваще (разговорное)	
агрятся (сленговое)	
епт (обсценное)	
ща (разговорное)	
херня x2 (вульгарное)	
спокойненько (разговорное)	
стремное (жаргонное)	
фигом (вульгарное)	*капец* (жаргонное)
мощный (разговорное)	
тусуют (жаргонное)	
классненько (разговорное)	
телку (жаргонное)	
красава (сленговое)	
съебал (обсценное)	
нормас (сленговое)	
хавчика (жаргонное)	
капец (жаргонное)	
суицидника x2 (разговорное)	
свезло (разговорное)	
рандомно (сленговое)	
четенько (сленговое)	
свалил (сленговое)	
бабла (жаргонное)	

С употреблением алкоголя менялась и интонация: если в нормальном состоянии говорящие меняли темп, эмоциональную окраску, делая речь более выразительной, то в состоянии алкогольного опьянения речь становилась невыразительной, монотонной, будто у говорящих пропадал интерес к процессу, им было сложнее выговаривать слова.

Другим нетипичным явлением для сказки стал анализ участниками происходящего. Внутри сказки существует своя схема построения, канонизированные герои и общие сюжеты, однако сказка сама по себе не претендует на реальность происходящего и не скрывает, что содержит вымысел. Сказитель знает, что сказочные события выдуманы и лишь являются частью общей истории, поэтому «нелогичность связи» его не останавливает и не сбивает во время рассказа [5]. Но во время пересказа истории про лягушку возникла одна особенность: участники эксперимента пытались искать логику в сюжете и критиковать происходящее, они делали комментарии, выражающие свою позицию по отношению к событиям сказки, или же разрушали жанр, проводя связи с реальной жизнью (таблица 6).

Таблица 6. **Анализ сказки**

Сюзанна	Даже собака умудрилась упасть с балкона. При чем, опять-таки с мордой в банке – *это ужас какой-то, бедная собака*.
Алиса	Вот, *странный*, конечно, *придурок*, ну в общем, соль в том, что… пацан спит с собакой в-на кровати.
Римма	По имени Боб. *Боб – это из Куплинова взято*.
Миша	Быстро одевается, бежит ее искать, выглядывает в окно и орет: «Лягушка-а-а-а!». *Зачем он орал – не понятно, лягушка ж не понимает русского языка.*
Миша	Вот, а потом *по непонятной причине* она взяла и, короче, вместе с банкой на голове прыгнула вниз, ну, *она на суицидника не была похожа*… значит, *видимо, случайно упала.*

Фольклорная традиция как народное творчество подразумевает отсутствие автора, а подобные действия включают сказителя в рассказ, нарушая рамки жанра. Сказитель может выступать только как пассивный участник сказки в формулах «И я там был, мед, пиво пил» [4].

Поиск логики в событиях сказки противоречит ее жанровым особенностям, нарушает целостность истории и ухудшает ее восприятие аудиторией. И причина этих изменений – алкоголь. Проанализировав разницу между текстами, записанными в двух состояниях, их особенности и различия, количество содержания в них элементов, деструктивно влияющих на целостность сказки, можно заключить, что существует прямая корреляция между состоянием человека и владением первичным жанром.

В связи с тем, что в основе сказки лежит коммуникация между сказителем и аудиторией, по возможности рассказывать историю можно судить о способностях человека общаться с собеседником. Выход за пределы рамок первичного жанра превращает сказку в нечто иное, затрудняя восприятие аудитории, разрывая связь сказитель-слушатель, мешая человеку, употребившему алкоголь, понимать других людей и быть понятым.

Проанализировав данные корпуса устной речи, с уверенностью можно говорить о деструктивном влиянии алкоголя на человека. После принятия алкоголя ассоциативное мышление ускоряется, индивид в меньшей степени контролирует речевой поток, больше концентрируясь на донесении смысла, нежели форме высказывания. Количество используемых филлеров сокращается, убирая необходимое расстояние между высказываниями, интонация становится монотонной. Речь лишается образности, выразительности, чаще используется сниженная и ненормативная лексика. В сказке человек в состоянии алкогольного опьянения также использует меньше изобразительно-выразительных средств, меньше формул, характерных для фольклорной традиции, что говорит о потере его навыка владения первичными жанрами. Несостоятельность человека как сказителя разрывает отношения рассказчик-слушатель, адресант-адресат [6], переставая пользоваться ими в повседневной жизни для выражения своих чувств, человек рискует быть непонятым окружающими.

ЛИТЕРАТУРА

1. Бахтин М. М. Проблема речевых жанров; Проблема текста в лингвистике, филологии и других гуманитарных науках // Коммуникативные стратегии культуры: Хрестоматия к курсу «Введение в теорию коммуникации». Новосибирск, 2003.

2. Вэлком М. О. Состояние высших интегративных функций мозга и уровень гликемии у молодых людей, употребляющих алкогольные напитки. – 2013.

3. Герасимова Н. М. Прагматика текста: фольклор, литература, культура. – 2012.

4. Лорд А.Б. Сказитель. М.: Восточная литература, 1994.

5. Пропп, В.Я. Морфология волшебной сказки. Исторические корни волшебной сказки. М.: Лабиринт, 1998.

6. Якобсон Р. Лингвистика и поэтика // Структурализм: «за» и «против». М. – 1975. – Т. 10.

7. Amiridze N., Davis B. H., Maclagan M. (ed.). Fillers, pauses and placeholders // John Benjamins Publishing. – 2010. – Vol. 93.

8. Brennan S. E., Schober M. F. How listeners compensate for disfluencies in spontaneous speech // Journal of Memory and Language. – 2001. – Vol. 44 (2). – P. 274–296.

9. Church M. W. et al. Hearing, language, speech, vestibular, and dentofacial disorders in fetal alcohol syndrome // Alcoholism: Clinical and Experimental Research. – 1997. – Vol. 21 (2). – P. 227–237.

10. Cone-Wesson B. Prenatal alcohol and cocaine exposure: influences on cognition, speech, language, and hearing // Journal of communication disorders. – 2005. – Vol. 38 (4). – P. 279–302.

11. Corley M., Stewart O. W. Hesitation disfluencies in spontaneous speech: The meaning of um // Language and Linguistics Compass. – 2008. – Vol. 2 (4). – P. 589–602.

12. Pisoni D. B., Martin C. S. Effects of alcohol on the acoustic-phonetic properties of speech: perceptual and acoustic analyses // Alcoholism: Clinical and Experimental Research. – 1989. – Vol. 13 (4). – P. 577–587.

13. Schegloff E. A., Jefferson G., Sacks H. The preference for self-correction in the organization of repair in conversation // Language. – 1977. – Vol. 53 (2). – P. 361–382.

УДК 004.89

AUTOMATIC EXTRACTION OF IMPLICIT ATTITUDES FROM TEXTS

N. V. Loukachevitch[1], V. A. Karnaukhova[1], N. L. Rusnachenko[2]
[1]Lomonosov MSU, Moscow
[2]Bauman Moscow Technical University, Moscow
louk_nat@mail.ru, ssnicker@yandex.ru, kolyarus@yandex.ru

The paper describes the task of automatic extraction of attitudes between the subjects mentioned in the text, as well as their connection with the implicit expression of the author's attitude to these subjects. A vocabulary of sentiment frames RuSentiFrames is presented, in which the basic attitudes associated of Russian predicate words are described.

Keywords: sentiment analysis, sentiment attitudes, sentiment frame, Russian language.

АВТОМАТИЧЕСКОЕ ИЗВЛЕЧЕНИЕ ИМПЛИЦИТНЫХ ОЦЕНОК ИЗ ТЕКСТОВ

Н. В. Лукашевич[1], В. А. Карнаухова[1], Н. Л. Русначенко[2]
[1]МГУ имени М. В. Ломоносова, Москва
[2]МГТУ имени Н. Э. Баумана, Москва
louk_nat@mail.ru, ssnicker@yandex.ru, kolyarus@yandex.ru

В статье описана задача автоматического извлечения оценочных отношений между субъектами, упоминаемыми в тексте, а также ее связь с неявным выражением оценочного отношения автора к этим субъектам. Представлен словарь оценочных фреймов RuSentiFrames, в котором для слов-предикатов русского языка зафиксированы основные оценочные отношения, связанные с употреблением этих слов в тексте.

Ключевые слова: анализ тональности, оценочные отношения, оценочные фреймы, русский язык.

1. Введение

Автоматический анализ тональности является одним из активно исследуемых проблем в сфере автоматической обработки текстов. Основными методами в этой задаче являются инженерный подход и подход на основе машинного обучения [2, 10]. Инженерный под-

ход использует словари оценочных слов и выражений, в которых каждому слову поставлен в соответствие вес позитивности или негативности этого слова. Подход на основе машинного обучения использует обучающую выборку, в которой размечены тональности отдельных выражений, предложений или текстов в целом. При этом в качестве представления текста используется представление типа мешок слов или совокупность векторных представлений (embeddings) слов.

Вместе с тем в автоматическом анализе тональности большое значение имеют не только отдельные слова и выражения, но и их объединение в синтаксические структуры, для которых должна выводиться итоговая оценка. В инженерном подходе для этого используются правила, из которых наиболее известными являются правила изменения тональности на противоположную (как в случае отрицания), или усиления тональности (как при использовании конструкции со словом «очень»). В подходе на основе машинного обучения для учета этой проблемы могут использоваться представления в виде биграмм или подходы на основе нейронных сетей, которые могут анализировать последовательности слов в предложении.

В автоматическом анализе тональности есть много трудностей, наиболее известными из которых являются ирония или сарказм. Другим видом сложных случаев анализа тональности являются разнообразные способы выразить неявные оценки. Например, в предложении «Генсек ООН осудил бомбардировку мирных жителей» очевидно, что автор текста относится к генсеку ООН положительно, хотя формально рядом с выражением *Генсек ООН* находятся слова с негативными коннотациями. Из предложения «Президент Ирана Хасан Рухани **осудил бомбардировку** Сирии» можно предположить, что президент Ирана позитивно относится к Сирии, но, как относится автор текста к президенту Ирана, не очень понятно. Видно, что оба типа применяемых методов могут иметь проблемы с выводом таких оценок.

В данной статье рассматривается задача извлечения из текста оценочных отношений между упоминаемыми сущностями, которые часто используются и для выражения имплицитных оценок автора. Также мы опишем структуру нового лингвистического ресурса RuSentiFrames, в котором описываются оценочные фреймы для слов русского языка, назначение которых – использование в распознавании оценочных отношений.

2. Обзор близких работ

В настоящее время для разных языков создано большое количество словарей оценочной лексики, которые представляют из себя список слов и выражений с приписанными оценками их полярности, а также, возможно, оценками силы этих полярностей [3, 16]. Для русского языка также создано и опубликовано несколько словарей такого типа [4, 9, 11]. Для работы с оценочными отношениями нужны специализированные словари, которые описывают тональности, связанные с аргументами слов-предикатов

Предложенный лексикон [13] состоит из блока субъективных слов, служебных слов, модификаторов и модальных операторов. Ядро лексикона составляет SentiFul Database, 10657 слов с размеченной полярностью (полярность оценивалась от 0 до 1) и весами полярности. В соответствии с оценочной теорией (appraisal theory), Sentiful Database была расширена такими типами эмоции, как аффект, суждение и признание.

В блок служебных слов системы вошли:

– реверсивные прилагательные, существительные и глаголы (то есть такие слова, которые меняют полярность слова, к которому они относятся, на противоположную: *ограничивать, прекращать*);

– глаголы распространения полярности: распространяют полярность на свои аргументы;

– глаголы переноса: переносят полярность на аргумент другого типа (например, полярность объекта на полярность субъекта).

В работе Rashkin et al. [15] рассматриваются так называемые фреймы коннотаций [6] для глаголов (connotation frames), в которых записываются предположения о сущностях *X* и *Y*, когда они употребляются с глаголом V (*X V Y*): включая, как автор относится к *X* и как к *Y*, как *X* и *Y* относятся друг к другу, улучшается или ухудшается состояние *X* и *Y* после действия *V*. В целом рассматриваются следующие отношения:

1. Тональности (от автора к сущностям, от читателя к сущностям, между сущностями)
2. Ценность (сущности)
3. Эффект (т.е. влияние, которое оказывается на сущность)
4. Внутреннее состояние (сущности после произведенного действия).

Будучи зафиксированными, таким предположения дают возможность лучше оценивать отношение автора к упоминаемым сущ-

ностям, а также их отношения между собой. В рамках работы 900 наиболее частотных переходных глаголов английского языка были размечены по несколько предложений с помощью краудсорсинга.

Klenner et al. [7, 8] описывают оценочные фреймы для немецких глаголов. Каждый фрейм состоит из совокупности ролей, ассоциированных с глаголов, полярности, а также (положительных или отрицательных) эффектов, связанных с ролями. Также описывается так называемая сигнатура глагола. Сигнатура указывает на фактуальность ролей в зависимости от различных факторов (таких как отрицание, настроение и т. д.). Например, если глагол «verhindern» (затруднять) не отрицается в предложении, то его объект не находится в фактическом состоянии, и его возможная полярность не должна учитываться.

Dang и Wiebe [5] рассматривают события, которые положительно или отрицательно влияют на сущности (goodFor / badFor). Например, *снижение X* плохо для *X*, но *создать X* хорошо для *X*. В этой статье рассматривается вывод тональности, когда тональность выражается по отношению к плохим или хорошим событиям.

Для русского языка в работе [14] авторы представляют достаточно подробную классификацию лексики для задачи определения тональности. В частности, большое внимание уделяется классификации глаголов, которые делятся на восемь классов. Например, «3 и 4 класс – негативные и позитивные глаголы, определяющие тональность объекта независимо от окружения, но в зависимости от его роли (например, глаголы «сдаться» и «проиграть» приписывают негатив субъекту и позитив объекту, а глаголы «обуздать» и «повергнуть», наоборот, приписывают позитив субъекту и негатив объекту)». После распознавания тональности отдельных слов в предложении, для определения тональности предложения применяются ряд правил сочетаемости тональности. Правила представляют собой комбинации различных членов предложения между собой.

Отметим, что в данном подходе [14] не различаются негативные (или позитивные) явления и отношение к ним автора или другого участника ситуации. Например, предложение «Во Франции произошел теракт» упомянуто негативное событие, но это не значит, что есть негативное отношение к Франции, которой, наоборот, была выражена поддержка и соболезнование в подавляющем большинстве новостей, сообщающих о теракте.

В работе [12] создана коллекция аналитических статей в области международной политики, которая размечена оценочными отноше-

ниями между именованными сущностями, упомянутыми в тексте. Были протестированы методы машинного обучения, которые показали низкое качество извлечения отношений. Также были выявлены проблемы с экспертной разметкой, которая показала около 0.55 F-меры, если рассматривать разметку одного эксперта как идеальную, а другую сравнивать с этой идеальной разметкой.

3. Оценочные фреймы для русского языка RuSentiFrames

Для более точного извлечения оценочных позиций, цитируемых или выраженных в текстах авторами, недостаточно иметь словарь оценочной лексики, в котором проставлены только оценки тональности для этого слова. Анализируя слова-предикаты, которые ссылаются на некоторую ситуацию с несколькими участниками, необходимо различать оценочные позиции (мнения) автора, участников ситуации, описываемой словом, последствия ситуации для участников, которые отличаются от субъективных мнений.

Для работы с более точным описанием оценок слов-предикатов создается словарь оценочных фреймов для русского языка RuSentiFrames. Особенностями данного ресурса являются представление в фреймовом виде следующей оценочной информации, ассоциируемой с заданным словом или выражением:

– позитивное или негативное отношение автора к участникам ситуации, выражаемыми словом (слот polarity),

– позитивное или негативное отношение между участниками ситуации, выражаемыми словом (слот polarity),

– ценность одного из участников ситуации для другого участника, например, когда участник A1 преследует другого участника A2, то мы не можем утверждать, позитивно или негативно относится A1 к A2, но в любом случае A2 представляет собой некоторую ценность для A1 (слот value),

– негативные или позитивные последствия для участников ситуации (слот effect),

– внутреннее состояние участников ситуации (слот state).

Слоты представляют информацию о негативной или позитивной тональности. Если тональность может быть любой или нейтральной, то такие слоты в фрейм не вносятся.

В каждом фрейме перечисляются роли участников в ситуации, которые задействованы в фрейме. Роли обозначаются нумерацией A_1, A_2 и т.д., где A_1 соответствует основному участнику ситуации,

который обычно выражается подлежащим. A_2 соответствует второму участнику ситуации, который выражается дополнением. В фрейме есть отдельный раздел, в котором кратко поясняется, что означает каждая роль этого фрейма.

Явным образом указывается только позитивные или негативные значения. Каждый слот имеет оценки уверенности эксперта. В настоящее время имеется две оценки уверенности 1 (уверен) и 0.7 (по умолчанию).

На первом этапе для каждого слова создавался отдельный фрейм, однако много близких по смыслу слов могут иметь похожие оценочные фреймы. Поэтому сейчас с каждым фреймом связывается совокупность слов и выражений. В отличие от других проектов к фреймам приписываются не только глаголы, но и слова других частей речи, а также важным является описание устойчивых словосочетаний, которые могут иметь другие оценочные фреймы, чем их слова-компоненты. В настоящее время группы слов и словосочетаний, связанные с конкретным оценочным фреймом или похожими оценочными фреймами создаются на основе синонимом, гипонимов и гиперонимов, описанных в тезаурусе русского языка РуТез [1].

Пример фрейма для группы слов со значением «убить» выглядит следующим образом:

- Title: убить
- Comments:
- Variants: Убить, убивать, убийство, лишить жизни, унести жизнь, сразить, сражать, уложить, уносить жизнь, поубивать, умертвить, умерщвлять, умерщвление, умертвлять, искоренить, искоренять, уничтожить, уничтожать, прекратить существование, искоренение, уничтожение .. колоть, заколоть, закалывать, зарубать, зарубить.. Распять, распинать, распятие, вздернуть на виселицу, расстреливать, расстрелять, повесить, вешать, повешение, повесить на виселице, вешать на виселице, расстрел
- Roles:
  A_1 – убийца
  A_2 – убитый
- Frame:
  Polarity (A_1, A_2, neg, 1)
  Polarity (A_2, A_1, neg, 1)
  Effect (A_2, -, 1)
  State (A_2, neg, 1)
  Polarity (author, A_1, neg, 0.7)

Фрагмент алфавитного словника фреймов выглядит следующим образом:

авианалет – фрейм АТАКОВАТЬ
авиаудар – фрейм АТАКОВАТЬ
авиационный налет – фрейм АТАКОВАТЬ
авиационный удар – фрейм АТАКОВАТЬ
агрессия – фрейм ВООРУЖЕННАЯ АГРЕССИЯ
агрессия против – фрейм ВООРУЖЕННАЯ АГРЕССИЯ
апеллировать – фрейм ЖАЛОВАТЬСЯ
апелляция – фрейм ЖАЛОВАТЬСЯ
аплодировать – фрейм ПООЩРИТЬ, ПОХВАЛИТЬ
арест – фрейм АРЕСТОВАТЬ
арест по подозрению – фрейм АРЕСТОВАТЬ
арестовать – фрейм АРЕСТОВАТЬ
арестовать по подозрению – фрейм АРЕСТОВАТЬ
арестовывать – фрейм АРЕСТОВАТЬ
арестовывать по подозрению – фрейм АРЕСТОВАТЬ и т. д.

В настоящее время оценочные фреймы записаны для 4000 слов и выражений русского языка, включая существительные, глаголы, именные и глагольные группы.

4. Эксперимент по оценке качества оценочных фреймов

Для проверки оценочных фреймов, написанных «из головы», и реального распределения тональностей в предложениях, двум экспертам было выдано 50 случайно выбранных слов, для которых уже были написаны фреймы. Слова были выданы вместе с их прописанными ролями, но с незаполненными тональностями. Эксперт должен был выполнить поиск по слову в новостях, отобрать 10 разных предложений, в которые входило слово, и заполнить экземпляр оценочного фрейма по конкретному предложению. Затем оценки, полученные по предложениям, были усреднены. Усредненный результат фрейма был сопоставлен с исходным фреймом. Такой же эксперимент с описанием фреймов по предложениям был выполнен для глаголов, которые часто упоминаются в текстах военной тематики, по материалам публикаций информационного агентства РИА Новости.

В результате сопоставления усредненных фреймов и исходных в нескольких случаях было выявлено присутствие еще одного актанта, который выражает оценочные отношения. Например, для слова *экстрадиция* изначально было описано в фрейме два актанта (*кто*

экстрадирует и *кого экстрадируют*). Но обнаружилось, что в текстах есть еще и третий актант: *куда экстрадируют*, и имеется негативное отношение между высылаемым и местом высылки. В целом, усредненный фрейм по 10 предложениям для слова *экстрадиция* оказался следующим:

Экстрадиция: Усредненный фрейм по предложения:
A_1 – кто, A_2 – кого, A_3 – куда.
Polarity (A_1, A_2, neg, -0.7)
Polarity (A_2, A_1, neg, -0,7)
Polarity (A_2, A_3, neg, -0.87)
Polarity (A_3, A_2, neg, -0.87)
Effect (A_2, -, 1)
State (A_2, neg, 1)

Наибольшее расхождение между описанными фреймами и усредненными фреймами было выявлено для слова *понять*. При описании фрейма в RuSentiFrames казалось, что понимание чего-либо – это позитивный процесс, и субъект положительно относится к тому, что он понял. Такая интерпретация привела к следующему фрейму для слова *понять*:

Понять: *оценочный фрейм в RuSentiFrames*
A_1 – тот, кто понимает, A_2 – то, что понимают/
Polarity (A_1, A_2, pos, 0.7)
Effect (A_2, +, 1)
State (A_1, pos, 1)
Polarity (author, A_1, pos, 0.7)

Однако на практике, в новостных текстах усредненный фрейм получился совсем другим:

Понять: Усредненный фрейм по предложениям
Polarity (A_1, A_2, neg, -0.6)
Polarity (author, A_1, pos, 0.34)
Polarity (author, A_2, neg, -0.77)
Effect (A_1, -, -0.2)
State (A_1, neg, -0.31)

По усредненному фрейму мы видим противоположную картину: все выглядит негативным. Это объясняется тем, что в текущих новостях чаще обсуждается что-нибудь плохое, и субъекту, упоминаемому в тексте, приходится понять, что ситуация плохая. Например, для предложения: *Как только садоводы поняли, что как такового урожая не будет, то на импровизированных рынках сразу поднялись цены*, эксперт записывает следующий фрейм:

Понять: фрейм для конкретного предложения
Polarity (A_1, A_2, neg, -1)
Polarity (author, A_1, neg, -0.7)
Polarity (author, A_2, neg, -1)
Effect (A_1, -, -0.7)
State (A_1, neg, -0.7).

т. е. попадание негативной ситуации в актант фрейма трансформирует все связанные с фреймом тональности. Таким образом, необходимо еще развитие специальных правил, которые трансформируют исходный фрейм в зависимости от появления в качестве его актантов явно позитивных или негативных участников:

Примером такого правила может быть следующее:

Polarity (A_1, A_2, neg, 1) & Polarity (Author, A_2, neg, 1)-> Polarity (Author, A_1, pos, 1).

т. е. если между двумя субъектами в тексте указано отношение, которое свидетельствует о негативном отношении первого субъекта ко второму субъекту, а в качестве второго субъекта употреблено явно отрицательное слово, то автор относится к первому субъекту положительно. Например, из предложения «*По словам источника, армия отбила все атаки, ликвидировав боевиков*», фрейма глагола *ликвидировать*, который говорит о негативном отношении субъекта ко второму участнику ситуации, и употребления явно отрицательного слово *боевиков* в качестве объекта глагола *ликвидировать* можно вывести, что автор положительно относится в данном случае к *армии*, хотя это явным образом и не сказано.

Также с помощью такого рода правил можно объяснить примеры, которые были приведены во введении. При анализе предложения «Генсек ООН осудил бомбардировку мирных жителей» нужно учесть, что мирное население является ценностью, которая разделяется подавляющим числом современных людей. Бомбардировка мирного населения разрушает ценности и в глазах многих людей является негативным событием. Генсек ООН осуждает это негативное событие, и поэтому его образ в этом предложении позитивный.

Во втором предложении в качестве объекта бомбардировки выступает именованная сущность *Сирия*, которая может подразумевать разные аспекты, включая текущий правящий режим, к которому у разных людей разное отношение. Поэтому, как автор или читатель отнесутся к Роухани, зависит от не определенных в предложении факторов, и никакого вывода мы сделать не можем.

Заключение

В данной статье мы рассмотрели задачу автоматического извлечения оценочных отношений между субъектами, упоминаемыми в тексте, а также ее связь с неявным выражением оценочного отношения автора к этим субъектам. Представлен словарь оценочных фреймов RuSentiFrames, в котором для слов-предикатов русского языка зафиксированы основные оценочные отношения, связанные с употреблением этих слов в тексте.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект 16-29-09606).

ЛИТЕРАТУРА

1. Лукашевич Н.В. Тезаурусы в задачах информационного поиска. Изд-во Московского университета, 2011.

2. Лукашевич Н.В. Автоматические методы анализа тональности // Большакова Е.И. и др. Автоматическая обработка текстов на естественном языке и анализ данных. НИУ ВШЭ, 2017.

3. Baccianella S., Esuli A., Sebastiani F. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Proceedings of LREC-2010, Vol. 10, P. 2200–2204.

4. Chetviorkin I., Loukachevitch N. Extraction of Russian Sentiment Lexicon for Product Meta-Domain. Proceedings of COLING-2012, P. 593–610.

5. Deng L., Wiebe J. Sentiment propagation via implicature constraints. Meeting of the European Chapter of the Association for Computational Linguistics EACL-2014. 2014.

6. Feng S., Kang J. S., Kuznetsova P., Choi Y. Connotation Lexicon: A Dash of Sentiment Beneath the Surface Meaning. In ACL (1), 2013. pp. 1774–1784).

7. Klenner M., Amsler M., Hollenstein N. Verb polarity frames: a new resource and its application in target-specific polarity classification. In G. Faaß (Ed.), KONVENS-2014, 2014. pp. 106–115.

8. Klenner M., Amsler M. Sentiframes: A resource for verb-centered German sentiment inference //Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016). 2016.

9. Kotelnikov E., Peskisheva T., Kotelnikova A., Razova E. A Comparative Study of Publicly Available Russian Sentiment Lexicons // In Conference on Artificial Intelligence and Natural Language, 2018. pp. 139–151.

10. Liu B., Zhang L. A survey of opinion mining and sentiment analysis // Mining Text Data. Springer: US, 2012. pp. 415–463.

11. Loukachevitch N., Levchik A. Creating a General Russian Sentiment Lexicon // In Proceedings of Language Resources and Evaluation Conference LREC-2016, 2016.

12. Loukachevitch N., Rusnachenko N. Extracting Sentiment attitudes from analytical texts // In Proceedings of Computational Linguistics and Intellectual Technologies, Papers from the Annual Conference Dialog-2018, 2018 pp. 459–468

13. Neviarouskaya A., Prendinger H., Ishizuka M. Semantically distinct verb classes involved in sentiment analysis //IADIS AC (1), 2009. pp. 27–35.

14. Pazelskaya A. G., Soloviev A.N. The method to extract emotions in Russian texts // Computational linguistics and intellectial technologies, 2011.

15. Rashkin H., Singh S., Choi Y. Connotation Frames: A Data driven Investigation // Proceedings of Association for Computational LInguistics Conference ACL-2016, 2016. pp. 311–322.

16. Wilson T., Wiebe J., Hoffmann P. Recognizing contextual polarity in phrase-level sentiment analysis // Proceedings of the conference on human language technology and empirical methods in natural language processing, 2005, pp. 347–354.

УДК 81'32

ON THE PRINCIPLES OF CREATION OF THE RUSSIAN SHORT STORIES CORPUS OF THE FIRST THIRD OF THE 20th CENTURY

G. Ya. Martynenko[1], T. Yu. Sherstinova[1,2], T. I. Popova[1], A. G. Melnik[1], E.V. Zamirajlova[1]
[1]*St. Petersburg State University, St. Petersburg*
[2]*National Research University Higher School of Economics*
g.martynenko@spbu.ru, t.sherstinova@spbu.ru,
st064458@student.spbu.ru, tipopova13@gmail.com, kate_zam@mail.ru

The paper describes the basic principles of creating a representative Russian short stories corpus, which should contain texts of the maximum number of Russian writers who wrote and published in this period. The corpus is being created as part of a research project aimed at studying the changes that occur in the language at the turning points of history. The First third of the 20th century turned into a whole series of social upheavals for Russia – the Russo-Japanese War, the First Russian Revolution, the First World War, the February and October Revolutions of 1917, the Civil War, the formation of a new Soviet state, the New Economic Policy, collectivization, and the beginning of industrialization. The development of a linguistic resource, including the works by a large number of authors for each time period, will provide an opportunity to conduct linguistic and statistical analysis of both language and style in synchrony and diachrony.

Keywords: corpus linguistics, Russian literature, Russian language, Russian short story, the beginning of the 20th century.

О ПРИНЦИПАХ СОЗДАНИЯ КОРПУСА РУССКОГО РАССКАЗА ПЕРВОЙ ТРЕТИ XX ВЕКА

Г. Я. Мартыненко[1], Т. Ю. Шерстинова[1,2], Т. И. Попова[1], А. Г. Мельник[1], Е. В. Замирайлова[1]
[1]*Санкт-Петербургский государственный университет, Санкт-Петербург*
[2]*Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург*
g.martynenko@spbu.ru, t.sherstinova@spbu.ru,
st064458@student.spbu.ru, tipopova13@gmail.com, kate_zam@mail.ru

В статье описываются основные принципы создания представительного корпуса русского рассказа 1900–1930 гг., включающего творческое насле-

дие максимального количества русских писателей, писавших и публиковавшихся в данный период. Корпус создается в рамках исследовательского проекта, посвященного проблеме изучения изменений, происходящих в языке в переломные моменты истории. Первая треть XX в. обернулась для России целой чередой социальных потрясений – русско-японская война, Первая русская революция, Первая мировая война, Февральская и Октябрьская революции 1917 г., Гражданская война, становление нового советского государства, нэп, коллективизация, начало индустриализации. Разработка лингвистического ресурса, охватывающего произведения большого числа авторов для каждого временного среза, даст возможность проводить лингвостатистический анализ языка и стиля в синхронии и диахронии.

Ключевые слова: корпусная лингвистика, русская литература, русский язык, русский рассказ, начало XX века.

1. Введение

В статье рассматриваются основные принципы создания корпуса русского рассказа начала XX в., разрабатываемого на филологическом факультете Санкт-Петербургского государственного университета. Корпус создается в рамках исследовательского проекта, посвященного проблеме изучения изменений, происходящих в языке в переломные моменты истории. Каждая революционная эпоха меняет привычный жизненный уклад и сложившиеся социальные отношения, приводит к трансформации стереотипов поведения и общей системы ценностей, что не может не влиять на языковые изменения.

Первая треть XX в. обернулась для России целой чередой социальных потрясений: русско-японская война, Первая русская революция, Первая мировая война, Февральская и Октябрьская революции 1917 г., Гражданская война, становление нового советского государства, нэп, коллективизация, начало индустриализации. Без преувеличения можно сказать, что произошедшие в России 100 лет назад революционные события драматическим образом повлияли на русский язык – огромный пласт «отжившей» лексики сменился новыми словами, которые отражали новые понятия и идеи; многие слова «из прошлой эпохи» приобрели или новые значения, или новые коннотации; существенные изменения произошли в стилистике, произошла трансформация общепринятых речевых структур (в частности, поменялись частоты многих лексических единиц, сменился набор частотных коллокаций, появились новые синтаксические обороты и т. д.).

Помимо «естественного» процесса резких языковых изменений, неизбежно сопровождающих любую переломную эпоху, следует отметить и целенаправленные действия «новых властей» на изменение языковых норм с целью максимально отмежеваться от уходящей эпохи и подчинить языковую политику государства решению новых актуальных задач. Этим, в частности, была вызвана реформа русской графики и орфографии 1917–1918 гг., а также новая языковая политика Советской России.

Не излишне обратить внимание на то, что часть языковых изменений, произошедших в революционную эпоху, лежат на поверхности, другие же имеют скрытый характер. Однако для осознания масштабности и тех и других языковых сдвигов необходимо применение строгих количественных методов, обращение к представительному объему речевого материала и сравнение последовательности хронологических срезов. Только в результате такого исследования можно будет с уверенностью установить, в какой степени революционная эпоха привела к значимым трансформациям в русском языке, какие языковые уровни изменились в первую очередь и в чем конкретно выразились эти изменения [1].

Для проведения масштабных исследований подобного рода необходимо создание представительного корпуса. Принципам создания такого ресурса посвящена данная работа.

2. Художественная литература как материал для исследования языковых изменений

Во все времена поэты и писатели чутко реагировали на происходящие вокруг события. На страницах литературных текстов прослеживается становление языковых норм, отражается их развитие и изменение. Поэтому художественную прозу можно считать надежным источником информации, не только в содержательном, но и в качественном аспекте.

О важности и необходимости изучения языка художественной литературы хорошо сказал выдающийся лингвист, один из основателей корпусной лингвистики Джон Синклэр: «Литература является ярким примером использования языка; никакой систематический подход не может претендовать на описание языка, если он не охватывает также литературу; при этом она должна рассматриваться не

как некое причудливое образование, но как естественное составляющее в системе языка»[1][2].

Художественная проза, особенно жанр рассказа, оперативно реагирует на события в социальной, политической и культурной жизни, особенно в эпоху социальных преобразований и кардинальных исторических сдвигов. При этом интерес представляет вся совокупность произведений, публикуемых в изучаемую эпоху. Более того, мы полагаем, что исследователя должна интересовать вся фактура произведения: и авторская речь, отражающая множество точек зрения на происходящие события, и речь персонажей, отражающая живые языковые процессы.

На сегодняшний день реализовано большое количество корпусных проектов литературоведческого характера (см., например, [3], [4], [5]).

Исследования русской литературы проводятся как на материале широко известного Национального корпуса русского языка [6], так и на базе специализированных компьютерных ресурсов. В частности, в МГУ были разработаны корпус и частотный словарь языка художественных произведений А. П. Чехова [7], в ИРЯ – частотный словарь языка Ф. М. Достоевского [8] и проведен дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг. [9], в СПбГУ – подготовлена серия частотных словарей выдающихся русских писателей – А. П. Чехова [10], Л. Н. Андреева [11], А. И. Куприна [12], И. А. Бунина [13].

Однако стоит отметить, что цель многих подобных проектов – описание языка только одного выдающего автора (Лермонтова, Достоевского, Чехова, Шекспира, Диккенса, В. Вулф и др.). При этом абсолютное большинство писателей – как известных, так, тем более, менее известных или забытых – оказываются вне поля зрения специалистов по автоматическому лингвистическому анализу. Таким образом, современная компьютерная лингвистика еще не в полной мере использует потенциал художественной литературы, реализованной в виде электронных текстов, принадлежащих как выдающимся, так и периферийным авторам. Разрабатываемый корпус призван восполнить этот пробел. Мы рассчитываем, что он сможет удовлетворить информационные потребности и филолога-словесника, и лингвиста, и литературоведа.

---

[1] Перевод наш (авт.).

3. Жанр и временной срез художественной прозы разрабатываемого ресурса

Создаваемый корпус мыслится как однородный по жанру ресурс, ориентированный в сторону наиболее распространенного жанра художественной прозы – жанра рассказа. Этот жанр наиболее популярен среди прозаиков, он охватывает практически все литературные направления и вовлекает в свою орбиту практически всех писателей. Более того, рассказы значительно быстрее, чем более крупные прозаические жанры, проходят издательский цикл, причем достаточная большая их часть публикуется в литературных журналах, которые также вовлекаются в единый художественно-литературный процесс. Поэтому можно утверждать, что рассказ, как особый жанр, выполняет «разведочную» функцию и даже работает на опережение, чутко улавливая и реагируя на изменения в общественном сознании и культуре общества.

Представительного корпуса русского рассказа до сих пор не существует. К примеру, даже в крупнейшем на сегодняшний день Национальном корпусе русского языка [6] жанр рассказа за 1900–1917 гг. на 01.09.2017 был представлен только 54 писателями (645 рассказами), что составляет лишь незначительную часть авторов, творивших и публиковавшихся в рассматриваемую эпоху. Более того, выборка художественных текстов, представленных в НКРЯ, не является сбалансированной (например, для той же подвыборки 73 рассказа (что составляет 11%) принадлежат перу И. А. Бунина, а 67 рассказа – Н. А. Тэффи, в то время как рассказы многих известных писателей не присутствуют вовсе, не говоря уже о второстепенных прозаиках.

Разрабатываемый корпус посвящен, быть может, самому драматическому периоду в развитии русского языка и литературы. Историческим центром этого периода, его переломом, является Октябрьская революция. Все остальные события и процессы рассматриваются или как преддверие центрального события, или как его последствия. Это позволяет произвести количественный анализ языковых изменений в достаточно широких хронологических рамках и оценить, какие из возникших языковых изменений закрепились в языке и прочно вошли в сознание его носителей, а какие – ушли из языка по завершении революционной эпохи.

На материале корпуса можно будет исследовать язык трех первых десятилетий XX века (1900–1930), разделенный на три времен-

ных среза: 1) начало XX века и предреволюционные годы, включая Первую мировую войну, 2) революционные годы – Февральская и Октябрьская революции и Гражданская война, и 3) постреволюционные годы – с окончания Гражданской войны до 1930 г. Каждый из этих временных периодов будут исследоваться отдельно, а полученные результаты будут объединены в общую картину, отражающую развитие русского языка в первой трети XX в.

4. Отражение «литературно-художественной системы» эпохи как важнейший принцип разрабатываемого ресурса

При создании корпуса русского рассказа первой трети XX в. ставится задача включения в базу текстов максимального числа авторов, писавших в исследуемую эпоху. Такой подход будет способствовать объективности проводимых на его материале лингвистических исследований, поскольку позволит изучать творчество не только ведущих писателей рассматриваемой эпохи, но и множество второстепенных авторов. Художественное наследие последних позволит обогатить представления ученых как о разных сторонах общественной и культурной жизни, так и о характерных особенностях языка того времени.

Первым о необходимости подобного подхода еще в 20-е гг. XX в. заявил видный представитель русской формальной школы Ю. Н. Тынянов. Он выдвинул идею системности литературы, в частности – идею «литературно-художественной системы», включающей в себя всю литературно-художественную продукцию конкретной исторической эпохи [14]. Тогда эта масштабная задача была только поставлена. В те времена она и не могла быть реализована, поскольку для ее решения необходимы мощные средства обработки данных, которые появились только в самое последнее время.

Важную роль в формировании концепции реализуемой в проекте модели корпуса русского рассказа сыграли идеи Андрея Белого, касающиеся необходимости массового создания словарей писателей [15], классификационные представления В. В. Виноградова, предлагавшего строить лингвистические аналоги литературных школ, направлений, стилей на основании критерия лингвистической близости произведений различных авторов [16], а также предложенный В. М. Жирмунским способ описания мировосприятия писателя через совокупность «словесных тем» [17].

Первым опытом системного анализа стиля русской литературы было лингвостатистическое исследование русской литературы, проведенное Г. Я. Мартыненко на материале художественных произведений 100 русских писателей, творивших в начале XX в. [18]. В результате были получены предварительные количественные данные о стилистической близости и стилистических отличиях русских прозаиков, выявлены характерные черты их индивидуальных стилей и построены многомерные классификации авторского стиля (см. рис. 1) [там же].

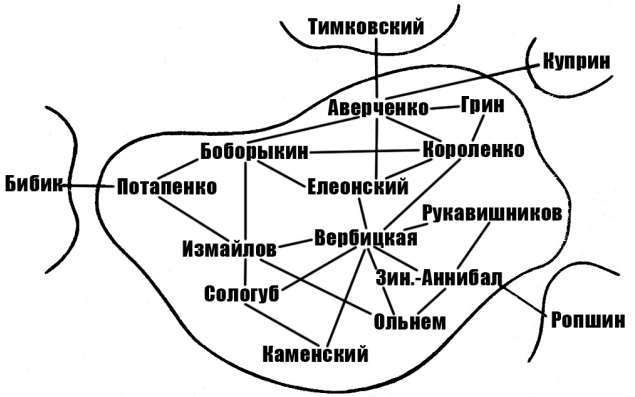


Рис. 1. Подграф с вершинами, принадлежащими к одному беллетристическому классу

Следует отметить, что в последнее время в гуманитарных науках возникли концепции эволюционных процессов, развивающие идеи Ю. Н. Тынянова о литературной эволюции, связанные с переориентацией ядра и периферии при переходе от одной литературно-художественной системы к другой. Речь прежде всего идет о динамике социокультурных систем: о циклических процессах, быстротекущих (несколько десятилетий) и медленно текущих (последовательность столетий), о соотношении понятий развития и эволюции, как чередования катастроф (резких социальных и культурных перемен) и монотонного развития.

Создание представительного корпуса русского рассказа первой трети XX в., охватывающего произведения большого числа (предположительно, несколько сотен) авторов для каждого временного среза, даст возможность проводить лингвостатистический анализ языка и стиля в синхронии и диахронии.

5. Основные задачи первого этапа разработки ресурса

На начальном этапе создания корпуса осуществляются следующие процедуры:

– Формирование представительного списка авторов и их произведений для представительного корпуса русского рассказа 1900–1930 гг.

– Каталогизация рассказов, написанных и опубликованных в исследуемый период.

– Поиск электронных версий текстов.

– Отбор представительной выборки текстов, подлежащих лингвистическому аннотированию.

– Оцифровка рассказов, для которых нет электронных версий.

– Перевод текстов из старой орфографии.

– Корректура электронных версий текстов.

– Метаразметка отобранных текстов.

– Сегментация текстов на структурные части (разделы, абзацы, предложения).

– Вычленение текста автора (нарратива), речи персонажей и авторских ремарок.

– Лингвистическое аннотирование текстов.

Первой важной задачей является получение максимально полного списка авторов, писавших рассказы в исследуемый временной период. Рассмотрим подробнее, как осуществляются ее решение.

6. Формирование представительного списка авторов и их произведений

Источниками для формирования списка авторов и их произведений являются:

1) Библиографии и литературные энциклопедии, словари писателей и пр.

2) Каталоги библиотек, в том числе онлайн каталоги (Российская национальная библиотека, Российская государственная библиотека, Научная библиотека Санкт-Петербургского государственного университета и др.).

3) Периодические издания исследуемой эпохи («Аполлон», «Шиповник», «Нива», «Огонек», «Новый мир» и мн. др.).

4) Электронные библиотеки – Lib.ru: «Классика» [19], Litres [20], Альдебаран [21], ЛитМир [22], электронная библиотека ИМЛИ РАН [23] и др.

5) Другие интернет-ресурсы (Проект «Совлит» [24], «3500 русских прозаических произведений (1800–1940)» [25], «Почти забытые» [26] и др.).

В корпус включаются только те литераторы, творческое наследие которых представлено хотя бы одним рассказом. Ставится задача вовлечения в исследование не только «столичных», но и региональных писателей, писавших на русском языке и проживавших на территории Российской империи (до 1917), а позже – на территории РСФСР и СССР. Было принято решение, что писатели, эмигрировавшие во время революционных событий, не включаются в системный анализ после их эмиграции. Также на данном этапе не планируется включение в корпус произведений детских писателей.

6.1. Использование энциклопедической информации

Первичный список авторов для включения в корпус формируется на основании энциклопедической информации о персоналиях (например, [27]), существующих библиографических указателей (например, [28], [29], [30], [31] и др.), антологий русского рассказа и сборников рассказов (например, [32]), а также публикаций в авторитетных периодических изданиях.

Сначала на материале Краткой литературной энциклопедии [27] была создана база данных, в которую были включены все русские писатели исследуемого периода, при этом отмечалась степень их вовлеченности в писательских процесс по трем периодам, соответствующим первым трем десятилетиям XX в. В табл. 1 это показано цифрами от 0 до 1: (1) – активный творческий период, (0) – отсутствие творческой деятельности или эмиграция. Например, значение 0,5 показывает, что творчество автора охватывает только половину соответствующего десятилетия. В правой колонке таблицы приводятся числовые данные о размере энциклопедической статьи в словах. Эти данные косвенно указывают о значимости автора с точки зрения литературоведов-составителей энциклопедии.

Таблица 1. **Фрагмент базы данных русских писателей по данным Краткой Литературной Энциклопедии**

ФАМИЛИЯ, Имя Отчество	Творчество до...	Десятилетия XX в.			Размер статьи
		I	II	III	
КОНИЧЕВ, Константин Иванович		0	0	0,1	190
КОПТЕЛОВ, Афанасий Лазаревич		0	0	0,6	392
КОРОЛЕНКО, Владимир Галактионович	1921	1	1	0	1989
КОСТЫЛЕВ, Валентин Иванович		1	1	1	307
КРАВЧЕНКО, Федор Тихонович		0	0	0,5	123
КРАШЕНИННИКОВ, Николай Александрович		1	1	1	203
КРЕПТЮКОВ, Даниил Александрович		0	0	0,5	154
КРЕТОВА, Ольга Капитоновна		0	0	0,4	118
КРЖИЖАНОВСКИЙ, Сигизмунд Доминикович		0	1	1	206
КРИНИЦКИЙ, Марк		1	1	1	297
КРУТИКОВ, Дмитрий Иванович		0	1	1	188
КРУШИНСКИЙ, Сергей Константинович		0	0	0,5	185
КРЮКОВ, Федор Дмитриевич	1920	1	1	0	300
КУЗМИН, Михаил Алексеевич		1	1	1	377
КУКЛИН, Георгий Осипович		0	0	0,4	140
КУМОВ, Роман Петрович	1919	1	1	0	130
КУПРИН, Александр Иванович	1919	1	1	0	1783
ЛАВРЕНЕВ, Борис Андреевич		0	0	0,6	580

ЛАДЫЖЕНСКИЙ, Владимир Николаевич	1919	1	1	0	144
ЛАНСКОЙ, Марк Зосимович		0	0	0,2	151
ЛАРРИ, Ян Леопольдович		0	0	1	168
ЛЕБЕДЕВ, Всеволод Владимирович		0	0	0,9	151
ЛЕБЕДЕНКО, Александр Гервасьевич		0	0,3	1	234

По данным КЛЭ был сформирован список, насчитывающий 273 персоналии. Ожидаемо, самые крупные статьи посвящены выдающимся русским писателям – Л. Н. Толстому (8688 слов), А. П. Чехову (6334), М. Горькому (3151), М. А. Шолохову (2946) и А. Н. Толстому (2063). Оказалось, что статьи, посвященные 20 наиболее значимым авторам, составляют 40% от всего объема текстов, посвященных всем 273 авторам, а 112 писателей характеризуются статьями, состоящими менее чем из 200 слов.

Однако полученный таким образом список авторов оказался одновременно как недостаточным, так и избыточным. С одной стороны, в КЛЭ не попали многие периферийные русские писатели, а с другой – по относительно небольшим статьям энциклопедии не всегда понятно, писал ли данный автор рассказы, и приходятся ли они на изучаемый временной период.

6.2. Использование библиографической информации

Поскольку для создания корпуса важное значение имеют не только известные авторы, включенные в энциклопедии и словари, но и менее известные, периферийные писатели, было решено далее пойти несколько иным путем – а именно, через каталогизацию произведений, представленных в каталогах крупных библиотек при заданном жанровом фильтре «рассказ».

В результате, на базе электронного каталога Российской национальной библиотеки была сформирована база данных всех рассказов, изданных отдельными брошюрами. Фрагмент такой базы представлен на рис. 2, а в табл. 2 приведен список русских авторов, полученный из данного каталога, рассказы которых в виде отдельных брошюр были изданы наибольшее количество раз (по данным каталога РНБ).

Код	Автор	Название	Библиография	Переиздани	Год издани
3152	Белоусов, Д.М.	По насту	Белоусов Д.М. ... По насту : [Рассказ]. - Новосибирск :		1927
3150	Белоусов, Д.М.	Тишка-волчатник	Белоусов Д.М. ... Тишка-волчатник : Рассказ. - Новоси		1927
428	Бельский, Леонид Петрович	Как тварь земная Богу молит	Бельский Л. П. Как тварь земная Богу молится : [Рассн		1903
2421	Беляев, Иван Дмитриевич	Голод	Беляев И.Д. Голод: [(День в умирающем городе)]: [(		1918
1258	Беляев, Павел	На диаконском штате	Беляев П. На диаконском штате : Рассказ / Свящ. Пав		1908
1404	Беляевская, Ольга Александ	Зина : Рассказ учительницы	Беляевская О.А. Зина : Рассказ учительницы / [Соч.] С		1909
1738	Беляков, Василий Борисови	Последний час с природой	Беляков В.Б. Последний час с природой : [Рассказ] / В		1912
3381	Беляков, Гр.	Ко дну	Беляков Гр. ... Ко дну : Рассказ. - Москва : Центросою		1929
3314	Беляков, Николай Диомидо	Волчья сторона	В полосе хвойных лесов : [Рассказ] / Н. Беляков; Рис.		1928
3317	Беляков, Николай Диомидо	Жучки и яблочки	[Рассказ] / Н. Беляков; Рис. и обложка В. Н. Берга. - М		1928
3316	Беляков, Николай Диомидо	Сенька куровод	Беляков Н.Д. ... Сенька куровод : [Рассказ] / Рис. и об		1928
3122	Беляков, Николай Диомидо	Сосунок	Беляков Н.Д. ... Сосунок : [Рассказ] .. / Н. Беляков. - М		1927
2934	Березовский, Феоктист Але	Варвара	Березовский Ф.А. Варвара : Рассказ / Феоктист Берез		1926
2741	Березовский, Феоктист Але	Во рву	Березовский Ф.А. Во рву : [Рассказ] / Ф.Березовский.		1925
3302	Березовский, Феоктист Але	За решеткой	Березовский Ф.А. ... За решеткой : [Рассказ] / Ф. Бере		1928
2861	Берзинь, Анна Абрамовна	Горбатенький	Берзинь А.А. Горбатенький : Рассказ / Ф.Ложкин [псе		1925
3505	Берман, Лазарь Васильевич	Кобылья вахта	Берман Л.В. ... Кобылья вахта : [Рассказ] / Л. Берман;		1930
3404	Берман, Лазарь Васильевич	Фабрика хлеба	Берман Л.В. ... Фабрика хлеба : [Рассказ].. / Б. [?] Бер		1929
2992	Бианки, Виталий Валентино	На волков с барабаном	Бианки В.В. На волков с барабаном : [Рассказ] / Вит. Б		1926
3369	Бианки, Виталий Валентино	За ястребом	Бианки В. В. ... За ястребом : [Рассказ] / В. Бианки. - [М		1928
2704	Бианки, Виталий Валентино	Кукушонок	Бианки В.В. Кукушонок : [Рассказ] / Виталий Бианки; [		1924
2682	Бианки, Виталий Валентино	Лесные домики	Бианки В.В. Лесные домики : [Рассказ] / Виталий Биан		1924
3055	Бианки, Виталий Валентино	Мал да удал	Бианки В.В. Мал да удал : [Рассказ] / Вит. Бианки; Рис		1926
2907	Бианки, Виталий Валентино	Однодневки	Бианки В.В. Однодневки : [Рассказ] / Виталий Бианки		1925

Рис. 2. Фрагмент базы данных русских рассказов, изданных отдельными изданиями в 1900–1930 гг. (по данным каталога РНБ)

Табл. 2. **Русские писатели, рассказы которых были изданы наибольшее количество раз в виде отдельных брошюр (по данным каталога РНБ)**

Писатель	Кол-во рассказов	Писатель	Кол-во рассказов
Зотов, Михаил	71	Куприн, Александр Иванович	24
Васильковский, Александр Константинович	64	Лукашевич, Клавдия Владимировна	23
Мамин-Сибиряк, Дмитрий Наркисович	54	Дмитриева, Валентина Иововна	23
Семенов, Сергей Терентьевич	44	Любич-Кошуров, Иоасаф Арианович	20
Тхоржевский, Корнелий Владиславович	41	Засодимский, Павел Владимирович	19
Шухмин, Христофор Алексеевич	40	Короленко, Владимир Галактионович	19
Горький, Максим	38	Чеглок, Александр Александрович	18
Немирович-Данченко, Василий Иванович	32	Подъячев, Семен Павлович	18
Гусев-Оренбургский, Сергей Иванович	30	Осетров, Захар Борисович	17

Серафимович, Александр Серафимович	29	Шмелев, Иван Сергеевич	16
Станюкович, Константин Михайлович	28	Ульянов, Алексей Николаевич	15
Андреев, Леонид Николаевич	28	Толстой, Лев Николаевич	15
Наживин, Иван Федорович	27	Петрова, Вера Константиновна	15
Зощенко, Михаил Михайлович	27	Митропольский, Иван Иванович	14
Круглов, Александр Васильевич	24	Караулов, Михаил Федорович	14

На рис. 3 приведены количественные данные об этих рассказах по годам.

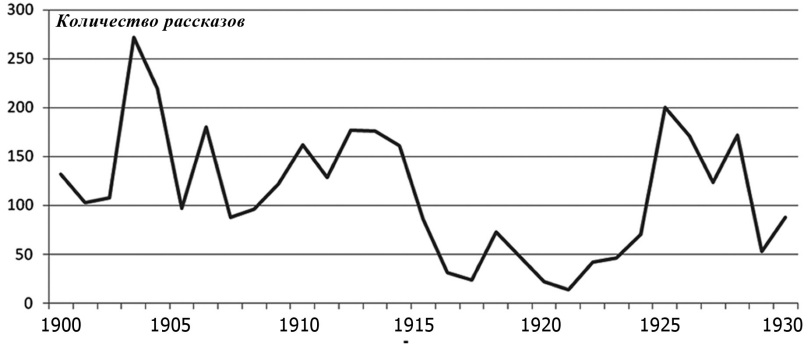


Рис. 3. Динамика публикаций рассказов по годам за 1900–1930 гг. (по данным РНБ, для изданий рассказов отдельными брошюрами)

6.3. Использование библиографической информации

Кроме того были выборочно каталогизированы периодические издания исследуемой эпохи («Аполлон», «Шиповник», «Нива», «Огонек», «Новый мир» и мн. др.) и доступные в сети электронные библиотеки [19–22] и некоторые другие интернет-ресурсы [24–26].

В результате был получен предварительный список имен писателей исследуемого периода, который насчитывает более 2500 персоналий.

Параллельно с формированием списка авторов происходит сведение электронных версий текстов (при их наличии), а малоизвестных и забытых авторов – оцифровка текстов (в первую очередь для периода 1900–1914), их ручное редактирование и внесение метаинформации об авторах и их произведениях.

Промежуточные итоги

К моменту подготовки статьи статистические характеристики корпуса выглядят таким образом:

Список имен прозаиков – более 2500[1].

Писатели, для которых есть хотя бы один рассказ в цифровом виде – 300,

из них тексты, специально оцифрованные для корпуса – 45.

Общее количество рассказов в цифровом виде – 3200.

Писатели, для которых найдено максимальное количество рассказов в электронной версии, представлены в табл. 3, а на рис. 4 показано распределение этих рассказов по годам.

Из таблицы видно, что далеко не все имена даже из этого списка известны широкому читателю.

Табл. 3. **Русские писатели и количество рассказов 1900–1930 гг. в цифровом виде**

Писатель	Кол-во рассказов	Писатель	Кол-во рассказов
Аверченко Аркадий Тимофеевич	740	Куприн Александр Иванович	44
Грин Александр Степанович	128	Лазаревский Борис Александрович	44
Серафимович Александр Серафимович	110	Гуро Елена	37
Амфитеатров Александр Валентинович	84	Щепкина-Куперник Татьяна Львовна	35
Бунин Иван Алексеевич	78	Брусянин Василий Васильевич	34

[1] Данный список следует считать предварительным, так как совсем не для всех авторов имеются библиографические данные, которые могут подтвердить, что данный автор действительно жил и творил в исследуемую эпоху. Кроме того, необходимо проверить этот список по словарю псевдонимов, и можно предполагать, что общее количество уникальных писателей в результате несколько снизится.

Андреев Леонид Николаевич	72	Чириков Евгений Николаевич	33
Будищев Алексей Николаевич	69	Гарин-Михайловский Николай Георгиевич	32
Горький Максим	60	Кржижановский Сигизмунд Доминикович	31
Романов Пантелеймон Сергеевич	60	Мамин-Сибиряк Дмитрий Наркисович	30
Дорошевич Влас Михайлович	59	Сологуб Федор	30
Шухмин Христофор Алексеевич	56	Потапенко Игнатий Николаевич	29
Федоров Александр Митрофанович	50	Авилова Лидия Алексеевна	25
Лейкин Николай Александрович	49	Баранцевич Казимир Станиславович	25
Гребенщиков Георгий Дмитриевич	46	Богданов Александр Алексеевич	25
Замятин Евгений Иванович	44	Гусев-Оренбургский Сергей Иванович	25

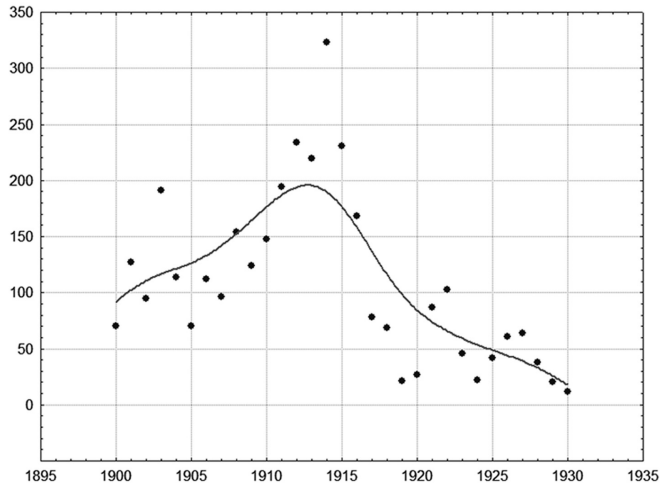


Рис. 4. Распределение корпуса электронных текстов рассказов по годам (для выборки в 3200 рассказов)

Заключение

Первый этап работы над созданием корпуса русского рассказа первой трети XX в. выявил неожиданно большое количество имен малоизвестных и фактически забытых авторов, в том числе и тех, которые в исследуемую эпоху печатались относительно много, и даже чаще других, считающихся относительно известными. Поэтому пополнение корпуса электронных текстов за счет оцифровки новых рассказов планируется осуществить в первую очередь для этих авторов. Каталогизация русских рассказов исследуемой эпохи будет продолжена с привлечением других периодических изданий. В настоящее время осуществляется формирование списка параметров для лингвистического анализа текстов и разработка программного обеспечения для обработки корпусных данных.

Максимально полная каталогизация рассказов изучаемой эпохи позволит проводить объективный статистический анализ состояния и тенденций литературного процесса, а получение цифровых текстов забытых авторов сделает их доступными для чтения и исследования, и даст возможность использовать для их анализа современные информационные, компьютерные и корпусные технологии.

Благодарности. Работа выполнена при поддержке Российского фонда фундаментальных исследований, грант № 17-29-09173 офи_м «Русский язык на рубеже радикальных исторических перемен: исследование языка и стиля предреволюционной, революционной и постреволюционной художественной прозы методами математической и компьютерной лингвистики (на материале русского рассказа)».

ЛИТЕРАТУРА

1. Мартыненко Г. Я., Шерстинова Т. Ю., Мельник А. Г., Попова Т. И. Методологичес**к**ие проблемы создания Компьютерной антологии русского рассказа как языкового ресурса для исследования языка и стиля русской художественной прозы в эпоху революционных перемен (первой трети XX века) / Компьютерная лингвистика и вычислительные онтологии. Выпуск 2 (Труды XXI Международной объединенной конференции “Интернет и современное общество, IMS-2018, Санкт-Петербург, 30 мая – 2 июня 2018 г. Сборник начных статей”). – СПб: Университет ИТМО, 2018. С. 99–104.

2. Sinclair, J. Trust the Text: Language, Corpus and Discourse, Routledge, 2004, 224 p.

3. Fischer-Starcke B. Corpus Linguistics in Literary Analysis : Jane Austen and Her Contemporaries. London; New York: Continuum, 2010.

4. Balossi G.. A Corpus Linguistic Approach to Literary Language and Characterization: Virginia Woolf's The Waves. Amsterdam; Philadelphia: John Benjamins Publishing Company, 2014.

5. CliC Dickens. URL:http://clic.bham.ac.uk/ (дата обращения: 22.12.2017).

6. Национальный корпус русского языка. URL: http://www.ruscorpora.ru (дата обращения 12.12.2017).

7. Частотный грамматико-семантический словарь языка художественных произведений А.П. Чехова. URL: http://www.philol.msu.ru/~lex/chehov.html (дата обращения: 15.12.2017).

8. Шайкевич А.Я., Андрющенко В.М., Ребецкая Н.А. Статистический словарь языка Достоевского. М.: Яз. слав. культуры, 2003.

9. Шайкевич А.Я., Андрющенко В. М., Ребецкая Н. А. Дистрибутивно-статистический анализ языка русской прозы, 1850–1870-х гг. М.: Языки славянской культуры, 2013.

10. Частотный словарь рассказов А. П. Чехова / Сост. А. О. Гребенников. Под ред. Г. Я. Мартыненко. СПб.: Изд-во Санкт-Петерб. ун-та, 1998.

11. Частотный словарь рассказов Л. Н. Андреева / Сост. А. О. Гребенников. Под ред. Г. Я. Мартыненко. СПб.: Изд-во Санкт-Петерб. ун-та, 2003.

12. Частотный словарь рассказов А.И. Куприна / Сост. А. О. Гребенников, Н. А. Данилова. Под ред. Г. Я. Мартыненко. СПб.: Изд-во С.-Петербургского университета, 2006.

13. Частотный словарь рассказов И. А. Бунина // Сост. А. О. Гребенников. Под ред. Г. Я. Мартыненко. СПб: Изд-во Санкт-Петерб. ун-та, 2012.

14. Тынянов Ю.Н. Архаиста и новаторы. Л.: Прибой, 1929.

15. Белый Андрей. Мастерство Гоголя. Л.: ОГИЗ, 1934. – 320 с.

16. Виноградов В. В. О языке художественной литературы. М., Гослитиздат, 1959. – 654 с.

17. Жирмунский В. М. Задачи поэтики / Жирмунский В. М. Теория литературы. Поэтика. Стилистика. Л.: Наука, 1977.

18. Мартыненко Г. Я. Основы стилеметрии. Издательство ЛГУ, 1988. – 176 с.

19. Lib.ru: «Классика» (Библиотека Максима Мошкова),. URL: http://az.lib.ru/ (дата обращения: 15.10.2018).

20. Электронная библиотека Litres, URL: https://www.litres.ru (дата обращения: 15.10.2018).

21. Электронная библиотека книг Альдебаран, URL: https://aldebaran.ru (дата обращения: 15.10.2018).

22. Электронная библиотека ЛитМир, URL: https://www.litmir.me (дата обращения: 15.10.2018).

23. Электронная библиотека ИМЛИ РАН, URL: http://biblio.imli.ru (дата обращения: 15.10.2018).

24. Проект «Совлит», URL: http://www.ruthenia.ru/sovlit/ (дата обращения: 28.10.2018).

25. «3500 русских прозаических произведений (1800–1940)», URL: https://doxie-do.livejournal.com/260117.html (дата обращения: 28.10.2018).

26. «Почти забытые», URL: https://slovesnik.org/chto-chitat/pochti-zabytye.html (дата обращения: 28.10.2018).

27. Краткая литературная энциклопедия в 9 т. М.: Советская энциклопедия, 1962–1978, URL: http://feb-web.ru/feb/kle/kle-abc/default.asp (дата обращения: 10.11.2018).

28. Владиславлев И. В. Русские писатели XIX – XX столетия. М., 1924.

29. Голубева О. Д. Из истории издания русских альманахов начала XX в. М., 1960.

30. Муратова К. Д. История русской литературы конца XIX – начала XX века, Библиографический указатель М.: АнСССР, 1963.

31. Русские писатели 1800–1917, Биографический словарь. В 7 т. Под ред. Николаева П. А. 1992–2000. М.: Научное издательство «Большая российская энциклопедия».

32. Нагибин Ю. М. (составитель). Антология русского советского рассказа. Предисловие Ю. Нагибина. Библиотечка «Книжного обозрения». М.: Книжное обозрение, 1987.

УДК 378.016:004.02

GENERATION OF LEARNING TASKS WITH USING DATABASES

C. B. Minnegalieva
Kazan Federal University, Kazan
mchulpan@gmail.com

The article describes methods for generating learning tasks based on information about the values of properties of objects and information about relations between objects.

Keywords: generation of learning tasks, electronic educational resource.

ГЕНЕРАЦИЯ УЧЕБНЫХ ЗАДАНИЙ С ИСПОЛЬЗОВАНИЕМ БАЗ ДАННЫХ

Ч. Б. Миннегалиева
Казанский федеральный университет, Казань
mchulpan@gmail.com

В статье описываются приемы генерации учебных заданий на основе информации о значениях свойств объектов и информации об отношениях между объектами.

Ключевые слова: генерация учебных заданий, электронный образовательный ресурс.

Создание необходимого количества учебных заданий всегда занимало время преподавателя и являлось достаточно трудоемкой работой. С внедрением в учебный процесс дистанционных технологий и развитием электронного обучения изменились требования к объему базы контрольных измерительных материалов. В настоящее время по одному и тому же дистанционному курсу могут обучаться сотни и более студентов одновременно. Для объективности контроля необходимо, чтобы задания, предъявляемые студентам, не повторялись. Поэтому актуальной является задача автоматической генерации необходимого количества учебных заданий для электронных образовательных ресурсов.

Если рассматривать дисциплины физико-математического цикла, по некоторым темам возможна генерация параметров и структуры задания при помощи систем компьютерной математики, библио-

тек математических функций языков программирования [1], [2].

По другим дисциплинам количество заданий, условия которых можно формализовать, меньше. Методы генерации на основе учебного материала можно разделить на две группы: использующие методы компьютерной лингвистики и методы, предусматривающие построение локальных фрагментов представления предметной области [3]. Во вторую группу можно включить генерацию учебных заданий на основе перечислений, на основе информации об отношениях между объектами, на основе информации о значениях свойств объектов. Информацию об объектах, их свойствах удобно хранить в структурированных базах данных. Табличная форма представления информации самая простая. В таблице можно хранить набор терминов, исторических дат, географических данных. Рассмотрим этот прием на примере генерации заданий по географии.

В таблице «Объекты» хранятся сами географические объекты: страны, республики, города, океаны, моря, реки и озера. Помимо названий сюда же записываются географические координаты и тип объекта. Типы объектов: страна, республика, город, океан, море, река и озеро. Таблица «Свойства» содержит все основные параметры географических объектов. К ним относятся числовые характеристики: население, плотность населения, площадь и процент водной поверхности. В системе генерации заданий могут быть другие вспомогательные таблицы для осуществления связи между данными. Вопросы можно разделить на две группы: вопросы, связанные с числовыми характеристиками объектов и вопросы на принадлежность объектов друг к другу.

Приведем примерные шаблоны для генерации вопросов по числовым характеристикам объекта:

– Как[ой] [город] имеет [наименьшую] [площадь]?

Слова, заключенные в квадратные скобки, меняются. Город – страна – республика – море и т. д. Наименьшую – наибольшую, площадь – плотность населения – процент водной поверхности. В зависимости от предполагаемой сложности вопроса из базы выбираются 40126 объектов. В программе предусматривается, чтобы они были из одного диапазона. Например, если задается вопрос про площадь городов, то наличие одном вопросе городов Пекин и Чистополь будет служить подсказкой для обучаемого. Поэтому предварительно объкты сортируются по указанному свойству.

– Расположите [страны] в порядке [убывания] их [плотности населения]. В этом случае названия стран (городов, озер) также вы-

бираются так, чтобы их плотность населения (площадь, процент водной поверхности) отличались друг от друга на небольшие величины.

На принадлежность объектов друг другу можно задать следующие типы вопросов:

– В какой [стране] находится [город]?

Аналогично случаям, рассмотренным выше, можно изменить страна – город, город – озеро – море и объекты выбираются так, чтобы не был очевиден ответ. Например, вопрос «В какой стране находится город Рим?» для большинства обучаемых окажется легким.

– Выбрать [озера], расположенные в [стране].

Генерация вопросов происходит аналогично рассмотренным примерам, в программе или при проектировании базы данных предусматривается проверка принадлежности данных одному диапазону. Если надо выбрать озера, расположенные в Норвегии, то наличие среди возможных ответов «озеро Байкал» будет служить непреднамеренной подсказкой для студента.

Имея большую базу данных по географическим объектам, можно составить другие шаблоны и разработать достаточное количество вопросов, учебных заданий.

Созданную структуру базы данных и шаблоны можно использовать для заполнения базы учебных заданий электронных образовательных вопросов по другим дисциплинам, таким, как история, химия, физика, информатика. Используя приемы, перечисленные выше, можно составить вопросы по истории «Соотнесите события с датами», «Выберите правильную дату события», «Укажите имя исторической личности, о ком идет речь», «Расположите в хронологической последовательности исторические события» и т. д. [4]

По химии можно составить вопросы, например, с такой формулировкой: «Из указанных в ряду химических элементов выберите три элемента, которые в Периодической системе химических элементов Д. И. Менделеева находятся в одном периоде. Расположите выбранные элементы в порядке возрастания их металлических свойств», «Выберите три элемента, которые в Периодической системе находятся в одном периоде, и расположите эти элементы в порядке увеличения радиуса атома. Запишите в поле ответа номера выбранных элементов в нужной последовательности». При составлении задач по химии можно воспользоваться Периодической системой химических элементов Д. И. Менделеева, представленной в удобном для разработчиков формате.

Пример вопроса по информатике: «Приведены запросы к поисковому серверу. Расположите номера запросов в порядке возрастания количества страниц, которые найдет поисковый сервер по каждому запросу. Для обозначения логической операции «ИЛИ» в запросе используется символ |, а для логической операции «И» – &. 1) принтеры & сканеры & продажа; 2) принтеры & продажа; 3) принтеры | продажа; принтеры | сканеры | продажа». Для части заданий по информатике можно использовать приемы для генерации заданий по математике [4].

Таким образом, имея информацию об объектах и их свойствах можно выполнить генерацию учебных заданий по разным дисциплинам. Пробные версии таких программ выполнены нами совместно со студентами. Созданные задания могут использоваться при заполнении базы учебных заданий электронного образовательного ресурса. Задания могут составляться в режиме реального времени при использовании дистанционных образовательных технологий.

ЛИТЕРАТУРА

1. Приемы повышения эффективности электронных образовательных ресурсов / Ч.Б. Миннегалиева //Образование и саморазвитие. – 2014. №2(40). – С.105–108.

2. Некоторые вопросы автоматизации контроля знаний / Ч.Б. Миннегалиева, Д.Р. Мухамедшин, К.В. Русецкий, А.В. Паркалов // Компьютерные инструменты в образовании. – 2014. – №6. – С. 52–59.

3. Башмаков А.И., Башмаков И.А. Разработка компьютерных учебников и обучающих систем. М.: «Филин», 2003.

4. Образовательный портал для подготовки к экзаменам решуегэ.рф [Электронный ресурс] https://ege.sdamgia.ru/. Дата обращения 02.10.2018.

УДК 004.891

ABOUT THE APPROACH TO TRANSLATION OF RDF/ OWL-ONTOLOGY TO THE GRAPHIC KNOWLEDGE BASE

V. S. Moshkin, A. A. Filippov, N. G. Yarushkina
Ulyanovsk State Technical University, Ulyanovsk
{v.moshkin, al.filippov, jng}@ulstu.ru

The paper presents the graphical knowledge base of the ontology repository. The original algorithm for translating the RDF/OWL-ontology into a graphical knowledge base is proposed.

Keywords: ontology, graph knowledge base, OWL.

ПОДХОД К ТРАНСЛЯЦИИ RDF/OWL-ОНТОЛОГИИ В ГРАФОВУЮ БАЗУ ЗНАНИЙ

В. С. Мошкин, А. А. Филиппов, Н. Г. Ярушкина
Ульяновский государственный технический университет, Ульяновск
{v.moshkin, al.filippov, jng}@ulstu.ru

В работе рассмотрена модель графовой базы знаний хранилища онтологии, а также предлагается оригинальный алгоритм трансляции RDF/ OWL-онтологии в графовую базу знаний.

Ключевые слова: онтология, графическая база знаний, OWL.

Введение

Современное постиндустриальное общество оперирует огромными массивами информации как в повседневной, так и в профессиональной деятельности. Подобные объемы информации приводят к трудностям при принятии различного рода решений в рамках жестких временных ограничений [1] [2].

Для решения данной проблемы применяются разнообразные интеллектуальные программные средства автоматизации деятельности человека. Однако, для эффективной работы таких средств необходима их адаптация к особенностям конкретной проблемной области (ПрО). Для описания особенностей ПрО обычно использу-

ются онтологии [3-9]. Онтология – модель ПрО, представленная в виде семантической паутины (графа) [10] [11].

В рамках нашей научной группы была предпринята попытка разработки хранилища онтологий, позволяющего:

1. Производить импорт онтологий в форматах RDF и OWL.
2. Формировать запросы к содержимому хранилища.
3. Не требовать от разработчика знаний в области онтологического проектирования и инженерии знаний.
4. Организовать взаимодействие с хранилищем онтологий с помощью протокола HTTP, сделав хранилище максимально независимым от используемого языка программирования и операционной системы.

Трансляция RDF/OWL-онтологии в графовую базу знаний

Для успешной трансляции онтологии, представленной на языках RDF и OWL, в содержимое ГБЗ необходимо выделить структурные элементы, которые будут отнесены к TBox (структура, схема) и ABox (наполнение, содержимое) ГБЗ соответственно.

Функции трансляции RDF/OWL-онтологии в ГБЗ можно представить следующими выражениями:

$$f_O^{RDF} : RDF \to O,$$
$$f_O^{OWL} : OWL \to O,$$

где $RDF = \left\langle C^{RDF}, I^{RDF}, P^{RDF}, S^{RDF}, R^{RDF} \right\rangle$ – множество сущностей онтологии в формате RDF,

$OWL = \left\langle C^{OWL}, I^{OWL}, P^{OWL}, S^{OWL}, R^{OWL} \right\rangle$ – множество сущностей онтологии в формате OWL,

O – множество сущностей онтологии ГБЗ. В таблице 1 представлено соответствие сущностей RDF/OWL-онтологии сущностям ГБЗ.

Таблица 1. **Соответствие сущностей RDF/OWL-онтологии сущностям ГБЗ**

RDF	OWL	ГБЗ
TBox		
rdfs:Resource	owl:Thing	$C^{T_i} = \left\{ C_1^{T_i}, C_2^{T_i}, \ldots, C_n^{T_i} \right\}$
rdfs:Class	owl:Class	$C^{T_i} = \left\{ C_1^{T_i}, C_2^{T_i}, \ldots, C_n^{T_i} \right\}$

Продолжение таблицы 1

RDF	OWL	ГБЗ
rdfs:subClassOf	owl:SubclasOf	$R_C^{T_i}$
rdf:Property	owl:ObjectProperty owl:DataProperty	$P^{T_i} = \left\{ P_1^{T_i}, P_2^{T_i}, \ldots, P_n^{T_i} \right\}$
rdfs:domain	owl:ObjectPropertyDomain owl:DataPropertyDomain	$R_P^{T_i}$
rdfs:range	owl:ObjectPropertyRange owl:DataPropertyRange	$R_P^{T_i}$
ABox		
rdf:type rdf:ID	owl:NamedIndividual	$I^{T_i} = \left\{ I_1^{T_i}, I_2^{T_i}, \ldots, I_n^{T_i} \right\}$
	owl:ClassAssertion	$R_I^{T_i}$
rdf:resource rdf:ID	owl:ObjectPropertyAssertion owl:DataPropertyAssertion	$S^{T_i} = \left\{ S_1^{T_i}, S_2^{T_i}, \ldots, S_n^{T_i} \right\} R_S^{T_i}$

Как видно из таблицы 1, основным сущностям онтологий в формате RDF и OWL соответствуют сущности онтологии ГБЗ. Сущности ГБЗ позволяют унифицировать различные форматы представления онтологий и сформировать модель данных, опираясь на которую, разработчик может формировать запросы к содержимому хранилища онтологий на языке запросов Cypher. Данный метод извлечения знаний из хранилища онтологий является более привычным для разработчика, чем работа с машинами логического вывода. Однако, при этом возможность логического вывода по содержимому нашего хранилища онтологий присутствует.

Заключение

В данной работе рассмотрен подход к трансляции RDF/OWL-онтологии в ГБЗ. Разработанное хранилище онтологий, основанное на ГБЗ, позволяет разработчику взаимодействовать с содержимым такого хранилища наиболее привычным для него путем – формировать запросы к содержимому ГБЗ.

При этом унифицированная модель данных ГБЗ позволяет импортировать онтологии в форматах RDF и OWL. Данные форматы часто используются при формировании описания особенностей ПрО в виде онтологии. TBox ГБЗ определяет доступные для формиро-

вания ГБЗ аксиомы, а также позволяет контролировать логическую целостность содержимого ГБЗ.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (гранты РФФИ 18-37-00450, 18-47-730035 и 18-47-732007).

СПИСОК ЛИТЕРАТУРЫ

1. Бова В.В., Курейчик В.В., Нужнов Е.В. Проблемы представления знаний в интегрированных системах поддержки управленческих решений // Известия ЮФУ. 2010. № 108(7).

2. Черняховская Л.Р., Федорова Н.И., Низамутдинова Р.И. Интеллектуальная поддержка принятия решений в оперативном управлении деловыми процессами предприятия // Вестник УГАТУ. Управление, вычислительная техника и информатика. 2011. Т 15. № 42(2).

3. Вагин В.Н., Михайлов И.С. Разработка метода интеграции информационных систем на основе метамоделирования и онтологии предметной области // Программные продукты и системы. 2008. № 1.

4. Гаврилова Т.А. Онтологический подход к управлению знаниями при разработке корпоративных информационных систем // Новости искусственного интеллекта. 2003. № 2.

5. Загорулько Ю.А. Построение порталов научных знаний на основе онтологии // Вычислительные технологии. 2007. Т. 12. № S2.

6. Карабач А.Е. Системы интеграции информации на основе семантических технологий // Наука, техника и образование. 2014. № 2(2).

7. Смирнов С.В. Онтологическое моделирование в ситуационном управлении // Онтология проектирования. 2012. № 2.

8. Тузовский А.Ф. Разработка систем управления знаниями на основе единой онтологической базы знаний // Известия Томского политехнического университета. 2007. Т. 310. № 2.

9. Филиппов А.А., Мошкин В.С., Шалаев Д.О., Ярушкина Н.Г. Единая онтологическая платформа интеллектуального анализа данных // Материалы VI международной научно-технической конференции OSTIS-2016, Минск, Республика Беларусь, 2016.

10. Ярушкина Н.Г., Мошкин В.С. Применение алгоритма логического вывода на основе FuzzyOWL-онтологии // Радиотехника. – 2015. – № 6. – С. 68–72.

11. Тарасов В.Б., Калуцкая А.П., Святкина М.Н. Гранулярные, нечеткие и лингвистические онтологии для обеспечения взаимопонимания между когнитивными агентами // Матер. второй междунар. научн.-техн. конф. «OSTIS-2012». – Минск, 2012. – С. 267–278.

УДК 81.32

EXTENDED LANGUAGE MODELING EXPERIMENTS FOR KAZAKH

B. Myrzakhmetov[1,2], Z. Kozhirbayev[1,3]
[1] *National Laboratory Astana, Nazarbayev university, Astana, Kazakhstan*
[2] *Nazarbayev university, School of Science and Technology, Astana, Kazakhstan*
[3] *L.N. Gumilyov Eurasian National University, Astana, 010008, Kazakhstan*
bagdat.myrzakhmetov@nu.edu.kz, zhanibek.kozhirbayev@nu.edu.kz

In this article we present dataset for the Kazakh language for the language modeling. It is an analogue of the Penn Treebank dataset for the Kazakh language as we followed all instructions to create it. The main source for our dataset is articles on the web-pages which were primarily written in Kazakh since there are many new articles translated into Kazakh in Kazakhstan. The dataset is publicly available for research purposes[1]. Several experiments were conducted with this dataset. Together with the traditional n-gram models, we created neural network models for the word-based language model (LM). The latter model on the basis of large parametrized long short-term memory (LSTM) shows the best performance. Since the Kazakh language is considered as an agglutinative language and it might have high out-of-vocabulary (OOV) rate on unseen datasets, we also carried on morph-based LM. With regard to experimental results, sub-word based LM is fitted well for Kazakh in both n-gram and neural net models compare to word-based LM.

Keywords: Language Modeling, Kazakh language, n-gram, neural language models, morph-based models.

РАСШИРЕННЫЕ ЭКСПЕРИМЕНТЫ ПО ЯЗЫКОВОЙ МОДЕЛИ ДЛЯ КАЗАХСКОГО ЯЗЫКА

Б. О. Мырзахметов[1,2], Ж. М. Кожирбаев[1,3]
[1] *Национальная Лаборатория Астана, Назарбаев университет, Астана, Казахстан*
[2] *Назарбаев университет, Школа Наук и Технологий, Астана, Казахстан*
[3] *Евразийский национальный университет имени Л. Н. Гумилева, Астана, 010008, Казахстан*
bagdat.myrzakhmetov@nu.edu.kz, zhanibek.kozhirbayev@nu.edu.kz

[1] https://github.com/Baghdat/LSTM-LM/tree/master/data/

В этой статье мы представляем набор данных для казахского языка для языкового моделирования. Это аналог набора данных Penn Treebank для казахского языка, поскольку мы следовали всем инструкциям по его созданию. Основным источником нашего набора данных являются статьи на веб-страницах, которые в основном были написаны на казахском языке, так как в Казахстане много новостных статей, переведенных на казахский язык. Набор данных общедоступен для исследовательских целей. Было проведено несколько экспериментов с данным набором данных. Вместе с традиционными моделями n-грамм, мы создали модели нейронных сетей для языковой модели на основе слов. Последняя модель на основе большой параметризованной долгой краткосрочной памяти (LSTM) показывает лучшую производительность. Поскольку казахский язык считается агглютинативным языком, и он может иметь большое количество слов вне словаря (out-of-vocabulary) по неизвестном наборе данных, мы также проводили эксперименты языковых моделей на основе морфемы. Что касается экспериментальных результатов, то языковой модель на основе подслов хорошо подходит для казахского как в n-граммовых, так и в нейронных сетях по сравнению с языковой моделью на основе слов.

Ключевые слова: Языковые модели, казахский язык, n-грамм, модели нейронных сетей, языковые модели на основе морфемы.

1. Introduction

The main task of the language model is to determine whether the particular sequence of words is appropriate or not in some context, determining whether the sequence is accepted or discarded. It is used in various areas such as speech recognition, machine translation, handwriting recognition [1], spelling correction [2], augmentative communication [3] and Natural language Processing tasks (part-of-speech tagging, natural language generation, word similarity, machine translation) [4, 5, 6]. Strict rules may be required depending on the task, in which case language models are created by humans and hand constructed networks are used. However, development of the rule-based approaches is difficult and it even requires costly human efforts if large vocabularies are involved. Also usefulness of this approach is limited: in most cases (especially when a large vocabulary used) rules are inflexible and human mostly produces the ungrammatical sequences of words during the speech. One thing, as [7] states, in most cases the task of language modeling is “to predict how likely the sequence of words is”, not to reject or accept as in rule-based language modeling. For that reason, statistical probabilistic language models were developed.

A large number of word sequences are required to create the language models. Therefore the language model should be able to assign probabilities not only for small amounts of words, but also for the whole sentence. Nowadays it's possible to create large and readable text corpora consisting of millions of words, and language models can be created by using this corpus.

In this work, we first created the datasets for the language modeling experiments. We built an analogy of the Penn Treebank corpus for the Kazakh language and to do so we followed all preprocessing steps and the corpus sizes. The Penn Treebank (PTB) Corpus [8] is widely used dataset in language modeling tasks in English. The PTB dataset originally contains one million words from the Wall Street Journal, small portion of ATIS-3 material and tagged Brown corpus. Then [9] preprocessed this corpus, divided into training, validation and test sets and restricted the vocabulary size to 10k words. From then, this version of PTB corpus is widely in language modeling experiments for all state of the art language modeling experiments. We made our dataset publicly available for any research purposes. Since there are not so many open source corpora in Kazakh, we hope that this dataset can be useful in the research community.

Various language modeling experiments were performed with our dataset. We first tried traditional n-gram based statistical models, after that performed state-of-the-art Neural Network based language modeling experiments. Neural Network experiments were conducted by using the LSTM [10] cells. LSTM based neural network with large parameters showed the best result. We evaluated our language modeling experiments with the perplexity score, which is a widely used metric to evaluate language models intrinsically. As the Kazakh language is agglutinative language, word based language models might have high portion of out of vocabulary (OOV) words on unseen data. For this reason, we also performed morpheme-based language modeling experiments. Sub-word based language model is fitted well for Kazakh in both n-gram and neural net models compare to word-based language models.

2. Data preparation

We collected the datasets from the websites by using our manual Python scripts, which uses BeautifulSoup and Request libraries in Python. These collected datasets were parsed with our scripts on the basis of the HTML structure. The datasets were crawled from 4 web-pages, whose articles

originally written in Kazakh: egemen.kz, zhasalash.kz, anatili.kazgazeta.kz and baq.kz. These web-pages mainly contain news articles, historical and literature texts. There are many official web-pages in Kazakhstan which belong to state bodies and other quasi-governmental establishments where texts in Kazakh could be collected. However, in many cases, these web-pages provide the articles, which were translated from the Russian language. In these web-pages, the news articles at the beginning will be written in Russian, only then, these articles translated into Kazakh. These kind of datasets might not well show the inside nature of the Kazakh language, as during the translation, the structure of the sentences and the use of words changes. We barely see the resistant phraseological units of Kazakh in these translated articles, instead we might see the translated version of the phraseological texts in other language. [11] studied original and translated texts in Machine translation, and found out that original texts might be significantly differing from the original texts. For this reason, we excluded the web-pages which might have translation texts. We choose the web-pages whose texts originally written in Kazakh. The statistics of datasets is given in Table 1.

Table 1. **Statistics of the dataset: train, validation and test sets shown separately for each source**

Sources	# of documents	# of sentences	# of words
egemen.kz	950/80/71	21751/1551/1839	306415/22452/26790
zhasalash.kz	1126/83 /95	8663/694/751	102767/8188/9130
anatili.kazgazeta.kz	438/32/37	23668/1872/2138	311590/23703/27936
baq.kz	752/72/74	13899/1082/1190	168062/13251/14915
Overall	3266/267/277	67981/5199/5918	886872/67567/78742

After collection of the datasets, we preprocessed the datasets by following [9]. First, all collected datasets were tokenized using Moses [12] script. We added non-breaking prefixes for Kazakh in Moses, so as not to split the abbreviations. Next preprocessing steps involved: lowercasing, normalization of punctuations. After normalization of the punctuations, we removed all punctuation signs. All digits were replaced by a special sign “N”. We removed all sentences whose length is shorter than 4 and longer than 80 words and also duplicate sentences. After these operations, we restricted the vocabulary size with 10000: we found the

most frequent 10000 words and then replaced all words with '<unk>', which are not in the list of the most frequent words.

After preprocessing of the datasets, we divided our datasets into training, validation and testing sets. We tried to follow the size of the Penn Treebank corpus. Since our datasets were built from the four sources, we tried to split all sources in the same proportion into training, validation and test sets. Since, the contents in each source might differ (for example, in egemen.kz there are mostly official news, on the other hand anatili. kazgazeta.kz contains mainly historic, literature articles), we avoid having one source as training and others only for testing or validation. For this reason, we split each source with equal portions. Our datasets divided into training, validation and test sets on the document level. The statistics about training, validation and test sets is given in Table 2. Note, overall sentence and word numbers might not be the sum of all columns, because we exclude the repeated sentences. To compare the size, at the end, we provide the statistics of the Penn Treebank corpus.

Table 2. **Statistics about the training, validation and test sets**

Sources	Train set	Validation set	Test set
egemen.kz	306415	22452	26790
zhasalash.kz	102767	8188	9130
anatili.kazgazeta.kz	311590	23703	27936
baq.kz	168062	13251	14915
Overall	886872	67567	78742
Penn Tree Bank dataset	887521	70390	78669

3. n-gram based models

The main idea behind the language modeling is to predict hypothesised word sequences in the sentence with the probabilistic model. "N-gram models predict the next word from the previous N-1 words" and it is an N-token sequence of words, [13] for example, if we say two gram model (or more often it is called a bigram model) it is two-word sequence such as "Please do", "do your", "your homework" and three gram model consists of the three-word sequences and so on. As [13] states, in n-gram model, the model computes the following word from the preceding. The N-gram idea can be formulated as: given the pervious word sequence and find the probability of the next words. During the computing of probabilities of

the word sequences it's important to define the boundaries (punctuation marks such as period, comma, column or starting of the new sentence from the new line) in order to prevent the search from being computationally unmanageable.

Formulated mathematically, the goal of a language model is to find the probability of word sequences, $P(w_1, \ldots, w_n)$, and it can be estimated by the chain rule of a probability theory:

$$P(w_1, \ldots, w_n) = P(w_1) \times P(w_2|w_1) \times \ldots \times P(w_n|w_1, \ldots, w_{n-1}) \quad (1)$$

There is a notion about history, for example, in the case $P(w_4|w_1, w_2, w_3)$, (w_1, w_2, w_3) considered as the history. This probability is found based on frequency.

We can write the formula for all cases bigram and trigram models as:

$$P(w_i|w_1 \ldots w_{i-1}) \approx P(w_i|w_{i-1}), \quad (2)$$

$$P(w_i|w_1 \ldots w_{i-1}) \approx P(w_i|w_{i-2}w_{i-1}), \quad (3)$$

respectively.

This assumption helps to reduce the computation and allows probabilities to be estimated for a large corpus. Also the assumption probability of the word which depends on the previous n words (or previous 3 words for a trigram) is called a **Markov assumption**. This Markov model [14] assumes that it is possible to predict the probability of some future cases without looking deeply into the past.

By using a Markov assumption, we can find the probability of the sequence of words by the following formula:

$$P(w_1, \ldots, w_n) = \prod P(w_i|w_1 \ldots w_{i-1}) \approx \prod P(w_i|w_{i-1}) \quad (4)$$

for bigram model and for trigram:

$$\approx \prod P(w_i|w_{i-2}w_{i-1}) \quad (5)$$

Up to recently, n-gram language models widely used in all language modeling experiments. In Kazakh, n-gram based language models still used in Speech Processing [15] and Machine translation [16] tasks. We trained n-gram models with the SRILM toolkit [17] with adding 0 smoothing technique. For our dataset, using of the modified Kneser-Ney [18] or Katz backoff [19] algorithms showed poor results, (543.63 on the test set), as there are many infrequent words replaced by '<unk>' sign, and only high gram models might work well. Adding 0 smoothing technique showed best performance for n-gram models. The results are given in Table 3.

4. Neural LSTM based models

In this experiment, we performed Neural LSTM-based language models. There are many types of neural architectures, which also applied successfully for the language modeling tasks. Starting from the work of [20] there are many Recurrent Neural Architectures proposed. With Recurrent Neural Networks, it's possible to model the word sequences, as the recurrence allows to remember the previous word history. Recurrent Neural Network can directly model the original conditional probabilities:

$$P(w_1, \ldots, w_n) = \prod P(w_i|w_1 \ldots w_{i-1}) \quad (6)$$

To model the sequences, f function constructed via recursion, initial condition is given by $h_0 = 0$ and the recursion will be $h_t = f(x_t, h_{t-1})$. Here, h_t is called hidden state or memory and it memorizes the history from x_1 up to x_{t-1}. Then, the output function is defined by combination of h_t function:

$$P(w_1, \ldots, w_n) = g_w(h_t) \quad (7)$$

f can be any nonlinear function such as tanh, ReLU and g can be a softmax function.

In our work, we followed [21] who presented a simple regularization technique for Recurrent Neural Networks (RNNs) with LSTM [10] units. [22] proposed dropout technique for regularizing the neural networks, but this technique does not work well with RNNs. This regularizing technique is tent to have overfitting in many tasks. [21] showed that the correctly applied dropout technique to LSTMs might substantially reduce the overfitting in various tasks. They tested their dropout techniques on language modeling, speech recognition, machine translation and image caption generation tasks.

In general, LSTM gates' equations given as follow:

$$f_t = \sigma(W_f[C_{t-1}, h_{t-1}, x_t]+b_f]); \quad (8)$$

$$i_t = \sigma(W_i[C_{t-1}, h_{t-1}, x_t]+b_i]); \quad (9)$$

$$o_t = \sigma(W_o[C_t, h_{t-1}, x_t]+b_o]) \text{ and} \quad (10)$$

$$g_t = \tanh(W_g[C_t, h_{t-1}, x_t]+b_g]). \quad (11)$$

Then the state values computed by using the above gates:

$$c^l_t = f \odot c^l_{t-1} + i \odot g \text{ and} \quad (12)$$

$$h^l_t = o \odot \tanh(c^l_t). \quad (13)$$

The dropout method by [21] can be described as follows: if there is a dropout operator, then it forces the intermediate computation to be more robustly, as the dropout operator corrupts the information carried by the units. On the other hand, in order not to erase all the information from the units, the units remember events that occurred many time steps in the past.

We also implement our[1] LSTM based Neural Network models using TensorFlow [23]. We trained regularized LSTMs of three sizes: the small LSTM, medium LSTM and large LSTM. Small sized model has two layers and unrolled for 20 steps. Medium and large LSTMs have two layers and are unrolled for 35 steps. Hidden size differs in three models: 200, 650 and 1500 for small, medium and large models respectively. We initialize the hidden states to zero. We then use the final hidden states of the current minibatch as the initial hidden state of the subsequent minibatch.

Our experiments showed that the LSTM based neural language modeling outperforms the n-gram based models. Large and Medium LSTM models shows better results than the n-gram add 0 smoothing method (Note, for n-gram Kneser-Ney discounting method we got poor results). Our experiments shows that the using of the Neural based language models have better performance for Kazakh. The results are given in Table 3.

Table 3. **Word-based language modeling results**

	n-gram	Neural LM		
		small	medium	large
Train perplexity	93.81	68.522	67.741	**63.185**
Validation perplexity	129.6537	143.871	118.875	**113.944**
Test perplexity	123.7189	144.939	118.783	**115.491**

5. Sub-word based language models

In the last section, we experimented with the sub-word based language models. The Kazakh language as other Turkic languages is an agglutinative language, the word forms can be obtained by adding the prefixes. This agglutinative nature may lead on having the high degree of the out-of-vocabulary (OOV) words on unseen data. To solve this problem, depending on the characteristics of individual languages,

[1] https://github.com/Baghdat/LSTM-LM

different language model units were proposed. [24] studied different word representations, such as morphemes, word segmentation based on the Byte Pair Encoding (BPE), characters and character trigrams. Byte Pair Encoding, proposed by [25], can effectively handle rare words in Neural Machine Translation and it iteratively replaces the frequent pairs of characters with a single unused character. Their experiments showed that for fusional languages (Russian, Czech) and for agglutinative languages (Finnish, Turkish) character trigram models perform best. Also, [26] considered syllables as the unit of the language models and tested with different representational models (LSTM, CNN, summation). As they stated, syllable-aware language models fail to outperform character-aware ones, but usage of syllabification can increase the training time and reduce the number of parameters compared to the character-aware language models.

By considering these facts, in this section we experimented with the sub-word based models. Morfessor [27] is a widely tool to split the datasets into morpheme-like units. It used successfully in many agglutinative languages (Finnish, Turkish, Estonian). As for now, there is no syllabification tool for Kazakh, we also used Morfessor tool to split our datasets into morpheme like units.

After splitting the datasets, we performed language modeling experiments on morpheme like units. The results are given in Table 4. By looking at the results, we can say that splitting the words into morpheme-like units benefits in terms of OOV and perplexity in both n-gram and neural net based models.

Table 4. **Morph-based language modeling results**

	n-gram	Neural LM		
		small	medium	large
Train perplexity	32.39255	19.599	24.999	25.880
Validation perplexity	44.11561	50.904	41.896	40.876
Test perplexity	44.39559	47.854	38.180	**37.556**

6. Conclusion

In this work we created analogy of the Penn TreeBank corpus for the Kazakh language. To create the corpus, we followed all instructions for preprocessing and the size of the training, validation and test sets. This

dataset is publicly available for the research purposes. We conducted language modeling experiments on this dataset by using the traditional n-gram and LSTM based neural networks. We also explored the sub-word units for the language modeling experiments for Kazakh. Our experiments showed that neural based models outperforms the n-gram based models and splitting the words into morpheme-like units has advantage compared to the word based models. In future, we are going to create the hyphenation tool for the Kazakh language, as Morfessor's morpheme-like units are data-driven and sometimes there are incorrect morpheme-like units.

Acknowledgement. This work has been funded by the Nazarbayev University under the research grant No129-2017/022-2017 and by the Committee of Science of the Ministry of Education and Science of the Republic of Kazakhstan under the research grant AP05134272.

REFERENCES

1. Russell S. and Norvig P. Artificial Intelligence: A Modern Approach (2nd Ed.) // Pretice Hall. 2002.

2. Kukich K. Techniques for automatically correcting words in text // ACM Computing Surveys.1992. 24(4), pp. 377–439.

3. Newell A., Langer S. and Hickey M. The role of natural language processing in alter-native and augmentative communication // Natural Language Engineering. 1998. 4(1). pp. 1–16.

4. Church K.W. A stochastic parts program and noun phrase parser for unrestricted text // In Proceedings of the Second Conference on Applied Natural Language Processing. 1988. pp. 136–143.

5. Brown P.F., Cocke J., DellaPietra S.A., DellaPietra V.J., Jelinek F., Lafferty J.D., Mercer R.L., and Roossin P.S. A statistical approach to machine translation // Computational Linguistics. 1990. 16(2). pp. 79–85.

6. Hull J.J. Combining syntactic knowledge and visual text recognition: A hidden Markov model for part of speech tagging in a word recognition algorithm // In AAAI Symposium: Probabilistic Approaches to Natural Language. 1992. pp. 77–83.

7. Whittaker E. W. D. Statistical Language Modelling for Automatic Speech Recognition of Russian and English // PhD thesis, Cambridge University, Cambridge. 2000.

8. Marcus M.P., Marcinkiewicz M.A. and Santorini B. Building a large annotated corpus of English: The penn treebank // Computational linguistics. 1993. 19(2). pp. 313–330.

9. Mikolov T., Kombrink S., Burget L., Černocký J. and Khudanpur S. Extensions of recurrent neural network language model // In Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on. 2011. pp. 5528–5531. IEEE.

10. Hochreiter S. and Schmidhuber J̈. Long short-term memory // Neural computation. 1997. 9(8). pp. 1735–1780.

11. Lembersky G., Ordan N. and Wintner S. Language models for machine translation: Original vs. translated texts // Computational Linguistics. 2012. 38(4). pp. 799-825.

12. Koehn P., Hoang H., Birch A., Callison-Burch C., Federico M., Bertoldi N., Cowan B., Shen W., Moran C., Zens R. and Dyer C. Moses: Open source toolkit for statistical machine translation // In Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions. Association for Computational Linguistics. 2007. pp. 177–180.

13. Jurafsky D. and Martin J. H. Speech and Language Processing (2nd Ed.) // Pretice Hall. 2009.

14. Markov A.A. Primer statisticheskogo issledovaniya nad tekstom "Evgeniya Onegina", illyustriruyushchij svyaz' ispytanij v tsep'. [Example of a statistical investigation illustrating the transitions in the chain for the "Evgenii Onegin" text.] // Izvestiya Akademii Nauk. 1913. pp. 153–162.

15. Kozhirbayev Zh, Karabalayeva M. and Yessenbayev Zh. Spoken term detection for Kazakh language // in Proceedings of the 4-th International Conference on Computer Processing of Turkic Languages "TurkLang 2016". 2016. pp. 47–52.

16. Myrzakhmetov B. and Makazhanov A. Initial Experiments on Russian to Kazakh SMT // Research in Computing Science. 2017. vol. 117. pp. 153–160.

17. Stolcke, A. SRILM – an extensible language modeling toolkit // In Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP). 2002. pp. 901–904. URL: http://www.speech.sri.com/projects/srilm/.

18. Kneser R. and Ney H. Improved backing-off for m-gram language modeling // In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing. 1995. vol. 1. pp. 181–184.

19. Katz S. M. Estimation of probabilities from sparse data for the language model component of a speech recognizer // IEEE Transactions on Acoustics, Speech and Signal Processing. 1987. 35(3). pp. 400–401.

20. Bengio Y., Ducharme R., Vincent P. & Jauvin C. A neural probabilistic language model // Journal of machine learning research. 2003. pp. 1137–1155.

21. Zaremba W., Sutskever I. and Vinyals O. Recurrent neural network regularization // arXiv preprint arXiv:1409.2329. 2014.

22. Srivastava N., Hinton G., Krizhevsky A., Sutskever I. & Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting // The Journal of Machine Learning Research. 2014. 15(1). pp. 1929–1958.

23. Abadi M., Barham P., Chen J., Chen Zh., Davis A., Dean J., Devin M., Ghemawat S., Irving G., Isard M. and Kudlur M. Tensorflow: a system for large-scale machine learning // In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. USENIX Association. 2016. pp. 265–283.

24. Vania, C., & Lopez, A. From Characters to Words to in Between: Do We Capture Morphology? // In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017. Volume 1: Long Papers. Vol. 1, pp. 2016–2027.

25. Sennrich, R., Haddow, B., & Birch, A. Neural Machine Translation of Rare Words with Subword Units // In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016. Volume 1: Long Papers. Vol. 1, pp. 1715–1725.

26. Assylbekov Z., Takhanov, R., Myrzakhmetov, B., & Washington, J. N. Syllable-aware Neural Language Models: A Failure to Beat Character-aware Ones // In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017. pp. 1866–1872.

27. Smit P., Virpioja S., Grönroos S. A. & Kurimo M. Morfessor 2.0: Toolkit for statistical morphological segmentation // In The 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL), Gothenburg, Sweden, April 26–30, 2014. Aalto University.

УДК 004 827

RESOLUTION OF ELLIPSES IN PLANIMETRIC TASK TEXTS ON THE BASIS OF COGNITIVE OBJECT MODELLING

X. A. Naidenova[1], S. S. Kurbatov[2], V. P. Ganapol'skii[1]
[1]Military medical academy, Saint Petrsburg
[2]Research Center of Electronic Computer Engineering, Moscow
ksenniadd@gmail.com, curbatow.serg@yandex.ru, ganvp@mail.ru

The article describes the processing of ellipses in an automated system of solving planimetric tasks according to their description in natural language. The resolution of ellipses is based on using the syntactic structures and semantics of geometry in parallel. The types of ellipses most frequently encountered in geometric tasks are revealed. A new approach to recognizing and resolving ellipses in the framework of cognitive semantics is offered.

Keywords: ellipsis, planimetry, natural language processing, cognitive modelling.

РАЗРЕШЕНИЕ ЭЛЛИПСИСОВ В ТЕКСТАХ ГЕОМЕТРИЧЕСКИХ ЗАДАЧ НА ОСНОВЕ КОГНИТИВНЫХ МОДЕЛЕЙ ГЕОМЕТРИЧЕСКИХ ОБЪЕКТОВ

К. А. Найденова[1], С. С. Курбатов[2], В. П. Ганапольский[1]
[1]Военно-медицинская академия, Санкт Петербург
[2]НИИ электронной вычислительной техники, Москва
ksennaidd@gmail.com, curbatow.serg@yandex.ru, ganvp@mail.ru

В статье описывается обработка эллипсисов в автоматизированной системе решения планиметрических задач по их описанию на естественном языке. Разрешение эллипсиса базируется на параллелизме синтаксических структур и использовании семантики геометрии. Даны примеры обработки и описаны ее ограничения. Выделены типы эллипсисов, наиболее часто встречающиеся в геометрических задачах. Предложен подход к распознаванию эллипсисов и их разрешению в рамках когнитивной семантики

Ключевые слова: эллипсис, планиметрия, обработка естественно-языковых текстов, когнитивное моделирование.

1. Введение

Неоднозначность естественного языка, обусловленная омонимией, давно изучается компьютерной лингвистикой. Однако неоднозначность, связанная с пропуском в тексте подразумеваемой

языковой единицы (эллипсис), стала активно анализироваться при автоматической обработке естественного языка сравнительно недавно [1,2]. И хотя в теоретической лингвистике эллиптичность получила достаточное освещение [3,4], восстановление эллипсиса в системах синтаксического анализа развито явно недостаточно.

Во многом это связано, во-первых, с тем, что устранение эллиптичности носит подчиненный характер по отношению к собственно синтаксическому анализу, а во-вторых, со сложностью восстановления эллипсиса. Сложность связана с необходимостью учета ряда контекстов − текущее предложение, соседние предложения, уже установленные синтаксические связи и, наконец, с семантикой текста.

Данная работа состоит из двух частей, в первой описывается обработка эллиптичности в интегральной системе решения планиметрических задач по их описанию на естественном языке, реализованной в рамках проекта INTEGRO (INTEGRating Ontology) [5]. Общая схема целостной системы и принципы ее работы приведены в [6]. Во второй части предлагается подход к обработке эллипсисов на основе когнитивной семантики, ориентированный на значительно более широкий спектр применений (как в отношении предметных областей, так и разных естественных языков).

1. Обработка эллипсисов в системе автоматического решения геометрических задач

1.1. Синтаксическая анализ

Общая схема целостной системы приведена на Рис. 1. В данной работе описывается обработка эллиптичности блоком «Лингвистический транслятор».

Лингвистический транслятор формирует синтаксическую структуру предложения и определяет, что некоторые ее элементы нарушают нормы языка. Например, отсутствует существительное для прилагательного, предлог стоит в конце предложения, у числа отсутствует обязательная единица измерения и т. п. Основные критерии определения эллиптичности исследованы лингвистами [7]. Базируясь на этих критериях, записанных в онтологии системы, лингвистический транслятор выявляет элементы синтаксической трансляции, претендующие на эллиптичность. Далее с помощью алгоритма, описанного ниже в разделе 1.2., выполняется восстановление эллипсиса.

Рис. 1. Общая схема целостной системы

Например, для предложения «Радиус первой окружности равен 12 см, а второй 10 см» элементы «второй» и «10» определяют эллиптичность. В результате ее разрешения должны быть сформированы две синтаксические структуры, соответствующие предложениям:

- Радиус первой окружности равен 12 см;
- Радиус **второй** окружности равен **10** см.

Эти структуры далее обрабатываются системными механизмами перефразирования для получения их онтологического представления в формальных терминах предметной области [6].

Концепция перефразирования предложена известным русским лингвистом Апресяном [8]. Мы используем вариант этой концепции, адаптированный для целей рассматриваемой системы. Концепция перефразирования предполагает, что в классе перефраз есть одна простейшая, или каноническая перефраза, составленная исключительно из таких лексем, которые наиболее четко выражают базовые понятия. Концепция основана на гипотезе, что << Одно из фундаментальных свойств человеческих языков состоит в том, что если в них есть несколько синонимичных в широком смысле средств выражения какого-то концепта, ровно одно из них оказывается привилегированным, каноническим, или прототипическим способом выражения этого концепта.>> [8].

Правила перефразирования для геометрии обеспечивают получение для предложений типа «точка, находящаяся на прямой», «прямая, проходящая через точку», «принадлежащая прямой точка»,

«точка, лежащая на прямой» и т. п. канонического описания «точка принадлежит прямой». Далее это каноническое описание отображается в семантическое (онтологическое) представление, а именно в триплет <<точка лежит прямая>>. Подчеркнем, что входящие в триплет объекты и отношение не зависят от конкретного естественного языка. Соответствующее правило перефразирования в левой части содержит объекты, зависящие от языка, а в правой − формальные объекты, инвариантные по отношению к различным естественным языкам.

Правила перефразирования разбиваются на два класса, в первый входят правила, у которых обе части являются обобщенными синтаксическими структурами (то есть содержат концепты «существительное», «прилагательное», «причастие» и т. д.). Во второй − правила с каноническим описанием в левой части и семантическим описанием − в правой. Правила второго класса используются для синтеза естественно-языкового описания из онтологической структуры. Правила первого класса целесообразно применять для синонимических преобразований синтезированной структуры с получением наиболее приемлемого текста для данной предметной области.

1.2. Алгоритм обработки эллипсиса

Алгоритм обработки эллипсиса основан на знаниях в онтологии, отражающих семантическую иерархию словоформ синтаксической структуры и фиксирующих нормы естественного языка. В первом приближении алгоритм можно описать следующим образом:

- сегментировать синтаксическую структуру на полную и эллиптичную (в общем случае это множество именных групп);
- в эллиптичной структуре выявить элементы, которые целесообразно путем сопоставления с элементами полной структуры использовать для разрешения эллиптичности;
- в полной синтаксической структуре выявить кандидатов на замещение элементами, найденными на предыдущем шаге;
- выполнить замещение и получить полную синтаксическую структуру.

В примере из раздела 1.1. «*первый*» замещается на «*второй*», а «*12*» на «*10*», ибо они соответствуют одним концептам онтологии. Здесь разные экземпляры объектов одного и того же типа и один атрибут (длина). При разрешении данным алгоритмом эллипсиса в предложении «*периметр треугольника составляет 37 см, а пло-*

щадь − 20 см²» объект будет один, а атрибуты разные. При внешней простоте алгоритма он позволяет успешно восстанавливать не только геометрические эллипсисы, но и ряд других, описанных, например, в [2]: «*двадцать лет такого танца составляют эпоху, сорок – историю*». Здесь «двадцать» заменяется на «сорок», а «эпоху» заменяется на «историю». Естественно, что онтология должна включать концепт «*временной_интервал*», связывающий в иерархии «эпоху» и «историю».

Рассмотрим пример «*он пошел в аптеку, а его брат – на почту».* Здесь «*он*» замещается именной группой «*брат чей? его*». Именная группа может быть расширена, например, «*его сводный брат по материнской линии*». Сопоставление в онтологии формально выполняется для вершин именных групп. Именно поэтому корректно восстанавливается структура для предложения «Треугольник ABC находится внутри окружности, а квадрат − вне».

1.3. Ограничения

Многие случаи эллиптичности не смогут обрабатываться вышеприведенным алгоритмом. Например, в предложении «*Имеется 7 окружностей. Радиус одной 5 см, двух других − 3, а остальных − 10 см*» содержится множественная эллиптичность. Аналогичный пример из [2]: «*Актинии отбрасывают щупальца, раки – клешни, ящерицы – хвост*». В ряде случаев возникает неоднозначность на уровне сопоставления. Для разрешения неоднозначности анализировались два варианта:

- продолжение работы и устранение неоднозначности на этапе семантической обработки (канонические синтаксические структуры);
- ввод в онтологию не только знаний об иерархии концептов, но и правил предпочтения при выборе кандидата на замещение.

Решение о целесообразности выбора одного из вариантов (или их комбинации) является предметом дополнительного исследования.

Отметим, что описанный алгоритм тестировался не только в геометрических текстах, но и в ряде других, в частности, в текстах так называемой «космической верфи» [9]. В соответствии с иерархией объектов этой области успешно восстанавливаются предложения типа «*Радиус первой цистерны равен 1.7 см*, *а второй 5.8*» или «*Длина солнечной панели равна 15 см*, *а вес 40 граммам*».

Тип эллипсиса с пропущенными модальными глаголами с отрицанием и без отрицания для английского языка рассматривается

в работах: [10; 11; 12; 13;14]. В работе [10] отмечается, что разрешение глагольных эллипсисов плохо поддается методам машинного обучения. Однако выделены случаи с модальными глаголами и для них предложены полностью автоматические методы резолюции эллипсисов без проведения предварительной обработки или аннотации корпуса текстов. Эти случаи также не требуют сложных семантических и прагматических рассуждений. Резолюция эллипсиса выполняется системой OntoSem с когнитивной структурой OntoAgents [11].

В этих статьях текст с глагольным эллипсисом имеет следующую структуру, состоящую из 2х частей, стоящих справа и слева от «тире» (обе части находятся в одном предложении). Правая часть содержит пропуск глагола, левая часть (антецедент) содержит глагол, которым нужно дополнить правую часть. Резолюция такого эллипсиса делится на три этапа:

- Собственно, распознать наличие эллипсиса, локализовать его и выделить его части;
- Найти ближайший слева глагол в антецеденте;
- Разрешить эллипсис (то есть, вставить найденный глагол на место тире).

Недавно появилась работа [15], в которой предлагается новый метод разрешения множественных эллипсисов в таких предложениях, как:

- Unemployment has reached 27.6% in Azerbaijan, 25.7% in Tadzhikistan, 22.8% in Uzbekistan, 18.8% in Turkmenia, 18% in Armenia and 16.3% in Kirgizia;
- Eve gave flowers to Al and Sue to Paul;
- Eve gave a CD to Al and roses to Sue;
- Arizona elected Goldwater Senator, and Pennsylvania Schwelker;
- I want to try to begin to write a novel and Mary a play.

Разрешение эллипсисов основано на использовании деревьев зависимости между клаузами и достигает точности в 65–75 % случаев.

Следует отметить, что вопрос о четком критерии эллиптичности и методах восстановления полной структуры предложения не решен полностью в рамках общепринятой лингвистической теории, основанной на семантико-синтаксическом разборе предложений. Синтаксис выявляет структуру эллипсиса и аналогичной части предложения без эллипсиса, семантика имеет дело со значениями слов. Однако, как показывает пример из [16], разрешение эллипси-

сов основано на понимании контекста (темы текста), смысла слов и словосочетаний: «*Charles makes love with his wife twice a week. So does John*».

В следующих разделах рассматриваются тексты планиметрических задач и предлагается подход к разрешению различных типов эллипсисов на основе когнитивных моделей геометрических объектов, действий и отношений между ними.

2. Обработке эллипсисов на основе когнитивной семантики

2.1. Классификация эллипсисов в геометрических задачах

Для изучения типологии эллипсисов в геометрических задачах мы использовали корпус текстов, включающий следующие источники: (Ш) И.Ф. Шарыгин «Задачи по геометрии. Планиметрия». – Москва: Наука, Библиотека «Квант», выпуск 17, 1982; (Г) Э.Г. Готман и З.А. Скопец «Решение геометрических задач аналитическим методом». Пособие для учащихся 9 и 10 классов». – Москва: Просвещение, 1979; (А) Н.П. Антонов [и др.]. Сборник задач по элементарной математике. – Москва: Наука, 1979; (К) А.П. Киселев. Геометрия. – Москва: ФИЗМАТЛИТ, 2014; (КОЛМ) А.Н. Колмогоров [и др.]. Геометрия. Учебное пособие для 8 класса средней школы. – Москва: Просвещение, 1980; (Б) И.Л. Бабинская. Задачи математических олимпиад. – Москва: Наука, 1975. Далее примеры задач приводятся в тексте с указанием первой буквы фамилии автора, номера задачи и страницы, на которой она находится.

Нами выделено несколько типов эллипсисов: эллипсисы с использованием знака « – » (эллипсисы с пропущенным предикатом, эллипсисы с пропущенным глаголом – Тип 1.А и Тип 1.Б), эллипсисы без использования знака « – » (эллипсисы с пропущенным существительным, с пропущенным местоимением, с пропущенным предикатом – Тип 2).

Отметим, что в некоторых задачах встречается только один из видов эллипсисов, но существуют задачи, в которых имеются несколько эллипсисов разных типов. Кроме того, один и тот же тип эллипсиса может встретиться в задаче несколько раз.

Рассмотрим выделенные типы эллипсисов с точки зрения их свойств и структуры.

Тип 1.А Пропущен предикат при наличии тире

Примеры:

(Ш56, стр. 28) *В треугольнике ABC даны R и r − радиусы описанной и вписанной окружностей. A_1, B_1, C_1 − точки пересечения биссектрис треугольника ABC с описанной окружностью*;

(Ш22, стр. 24) *...Пусть O_1 O_2 O_3 O_4 − центры окружностей, описанных соответственно около треугольников ABM, BCM, CDM и DAM*;

(Ш93, стр. 33) *Основания перпендикуляров, опущенных из B и D на AC − M и N*;

(Ш124, стр.37) Дан прямоугольный треугольник ABC, угол *C – прямой*, O *– центр вписанной окружности*, *M – точка касания вписанной окружности с гипотенузой*;

(Ш162, стр. 43) Доказать, что середины сторон треугольника, основания высот и середины отрезков высот от вершин до точки их пересечения лежат на одной *окружности* – «*окружности девяти точек*» (задача Эйлера);

(Ш128 стр. 37) Дан треугольник ABC*, D − произвольная точка плоскости*;

(Ш122 стр. 37) Доказать, что *данный четырехугольник – ромб*;

(КОЛМ4 стр. 45) Докажите, что: а) всякая трапеция, вписанная в окружность, равнобедренная; б) *всякий параллелограмм, вписанный в окружность, – прямоугольник*; в) *всякий ромб, вписанный в окружность*, *– квадрат*;

(Ш124, стр. 37) Дан прямоугольный треугольник ABC, *угол C – прямой*;

Ш275, стр. 56) В треугольнике ABC *угол B − средний по величине:* A < B < C;

(Б301 стр 34) Среди точек данной прямой *l* найти такую, что *сумма расстояний от нее до двух данных точек A, B – минимальная*;

(Ш85, стр. 32) Пусть a, b, c и d – длины сторон вписанного четырех угольника (a, c – противоположные стороны), ha, hb, hc hd – расстояния от центра описанного круга до соответствующих сторон. Доказать, что если центр круга – внутри четырехугольника, то ahc cha = bhd dhb.

Структурными компонентами этих эллипсисов являются Именные Группы и Предложные Группы (ПГ). Под именной группой (ИГ) мы будем понимать фрагмент предложения, включающий существительное или местоимение с определяющими его словами, включая причастные и уточняющие обороты с предложными группами. Выделение ИГ в предложениях реализуется в системе OntoIntegrator

[17] в рамках проекта по созданию the World Digital Mathematical Library – WDML.

Рассмотрим этот тип эллипсиса более подробно. Этот тип эллипсиса имеет следующую структуру:

а) <именная группа>< – ><обозначение(я)> (*Основания перпендикуляров, опущенных из B и D на AC, − M и N*);

б) <обозначение(я)>< – ><именная группа> (*D − произвольная точка плоскости*);

в) <именная группа>< – ><именная группа> (*данный четырехугольник – ромб*);

г) <именная группа>< – ><предложная группа> (*C между A и B*);

д) <именная группа>< – ><свойство, выраженное прилагательным> (*угол C – прямой*).

Если ИГ начинается с существительного в именительном падеже единственного или множественного числа, то можно рассматривать упрощенные структурные формы:

а*) обозначение(я) – объект(ы) (O1 O2 O3 O4 − центры ……);

б*) объект(ы) – обозначение(я) (основания …. – M N).

На месте « – » в случаях а), а*), б) и б*) подразумевается форма глагола «быть»: ЕСТЬ.

В случае а) тире можно удалить, если именная группа содержит только наименование объекта(ов): объект(ы)обозначение(я). Аналогично, в случае б) тире можно удалить, если именная группа содержит только наименование объекта(ов): объект(ы)обозначение(я).

Случаи а) и б) не требуют анализа контекста предложения, фрагменты текста, содержащие эллипсис, связаны с введением в рассмотрение определенных объектов и их обозначений.

Разрешение эллипсиса можно проводить по схеме:

- Выделение именных групп;
- Распознавание вершин именных групп как геометрических объектов;
- Распознавание обозначений;
- Распознавание тире между обозначением(ями) и именной группой(ами);
- Проверка по правилам онтологии соответствия между обозначением и объектом (вершиной ИГ);
- Разрешение эллипсиса с заменой тире на «это есть» или удалением тире.

Случаи в) и г) в геометрических задачах относительно редки (по крайней мере, в рассматриваемом нами корпусе задач).

Тире в русском языке ставится в разнообразных ситуациях. Однако в случае в), согласно справочнику Розенталя [18], мы имеем ситуацию, когда тире ставится между подлежащим и сказуемым при отсутствии связки, если оба члена предложения (обе вершины именных групп) выражены существительным в форме одного и того же падежа, например, «*одиночество в творчестве – тяжелая штука*», «*следующая станция – Мытищи*». В геометрических задачах случай в) имеет характер логического определения (например, «*геометрия – раздел математики, изучающий пространственные формы и отношения тел*») или тождества, когда подлежащее и сказуемое выражают одно и то же понятие:

- точки их пересечения лежат на одной окружности – окружности девяти точек;
- данный четырехугольник – ромб.

На месте « – » в случае в) подразумевается отношение ЭТО ЕСТЬ или просто ЕСТЬ.

Отметим, что в геометрических задачах в случае в) существительные, соответствующие объектам, не обязательно стоят в именительном падеже. В этом случае если вершины ИГ представляют собой один и тот же объект, что выражается одним и тем же словом (единственного или множественного числа), то разрешение эллипсиса не представляет сложности. Если вершины ИГ не выражаются одним и тем же словом, то выполнение отношения ЭТО ЕСТЬ необходимо проверить через логический вывод в онтологии.

Случай г) более сложный: *центр круга – внутри четырехугольника; C между A и F*.

Выражение с эллипсисом состоит из двух частей со значением субъекта и определяющего обстоятельства, и построено по схеме «что – где» [18]. Для разрешения подобного эллипсиса необходимо учитывать контекст задачи. В задаче (Ш85 стр. 32) контекст ясен уже из первых двух предложений: речь идет о вписанном четырехугольнике (первое предложение задачи) и описанной вокруг четырехугольника окружности (второе предложение задачи). Таким образом, в третьем предложении может идти речь только о центре описанной окружности и о месте его нахождения по отношению к четырехугольнику. Отметим, что учет контекста подразумевает включение в разрешение эллипсиса правдоподобных рассуждений.

Эллипсис с пропущенным глаголом «находиться», «быть в чем-то», «быть где-то» разрешается с помощью вывода в онтологии описанной выше системы.

В случае г) (<именная группа>< – ><свойство>): *угол С – прямой*; *угол В – средний по величине*; сумма расстояний от нее до двух данных точек А, В – *минимальная*. Свойство выражено через прилагательное. На месте « – » в этом случае подразумевается форма глагола «быть»: ЕСТЬ.

Тип 1.Б Пропущен глагол, тире присутствует

Примеры:

(Ш35, стр. 25) *В треугольнике АВС взяты точки М, N и Р: М и N − на сторонах АС и ВС, Р − на отрезке MN, причем |AM|/|MC| = |CN|/|NB| = |MP|/|PN|*;

(Ш62, стр. 29) *На плоскости даны две прямые, пересекающиеся в точке О, и две точки А и В. Обозначим основания перпендикуляров, опущенных из А на данные прямые, через М и N, а основания перпендикуляров, опущенных из В, − через К и L*;

(КОЛМ16, стр. 57) Внутри квадрата A_1 A_2 A_3 A_4 взята точка Р. Из вершины A_1 проведена прямая, перпендикулярная к прямой A_2Р, *из вершины* A_2 *– к прямой* A_3Р, *из вершины* A_3 *– к прямой* A_4Р *и из вершины* A_4 *– к прямой* A_1Р;

(КОЛМ12, стр. 41) Докажите, что величина угла с вершиной внутри круга равна полу сумме угловых величин двух дуг, *из которых одна заключена между сторонами этого угла, а другая – между продолжениями сторон*.

В общем случае с пропущенным глаголом необходимо восстановить всю глагольную группу (ГГ), с входящими в нее именными и предложными группами, причем часто – не один раз (как в задаче КОЛМ16).

Рассмотрим один из трудных случаев глагольного эллипсиса Ш35. В этом предложении мы имеем неполные глагольные группы. Фраза «В треугольнике АВС взяты точки М, N, и Р» может рассматриваться как предпосылка (presupposition), и глагольная группа на самом деле включает три утверждения:

В треугольнике АВС взята точка М на стороне АС;

В треугольнике АВС взята точка N на стороне ВС;

В треугольнике АВС взята точка Р на отрезке MN.

Реконструкция этого предложения поддерживается мыслимой геометрической ситуацией (назовем ее **когнитивной моделью геометрической ситуации**). Восстановление идет последовательно, но с одновременным моделированием: временных (раньше, позже), референциальных (обозначение отсылает к объекту, местоимение

ссылается на объект), пространственных (в треугольнике, на стороне), лингвистических (связи отношений, объектов, свойств с определенными словами и словосочетаниями), количественных и т. д. В нашем примере мы имеем (→ обозначает ссылку):

Треугольнике ABC → треугольник → обозначение = ABC;

Треугольнике ABC → один → этот → заданный;

В треугольнике ABC → сторона AC → одна, первая; сторона BC → два, вторая; сторона AB → три, третья;

В треугольнике ABC взяты точки M, N и P;

Точка один → обозначение = M → первая; точка два → обозначение = N → вторая;

Точка три → обозначение = P → третья.

Теперь мы нуждаемся в понимании действия "взять точку в треугольнике" и выдвижении гипотез "Где?". В соответствии с одной из гипотез, можно рассмотреть следующие случаи:

В треугольнике ABC взята точка M (первая) на стороне AC (первой);

В треугольнике ABC взята точка N (вторая) на стороне BC (второй);

И по аналогии:

В треугольнике ABC взята точка P (третья) на отрезке MN.

Отрезок → обозначение = MN → соединяет точки M и N (поддерживается знанием о том, как порождается отрезок).

Принятая гипотеза согласуется с текстом задачи (с геометрической ситуацией).

В результате порождается полный текст этой задачи: *В треугольнике ABC взята точка M на стороне AC*, *взята точка N на стороне BC и взята точка P на отрезке MN.*

Процесс связывания объектов во время их конструирования поддерживается когнитивными моделями объектов и операциональным знанием. Как отмечает Д. Сулейманов в [19], «надо идти не от текста, а от задачи». Все когнитивные модели могут быть явно определены на основе геометрической семантики и ассоциированы с частями речи и типичными словосочетаниями с их грамматическими ролями на уровне предложения.

Восстановление полного текста требует рассуждений по аналогии и понимания смысла действий с геометрическими объектами. Именно, аналогичные действия предполагаются с аналогичными объектами, и поэтому пропускаются слова. На практике большин-

ство пропущенных слов является избыточным для понимания смысла предложения. Человек пропускает слова сознательно. Однако если пропущенная информация не является избыточной, то заполнение пробелов представляет проблему, которая разрешается с помощью анализа возникающих геометрических ситуаций.

Мы должны предполагать существование когнитивной аналогии, которая порождает и конструктивную ГЕОМЕТРИЧЕСКУЮ АНАЛОГИЮ. Когнитивные структуры соответствуют смысловым структурам, излагаемых в тексте геометрических ситуаций, и именно поэтому они определяют структуру ИГ, ПГ и ГГ. Поддержке разрешения эллипсисов со стороны когнитивной семантики посвящается следующий раздел.

В любом случае процесс резолюции эллипсисов должен включать выделение именных и глагольных групп, а также их сравнение для определения аналогичных составляющих. Для неполных групп важно выявить возможные их границы, и чаще всего, одна из границ или даже обе в тексте проявлены.

Рассмотрим фрагмент задачи (Ш68, стр. 30): на AB взята точка M так, что угол MCA = 60°; на стороне CB − N так, что угол NAC = 50°. Найти угол NMA.

Здесь глагольная группа имеет следующую структуру:

«на ИМЕННАЯ ГРУППА-1 взята ИМЕННАЯ ГРУППА-2 так, что ИМЕННАЯ ГРУППА-3 = 60°»; часть с эллипсисом имеет вид:

«на ИМЕННАЯ ГРУППА-4 − ИМЕННАЯ ГРУППА-5 так, что ИМЕННАЯ ГРУППА-6 = 50°».

Отметим, что именные группы 1 и 4, 2 и 5, 3 и 6 имеют в качестве вершин аналогичные геометрические объекты, и по аналогии на место тире следует поставить глагол ВЗЯТЬ.

Тип 2 Пропуск существительного, местоимения или предиката при отсутствии тире

Дадим только некоторые примеры эллипсисов этого типа.

Пропущено существительное

(Ш63, стр. 29) Две окружности касаются друг друга внутренним образом в точке A. Из центра большей окружности проведен радиус OB, касающийся *меньшей* в точке C. Найти угол BAC. (После слова «меньшей» пропущено слово «окружности»).

(Ш79, стр. 31) В треугольнике проведены три прямые, параллельные его сторонам и касающиеся вписанной окружности. Они отсекают от *данного* три треугольника. (После слова «данного» пропущено слово «треугольника»).

(Б312 стр. 35) Через середину гипотенузы прямоугольного треугольника проведен перпендикуляр. Отрезок этого перпендикуляра, заключенный внутри треугольника, равен 3 см, а *вне* треугольника (до пересечения с продолжением другого катета) 9 см. (Перед словом «вне» пропущена целиком ИГ: «отрезок этого перпендикуляра, заключенный», кроме того, пропущено слово «равен»).

Пропуск местоимения

(Ш80, стр. 31) В окружности радиуса R проведены две хорды АВ и АС. На АВ или на ее продолжении взята точка М, ….Аналогично, на АС или *на продолжении* взята точка N. (Между словами «на» и «продолжении» пропущено местоимение «ее»).

Пропуск предиката

(Ш29, стр. 25) Сторона ВС треугольника АВС равна *a*, радиус вписанного круга *r*.

Для разрешения подобных эллипсисов требуется введение последовательно геометрических объектов по мере анализа предложения, правдоподобных рассуждений типа «если введен в рассмотрение единственный треугольник, то он является *данным* треугольником», рассуждений по аналогии: «если отрезок перпендикуляра заключен внутри треугольника, то вне треугольника возможно надо рассматривать другой отрезок перпендикуляра, лежащий на его продолжении». Полное разрешение эллипсиса равнозначно в этом случае построению **полной модели геометрической ситуации**, что будет включать знания о том, что перпендикуляр, восстановленный из середины гипотенузы, при продолжении пересекает сначала противолежащий катет (один из катетов), затем пересекает продолжение другого катета, что при пересечении линий образуются точки пересечения, между точками на линии образуются отрезки и т. п. Таким образом приходим к выводу, что на основе только синтаксического анализа и онтологического знания невозможно разрешение трудных случаев эллипсисов и полного понимания текста геометрических задач.

Анализ эллипсиса в задаче (Ш63 стр. 29) показывает, что разрешение эллипсиса может потребовать общезначимых знаний и применения других правдоподобных рассуждений, кроме рассуждения по аналогии, а именно рассуждений на основе импликаций (Modus Ponens). В первом предложении заданы две касающиеся друг друга окружности. Во втором предложении «из центра большей окружности проведен радиус»: окружностей две, одна из них большая,

следовательно, другая окружность – меньшая. Это рассуждение позволяет заключить, что проведенный радиус относится к меньшей окружности. Возможно разрешить этот эллипсис и на основе когнитивной модели: две окружности касаются друг друга внутренним образом, следовательно, одна окружность включает другую (знание из когнитивной модели внутреннего касания окружностей), и тогда одна окружность больше другой (импликативный вывод); мы имеем радиус OB, радиус это линия, она может касаться только окружности, следовательно, после слова «меньшей» должно стоять слово «окружности».

В задаче (Ш29 стр. 35) сложность в том, что необходимо понять, что *r* это не обозначение радиуса вписанного круга, а его величина (длина). Это действительно довольно сложный случай, который требует внелингвистических знаний, например, что обозначения (идентификаторы) это наборы заглавных букв, а величины обозначаются прописными буквами, причем каждая величина обозначается только одной буквой.

Анализ трудностей в разрешении эллипсисов показывает необходимость разработки нового подхода к анализу текстов планиметрических задач, основанного на когнитивных (мыслимых) моделях геометрических объектов, которые включают знания о каждом геометрическом объекте, о его составных элементах, действиях, его порождающих и преобразующих, действиях, которые объект может сам производить, его отношениях с другими объектами, о результатах геометрических построений и т. п. Такие модели также необходимо связывать с определенными языковыми конструкциями в данной проблемной области. Словарь терминов и выражений может создаваться на основе рассматриваемого корпуса текста.

Анализ текста становится когнитивно-управляемым, синтаксический анализатор играет подчиненную роль в этом процессе. Онтология содержит теоретическое знания в области планиметрии и знания методов решения планиметрических задач различного типа (вычислительных, на построение, на доказательство). Онтология берет на себя бремя решение задачи и формирования сопровождающего решение чертежа (визуализации решения).

Когнитивный анализатор текста работает (по замыслу) в пошаговом режиме и передает преобразованный и осмысленный текст в онтологию на языке, требуемом в онтологическом блоке. Схема когнитивно-управляемого анализа текста дана на Рис. 2.

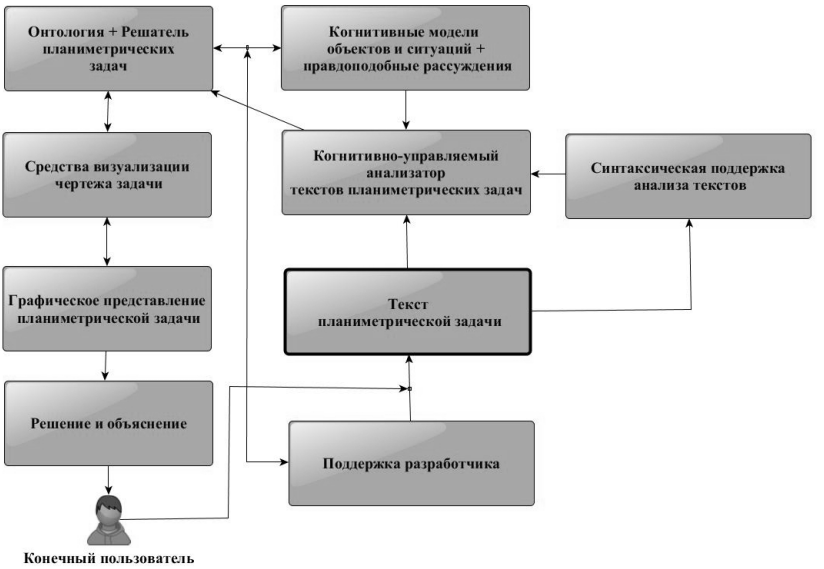


Рис. 2. Схема когнитивно-управляемого анализа текста

Взаимодействие когнитивных моделей и анализируемого текста должно обеспечить принцип «когнитивной ожидаемости» и «детерминированности контекста» [20].

Предлагаемый подход, на основе когнитивной семантики, согласуется с предложенным Сулеймановым Д.Ш. прагматически-ориентированным подходом к разработке лингвистических моделей и реализации систем обработки ЕЯ текстов [19].

2.3. Структура когнитивных моделей фигур и действий в планиметрии

Введем понятие когнитивной модели геометрического объекта. Когнитивная модель будет отображать, как объект образован и его положение в пространстве по отношению к другим объектам. Причем всем составляющим когнитивной модели объекта будут соответствовать слова, которые действительно появляются в рассматриваемом корпусе текстов. Иными словами, когнитивная модель определяет структуру текста.

В конструкции когнитивной модели важны следующие отношения:

• объект может совершать некоторые действия;

• объект может быть объектом других действий;

• объект состоит в пространственных и временных отношениях (раньше, позже, далее, уже построен, уже дан) с другими объектами;

• объект может состоять из других объектов;

• объект может быть частью других объектов;

• объект обладает свойствами, некоторые из которых (назовем их **актантными**) связаны с действиями, которые объект совершает (пересекает – пересекающий, лежит – лежащий) или с действиями, которые он испытывает (дан – данный, проведен – проведенный, образован – образованный, отсечен – отсеченный, вложен – вложенный);

отношения между свойствами одной фигуры и свойствами разных фигур; между свойствами фигуры и ее частей реализуются импликации: если *центры* – то *окружностей*, если *радиусы* – то *окружностей*, если *окружности* – то *описанные около* или *вписанные в*, если *вписанные в* – то «*в объект*», если биссектриса – то угол, если биссектриса и угол – то вершина, из которой исходит биссектриса, если биссектриса – то угол, из вершины которого она исходит, разделен пополам, если биссектриса – то она является осью симметрии угла, из вершины которого она исходит.

С когнитивными схемами мы связываем такое явление как **когнитивное ожидание** появления в тексте определенных слов и повествовательных конструкций.

Основные свойства объекта связаны и с его способом построения (треугольник прямоугольный, равнобедренный; равнобокая трапеция). Когнитивные модели строятся для каждого действия, совершаемого в планиметрии: ОБЪЕКТ – действие – ОБЪЕКТ = РЕЗУЛЬТАТ. Например, ОБЪЕКТ – пересечение с – ОБЪЕКТ = ТОЧКА(И) ПЕРЕСЕЧЕНИЯ; пересечь ОБЪЕКТ с ОБЪЕКТОМ = ТОЧКА(И) ПЕРЕСЕЧЕНИЯ.

Одним из инструментов для моделирования когнитивных геометрических конструкций может служить гипертекст. Гипертекстовая структура позволяет избежать рекурсии и зацикливания при анализе текста задачи. ИГ одного объекта может включать ИГ-ы других объектов, образованные через гипертекстовые ссылки. Например, текст «Биссектриса внутреннего угла треугольника АВС делит…» = (биссектриса (внутреннего угла (треугольника АВС))) делит… преобразуется в структуру именных групп: ИГ(биссектриса

ИГ(внутренний угол (ИГ(треугольник ABC)). Структура именных групп останется той же при изменении порядка слов в тексте. Текст «в треугольнике ABC биссектриса внутреннего угла делит…» также будет преобразован в структуру ИГ(биссектриса ИГ(внутренний угол (ИГ(треугольник ABC)).

По сути, ИГ это связанная (по падежам, числам, обозначениям) цепочка слов, которая описывает знания об объекте в рамках соответствующей когнитивной модели.

Другими инструментами для описания когнитивных моделей могут служить фреймы или семантические универсалии, описанные в [20] как структуры данных, которые служат для представления стереотипных ситуаций. «Фреймы это представления знаний о «мире», которые дают нам возможность совершать такие базовые когнитивные акты, как восприятия, понимание языковых сообщений и [языковые] действия» (цитируется по Умберто Эко, стр. 52). Фреймы (когнитивные модели) дают возможность производить абдуктивные рассуждения, то есть выдвигать гипотезы и искать их подтверждение в тексте.

Возможны и другие подходы к представлению когнитивных моделей, например, объектно-ориентированные базы знаний с ассоциативными связями между схемами объектов или когнитивное моделирование на основе референциальной концептуализации, предложенной в [21]. В дальнейшей работе предполагается проанализировать перечисленные подходы к реализации когнитивных моделей и разработать вариант реализации, наиболее адекватный в данной проблемной области.

В заключение опишем элементы когнитивной модели объекта «биссектриса» (Таблицы 1 и 2).

Таблица 1. **Глагольные группы с биссектрисой**

Биссектриса(ы)	Гиперссылка на действие	Гиперссылка на объект
Делящая	Делит (делить на)	Сторону треугольника
Перпендикулярная	Перпендикулярна (быть перпендикулярным)	Медиане треугольника
Рассекающая	Рассекает (рассекать)	Сторону параллелограмма на отрезки
Пересекающая	Пересекает (пересекать в)	Биссектрису(ы) треугольника в точке пересечения

Пересекающиеся	Пересекаются (пересекаться в, на)	В точке (пересечения) с окружностью (Окружность)
Ограничивающие	Ограничивают (ограничивать)	Площадь четырехугольника
Встречающая(щие)	Встречает(ют) (встречать)	Окружность в точке
Содержащая	Содержит (содержать)	Точку пересечения прямых
Лежащая	Лежит (лежать на, в)	На биссектрисе
Данная	Дана (быть данной)	Угла в треугольнике в параллелограмме
Исходящая	Исходит (исходить из)	Вершины угла
Образующие	Образуют (образовать при пересечении)	Квадрат

Таблица 2. **Именные группы биссектрисы**

Биссектриса(ы)	Гиперссылка на объект	Гиперссылка на объект
	Угла	
	Угла	Треугольника
	Угла (обозначение)	Треугольника (обозначение)
Острого	Угла	Прямоугольного треугольника
Внутреннего	Угла	Треугольника
	Угла	При основании равнобедренного треугольника
		В треугольнике
	Одной вершины	Вписанного треугольника
	Углов (прилежащих к одной стороне)	Параллелограмма
		Треугольника (обозначения)
Внутренних	4-х углов	В параллелограмме
	Углов (обозначения)	Выпуклого четырехугольника
	Углов	Прямоугольника

Таблицы составлены по текстам рассматриваемых геометрических задач.

Заключение

В статье дана классификация эллипсисов в задачах по планиметрии с присутствием в тексте тире (пропуск предиката, пропуск глагола), и без тире в тексте задачи (пропуск существительного или предиката).

Выявлена структура эллипсисов на основе выделения именной и глагольной групп.

Введено понятие когнитивных моделей геометрических объектов и действий с ними.

Показана существенная роль правдоподобных рассуждений при разрешении эллипсиса с пропущенным глаголом.

Продолжение работы предполагает разработку автоматизированного анализа текста геометрической задачи на основе когнитивно-управляемого синтаксического разбора.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект № 18-07-00098А).

ЛИТЕРАТУРА

1. Мальковский М.Г. [и др]. Восстановление эллипсиса как задача автоматической обработки текстов. Программные продукты и системы. – 2014. – № 3. – С. 32–36.

2. Кобзарева Т., Епифанов М., Лахути Д. Поиск антецедента эллипсиса в фрагменте со сказуемым (автоматический анализ русского предложения) // Труды 15ой Национальной конференции по искусственному интеллекту с международным участием (КИИ-2016). – Vol. 2. – 2016. – С. 56–62.

3. Адамец П. Семантическая интерпретация “значимых нулей” в русских предложениях // Язык и стих в России: сборник в честь Дина С. Ворта к его 65-летию. – М.: Восточная литература РАН, 1995. – С. 9–18.

4. Вардуль И. Ф. К вопросу о явлении эллипсиса // Инвариантные синтаксические значения и структура предложения. – М.: Наука, 1969. – С. 59–70.

5. Курбатов С.С., Найденова К.А., Хахалин Г.Г. Интегрирование интеллектуальных систем анализа/синтеза изображений и текста: контуры проекта INTEGRO // Труды международной научной конференции «Открытые семантические технологии для интеллектуальных систем (ОСТИС-2011). – Минск: БГТУ, 2011. – С. 213–232.

6. Курбатов С., Воробьев А. Онтологический решатель геометрических задач по естественно-языковому описанию // Труды пятнадцатой национальной конференции по искусственному интеллекту с международным участием (КИИ-2016). – Vol. 1. – 2016. – С. 56–63.

7. Ларькина А. Эллипсис в современном французском языке: автореферат диссертации на соискание степени кандидата филологических наук (специальность 10.02.05). – СПБ: СПБ ГУ, 2009. – 175 с.

8. Апресян Ю., Богуславский И., Иомдин Л. [и др.] Лингвистическое обеспечение системы ЭТАП-2. – М.: Наука, 1989. – 295 с.

9. Лобзин А., Хахалин Г., Курбатов С., Литвинович А. Интеграция на базе онтологии естественного языка и изображений в системе TEXT-TO-PICTURE // Труды 8-ой Международной научно-практической конференции «Интегрированные модели и мягкие вычисления в искусственном интеллекте». – М.: Физматлит, 2015. – С. 296–305.

10. McShane, M. & Babkin, P. Automatic Ellipsis Resolution: Recovering Covert Information from Text // Proceedings of the Twenty Ninth AAAI Conference on AI. – 2015. – P. 572–578.

11. McShane, M., Nirenburg, S., Beale, S., Johnson, B. Resolving Elided Scopes of Modality in OntoAgent // Advances in Cognitive Systems. – Vol. 2. – 2012. – P. 95–112.

12. Kenyen-Dean, K., Cheung, J.C.K., Precup, D. Verb Phrase Resolution Using Discriminative and Margin-Infused Algorithms // Empirical Methods in NLP: Proc. of the 2016 Conf. – 2016. – P. 1734–1743.

13. McShane, M., Babkin, P. Detection and Resolution of Verb Phrase Ellipsis // LiLT. – Vol. 13. – №1. – 2016. – P. 1–36.

14. Liu, Z., Gonzalez, E., Gillick., D. Exploring the steps of VPE //Proc. of the Workshop on Conference Resolution Beyond OntoNotes (CORBON 2016), co-located with NAACL. – 2016. – P. 32–63.

15. Schuster, S., Nivre, J., Manning, Ch. D. Sentences with Gapping: Parsing and Reconstructing Elided Predicates. arXive: 1801.06922v1 [cs. CL] 18 Apr 2018.

16. Эко У. Роль читателя. Исследование по семиотике текста. – М.: АСТ, 2016. – 638 с.

17. Невзорова О.А., Невзоров В.Н. Интеллектуальная инструментальная система OntoIntegrator для задач автоматической обработки текстов // Тринадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2012): Труды конференции. – Т. 4. – Белгород: Изд-во БГТУ, 2012. – С. 92–99.

18. Розенталь Д.Э. Справочник по русскому языку: орфография и пунктуация. – М.: АСТ, 2013. – 736 с.

19. Сулейманов Д.Ш. Прагматически-ориентированные лингвистические модели как основа систем и технологий обработки естественного языка. Аналитический обзор // Формальные модели и системы в вычислительной лингвистике / под ред. Соснина П.И. и Невзоровой О.А. – Казань: АН РТ, 2016. – Глава 1. – С. 8–59.

20. Сулейманов Д.Ш., Гатиатуллин А.Р. Система семантических универсалий и их реализация в виде реляционно-ситуационного фрейма // Когнитивно-семиотические аспекты моделирования в гуманитарной сфере /под ред. В. Стефанюка и Э. Тайсиной. – Казань: АН РТ, 2017. – С. 185–210.

21. Theodorakis, M., Analyti, A., Constantopoulos, P., and Spyratos, N. Contextualization as an Abstraction Mechanism for Conceptual Modeling // Conceptual Modeling: Proceeding of 18th Int. Conf. (ER'99). – Springer, 1999. – P. 476–489.

УДК 004.822

STUDY OF ONTOMATHEDU ONTOLOGY STRUCTURE: FIRST RESULTS

O. A. Nevzorova[1], V. N. Nevzorov[2], L. R. Shakirova[3], M. V. Falileeva[3]
[1]Kazan Federal University, Tatarstan Academy of Sciences, Kazan
[2]Kazan National Research Technical University named after A.N. Tupolev, Kazan
[3]Kazan Federal University, Kazan
onevzoro@gmail.com, nevzorovvn@gmail.com, liliana008@mail.ru, mmwwff@yandex.ru

This article presents the results of the linguistic-statistical analysis of the ontological resource OntoMathEdu, which is built in the field of elementary geometry. We used the analytical software of the "OntoIntegrator" ontological-linguistic system. The ontology OntoMathEdu contains both the taxonomy of the main objects of planimetry and the taxonomy of the main typical tasks studied in the school course of geometry. Frequency analysis of educational texts allows us to choose the most important concepts of ontology, which can later be used in the ranking of ontological concepts in the process of studying geometry.

Keywords: ontology, concept, elementary geometry, "OntoIntegrator" ontological-linguistic system, linguistic-statistical analysis.

ИССЛЕДОВАНИЕ СТРУКТУРЫ ОНТОЛОГИИ ONTOMATHEDU: ПЕРВЫЕ РЕЗУЛЬТАТЫ

О. А. Невзорова[1], В. Н. Невзоров[2], Л. Р. Шакирова[3], М. В. Фалилеева[3]
[1]Казанский федеральный университет, Академия наук Республики Татарстан, Казань
[2]Казанский национальный исследовательский технический университет им. А.Н. Туполева, Казань
[3]Казанский федеральный университет, Казань
onevzoro@gmail.com, nevzorovvn@gmail.com, liliana008@mail.ru, mmwwff@yandex.ru

В статье проведен лингво-статистический анализ нового онтологического ресурса по элементарной геометрии OntoMathEdu, выполненный с использованием аналитических инструментов онтолого-лингвистической системы «OntoIntegrator». Онтология OntoMathEdu содержит как систематизацию основных объектов планиметрии, так и основные типовые зада-

чи, изучаемые в школьном курсе геометрии. Частотный анализ учебных текстов позволил выявить наиболее важные понятия онтологии, что впоследствии может быть использовано при ранжировании онтологических понятий в процессе изучения геометрии.

Ключевые слова: онтология, концепт, элементарная геометрия, онтолого-лингвистическая система «OntoIntegrator», лингво-статистический анализ.

Введение

«Умение самостоятельно построить целостную картину дисциплины» является необходимой компетенцией, принятой в современном Российском федеральном государственном образовательном стандарте высшего профессионального образования по направлению подготовки «Математика» для магистров [1]. В связи с этим формирование системного мышления в средней и высшей школе является приоритетной целью обучения. Для реализации этой цели необходимо выстроить системную картину математических знаний, в которой были бы отражены взаимосвязи различных составных частей, образующих целостную картину. В [2] отмечается, что в массовом педагогическом опыте формирование системного мышления еще не стало объектом теоретико-методологического сознания и практической реализации. Одной из причин является отсутствие разработанных моделей системной организации знаний, применяемых в педагогической практике.

Одним из возможных подходов к систематизации знаний является онтологический подход. В самом простом случае онтология описывает иерархию связанных понятий. В более сложных случаях в описание добавляются аксиомы, выражающие связи между понятиями и ограничивающие их интерпретацию. Выделяются различные виды онтологий, такие как:

- онтологии верхнего уровня, описывающие наиболее абстрактные концептуализации;
- предметные онтологии, описывающие понятийный состав конкретных предметных областей;
- онтологии задач, описывающие задачи (деятельность) в конкретных предметных областях;
- онтологии-приложения, ориентированные на применение знаний предметных онтологий и онтологий задач для специфических приложений.

Для образовательных целей, прежде всего, требуется разработка предметных онтологий по учебным предметам, в которых системно выделены основные понятия учебной дисциплины и взаимные связи понятий. С другой стороны, развитие инициативы Open Linked Education "Открытые связанные данные в образовании" (https://www.w3.org/community/opened/) также требует дополнительной онтологизации предметных областей, поскольку одной из главных текущих целей данной инициативы в настоящее время является сбор данных, связанных с образованием, в том числе разработка специализированных терминологических словарей и приложений, которые используют терминологические ресурсы.

Для моделирования предметных областей в области математики ранее была разработана онтология профессиональной математики OntoMathPRO [3-6]. Эта онтология покрывает ряд областей математики, таких как алгебра, математический анализ, геометрия, дифференциальные уравнения, численный анализ, теория вероятностей и математическая статистика. Новый онтологический ресурс – онтология OntoMathEdu по планиметрии ориентирована на математические знания курса геометрии средней школы и может быть эффективно использована в преподавании школьной геометрии. В настоящей статье анализируются структурные характеристики нового ресурса и определяются направления его дальнейшего развития.

Онтология OntoMathEdu

Онтология OntoMathEdu в текущей версии разработана для систематизации знаний в области элементарной геометрии. Элементарная геометрия является уникальным разделом математики, который имеет интересную и богатую историю развития, большие области применения в жизни.

В настоящее время подготовка учащихся по геометрии является наиболее проблемной областью математического образования. Несмотря на аксиоматический подход в построении школьного курса планиметрии, анализ педагогической практики показал неполноту, отсутствие системности в построении взаимосвязей между геометрическими понятиями. Одним из путей решения данной проблемы является конструирование содержания курса планиметрии, построенного на основе системного подхода с опорой на онтологию с концептами, связанными системой ключевых отношений, необходимых для построения полной картины геометрического знания. Та-

кая онтология, по нашему мнению, позволит методистам, учителям выработать более продуктивные методики обучения геометрии.

Проектирование онтологии OntoMathEdu включало на данном этапе построение формальной таксономии геометрических понятий. Геометрические понятия курса планиметрии (7–9 класс средней школы) извлекались вручную из учебников по геометрии для общеобразовательных учреждений. Использовались учебники Атанасяна Л.С, Бутузова В.Ф. и др. [7], Погорелова А.В. [8], Смирновой И.М и Смирнова В.А. [9], Шарыгина И.Ф. [10] и учебно-методические материалы. Выделено 585 концептов, организованных в 14 иерархий, установлено 583 связи типа «класс-подкласс».

Структурные свойства нового онтологического ресурса были проанализированы с помощью аналитических программных инструментов системы «OntoIntegrator». Ниже будут представлены некоторые предварительные оценки построенного ресурса.

Аналитические программные инструменты системы «OntoIntegrator»

Система «OntoIntegrator» является онтолого-лингвистической научно-исследовательской инструментальной средой, ориентированной на задачи автоматической обработки текстов с использованием различных онтологических моделей. Она относится к классу специализированных систем, основные функциональные возможности которых обеспечивают:

- проектирование онтологических моделей произвольной структуры с широкими возможностями визуализации данных;
- моделирование прикладных задач, связанных с обработкой текстов;
- обработку текстов на основе онтологических и лингвистических моделей.

Система ориентирована на разработку приложений, связанных с эффективной автоматизацией прикладных задач анализа текстов, и реализует онтолингвистический подход [11].

Спектр применений разрабатываемых технологий весьма разнообразен и включает: структурирование и визуализацию текстовых документов, автоматическое аннотирование документов, текстовый поиск, классификацию документов, извлечение знаний из текстов.

Аналитические инструменты позволяют исследовать различные структурные свойства онтологий. При использовании этих инстру-

ментов для анализа онтологии OntoMathEdu были получены количественные и качественные результаты, позволившие выявить некоторые структурные недостатки, а также определить конкретные шаги по внесению дополнений и изменений в онтологию.

Структурные свойства онтологии OntoMathEdu

Всего в онтологии OntoMathEdu с помощью программных инструментов системы «OntoIntegrator» было выделено:

- 585 концептов;
- 14 иерархий;
- 616 текстовых входов;
- 583 связей типа «класс-подкласс»;
- 3 изолированные вершины;
- 6 многозначных концептов (принадлежащих разным иерархиям).

В онтологии OntoMathEdu определены следующие иерархии, описывающие основные тематические разделы и ключевые концепты элементарной геометрии:

1. Взаимное расположение геометрических фигур на плоскости.
2. Геометрическая фигура на плоскости.
3. Единица измерения.
4. Инструменты измерений и построений.
5. Конструктивные аксиомы и задачи на построение.
6. Метрическое свойство геометрической фигуры.
7. Основные понятия аксиоматического построения планиметрии.
8. Основные элементы геометрического преобразования.
9. Отношения между геометрическими фигурами.
10. Преобразование плоскости.
11. Признак или свойство геометрического преобразования.
12. Расстояние между геометрическими фигурами.
13. Средние величины в планиметрии.
14. Теоремы планиметрии.

Каждая из указанных выше иерархий включает понятия, связанные отношением «класс-подкласс». Число уровней в иерархии различно и представлено на рис. 1. Наиболее представительными являются базовые иерархии 2 (Геометрическая фигура на плоскости, 156 понятий) и 14 (Теоремы планиметрии, 152 понятия).

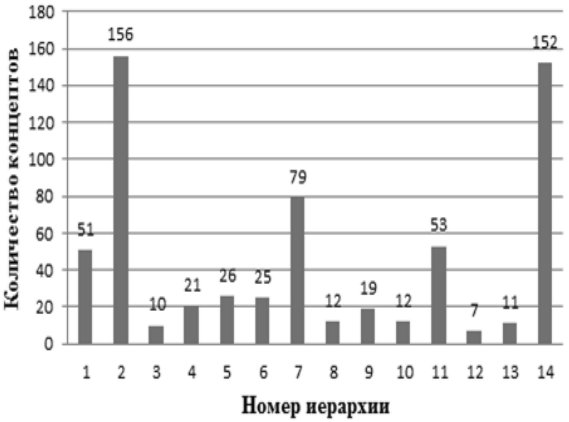


Рис. 1. Состав иерархий онтологии OntoMathEdu

Фрагмент иерархии «Геометрическая фигура на плоскости» приведен на рис. 2.

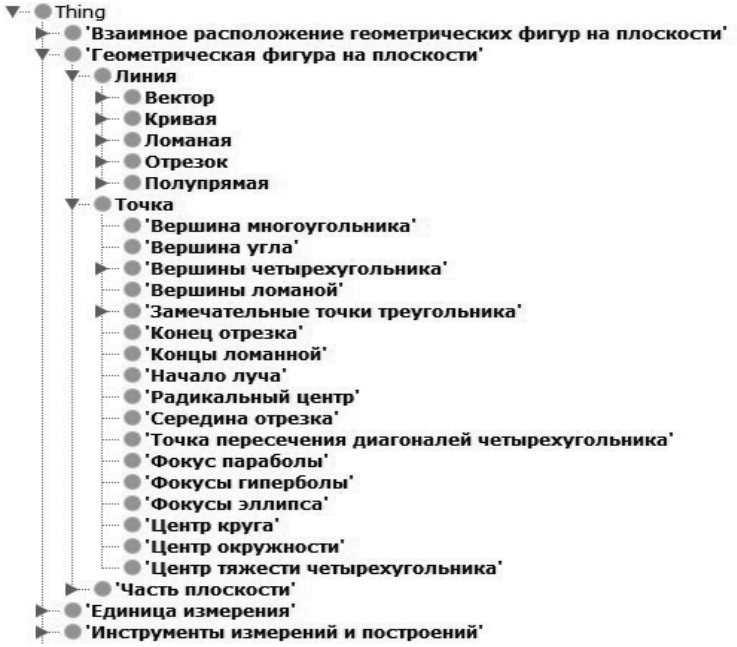


Рис. 2. Фрагмент иерархии «Геометрическая фигура на плоскости»

Рис. 3. Распределение имен концептов по составу

Иерархия «Геометрическая фигура на плоскости» выделяет первый подуровень, относящийся к основным классам геометрических фигур на плоскости (Линия, Точка, Часть плоскости), и затем детальную классификацию понятий по отношению «класс-подкласс» внутри соответствующих верхних классов. Несомненно, что построенные иерархии обладают значительным потенциалом в применении их в педагогической практике.

Как уже отмечалось, онтология OntoMathEdu строилась вручную на материалах школьных учебников и поэтому следующей задачей явилось применение автоматических методов извлечения терминологии, разработанных в системе «OntoIntegrator». Для извлечения математической терминологии из учебных текстов применялись различные алгоритмы обработки содержимого математических текстов: алгоритмы сегментации текста на предложения; методы распознавания объектов текста (выделение формул, числовых последовательностей, слов, знаков препинания, аббревиатур и др.); алгоритмы распознавания именных групп, содержащих термины прикладной онтологии; алгоритмы распознавания сложных синтаксических конструкций (групп сочинительного сокращения) и др. [12]. Дополнительно был выполнен лингво-статистический анализ имен онтологических понятий. На рис. 3 показано частотное распределение слов в названиях концептов онтологии. Наиболее частотными классами являются двух-, трехсловные именования, относящиеся к именам основных объектов предметной области. Более длинные названия

(более 5 слов) относятся фактически к формулировкам стандартных задач в данной предметной области. Таким образом, особенностью онтологии OntoMathEdu является не только систематизация объектов элементарной геометрии, но и систематизация типовых задач, что важно для ее применения в образовательном процессе. Примеры имен концептов приведены на рис. 4.

Кол-во слов в названии	Пример названия концепта
1	Вектор; Гипербола; Точка
2	Вписанный угол; Гомотетия луча
3	Внешняя область многоугольника Гармоническая четверка точек
4	Внутренняя касательная двух окружностей Вычисление расстояния между точками
5	Деление отрезка точкой внешним образом Косинус острого угла прямоугольного треугольника
6	Задачи на построение циркулем и линейкой Вычисление длины вектора по его координатам
7	Нахождение координат суммы двух и более векторов; Свойство точек касания вписанной и вневписанных окружностей
8	Второй распределительный закон умножения ненулевого вектора на число; Признак равенства прямоугольных треугольников по гипотенузе и катету
9	Нахождение площади многоугольника с вершинами в точках целочисленной решетки; Построение треугольника по двум сторонам и углу между ними
10	Построение треугольника по стороне и двум прилежащим к ней углам; Признак параллельных прямых по равенству накрест лежащих углов при секущей
11	Теорема о вписанных углах, опирающихся на одну и ту же дугу; Нахождение площади параллелограмма по стороне и высоте, проведенной к этой стороне
12	Правила, позволяющие по координатам векторов находить координаты их суммы, разности и произведения

Рис. 4. Примеры именований концептов в онтологии

При построении онтологии для образовательных практик было бы полезным использовать данные о степени значимости понятий в

учебном курсе. Данные о частотности понятий в учебных материалах, взаимосвязях высокочастотных понятий (контекстное окружение высокочастотных понятий) способствует выделению наиболее важных понятий учебной дисциплины. Последующее ранжирование понятий с точки зрения их значимости может быть полезным при формировании материалов для тестирования. На рис. 5 приведены высокочастотные понятия (с указанием частоты встречаемости) по двум школьным учебникам геометрии. На рис.6 даны низкочастотные понятия.

Самые высокочастотные концепты по учебнику Шарыгина		Самые высокочастотные концепты по учебнику Атанасяна	
Название	**Кол-во**	**Название**	**Кол-во**
точка	1566	точка	1620
треугольник	842	прямая	1017
прямая	835	угол	900
окружность	755	треугольник	878
угол	641	отрезок	587
теорема	345	вектор	493

Рис. 5. Высокочастотные понятия онтологии OntoMathEdu

Самые низкочастотные концепты по учебнику Шарыгина		Самые низкочастотные концепты по учебнику Атанасяна	
Название	**Кол-во**	**Название**	**Кол-во**
эллипс	1	семиугольник	1
сантиметр	1	ломаная	1
радиан	2	штангенциркуль	2
полупрямая	2	рулетка	2
свойство треугольника	2	рейсшина	2
признаки ромба	2	признаки параллелограмма	2

Рис. 6. Низкочастотные понятия онтологии OntoMathEdu

Общая оценка частотного распределения концептов онтологии приведена на рис.7.

Частота употребления	Кол-во концептов по учебнику Шарыгина	Кол-во концептов по учебнику Атанасяна
1000-1620	1	2
500-999	4	3
100-499	13	19
50-99	21	16
10-49	69	79
5-9	31	27
1-4	54	47
Всего терминов:	190	193

Рис. 7. Частотное распределение концептов онтологии OntoMathEdu

Заключение

В статье проведен лингво-статистический анализ нового онтологического ресурса по элементарной геометрии OntoMathEdu, выполненный с использованием аналитических инструментов системы «OntoIntegrator». Онтология OntoMathEdu содержит не только систематизацию основных объектов предметной области, но и включает в себя основные типовые задачи, изучаемые в школьном курсе геометрии. Последнее обстоятельство делает этот ресурс особенно полезным для применения в образовательном процессе. Частотный анализ учебных текстов позволил выявить наиболее важные понятия онтологии, что впоследствии может быть использовано при ранжировании онтологических понятий в процессе изучения геометрии.

Проведенный анализ является предварительным и позволяет определить направления для дальнейшего развития ресурса и его совершенствования.

Благодарности. Работа выполнена при финансовой поддержке РФФИ и Правительства Республики Татарстан в рамках научного проекта № 18-47-160007.

ЛИТЕРАТУРА

1. Федеральный государственный образовательный стандарт высшего профессионального образования по направлению подготовки 010100 математика (квалификация (степень) «магистр»). http://fgosvo.ru/uploadfiles/fgos/30/20110325143133.pdf.

2. Сагатеева Л.С. В Формирование системного мышления в обучении математики // Известия ВГПУ, 2008. № 9. С. 201–204. https://cyberleninka.ru/article/n/formirovanie-sistemnogo-myshleniya-v-obuchenii-matematike.

3. Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E. OntoMathPRO Ontology: A Linked Data Hub for Mathematics // Pavel Klinov, Dmitry Mouromstev (eds.). Proceedings of the 5th International Conference on Knowledge Engineering and Semantic Web (KESW 2014). Communications in Computer and Information Science, vol. 468. Springer, Cham, 2014. pp. 105–119. doi: https://doi.org/10.1007/978-3-319-11716-4_9.

4. Elizarov A., Kirillovich A., Lipachev E., Nevzorova O. Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management // Communications in Computer and Information Science, vol 706. Springer, Cham, 2017. pp. 33–46.

5. Елизаров А.М., Липачев Е.К., Невзорова О.А., Соловьев В.Д. Методы и средства семантического структурирования электронных математических документов // Доклады Академии Наук, 2014. Т. 457. № 6. С. 642–645.

6. Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачев Е.К. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов // Доклады РАН, 2016. Т. 467. № 4. С. 392–395.

7. Геометрия. 7–9 классы: учеб. для общеобразоват. организаций / [Л.С. Атанасян, В.Ф. Бутузов, С.Б. Кадомцев и др.]. – М.: Просвещение, 2018. – 383 с.

8. Погорелов А.В. Геометрия. 7–9 классы: учеб. для общеобразоват. организаций / А.В. Погорелов – М.: Просвещение, 2018. – 240 с.

9. Смирнова И.М. Геометрия. 7–9 классы: учеб. для общеобразоват. организаций / И.М. Смирнова, В.А. Смирнов. – М.: Мнемозина, 2015. – 376 с.

10. Шарыгин И.Ф. Геометрия. 7–9 классы: учеб. для общеобразоват. организаций / И.Ф. Шарыгин. – М.: Дрофа, 2018. – 364 с.

11. Невзорова О.А. Онтолингвистические системы: технологии взаимодействия с прикладной онтологией // Ученые записки Казанского государственного университета. Серия физико-математические науки. Т. 149. Кн. 2. 2007. С. 105–115.

12. Невзорова О.А., Невзоров В.Н. Методы аннотирования структурных элементов математического текста на основе информационных технологий системы «OntoIntegrator» // Информационные и математические технологии в науке и управлении / Труды XVI Байкальской Всероссийской конференции «Информационные и математические технологии в науке и управлении». Часть II. – Иркутск: ИСЭМ СО РАН, 2011. – С. 106–112.

УДК 004.021

KEYWORD BASED ALGORITHM OF NAMED ENTITY RECOGNITION IN THE TATAR LANGUAGE CORPUS

O. A. Nevzorova, D. R. Mukhamedshin, A. M. Galieva
The Tatarstan Academy of Sciences, Kazan
onevzoro@gmail.com, onevzoro@gmail.com, amgalieva@gmail.com

Named entities recognition is one of the urgent tasks in the researches of language using electronic language corpuses. This article discusses the main methods for solving this problem, including algorithms based on various machine learning models, regular expressions and dictionaries. Also in the article, the authors proposed their own algorithm, which allows named entities recognition on the basis of search queries using direct and reverse search. The results of the algorithm, presented in the article, suggest what additional functions are necessary to achieve the best results. The proposed algorithm is used in the "Tugan Tel" corpus management system and can be used both with the electronic corpus of the Tatar language and with corpuses of other languages.

Keywords: Named entity recognition, NER, Corpus management system, Text mining.

РАСПОЗНАВАНИЕ ИМЕНОВАННЫХ СУЩНОСТЕЙ В КОРПУСЕ ТАТАРСКОГО ЯЗЫКА НА ОСНОВЕ КЛЮЧЕВОГО СЛОВА ЗАПРОСА

О. А. Невзорова, Д. Р. Мухамедшин, А. М. Галиева
Академия Наук Республики Татарстан, Казань
onevzoro@gmail.com, damirmuh@gmail.com, amgalieva@gmail.com

Распознавание именованных сущностей – одна из актуальных задач в исследованиях языка с использованием электронных языковых корпусов. В этой статье рассматриваются основные методы для решения этой задачи, включая алгоритмы, основанные на различных моделях машинного обучения, регулярных выражениях и словарях. Также в этой статье авторы представляют свой собственный алгоритм, который позволяет извлекать именованные сущности на основе поисковых запросов с использованием прямого и обратного поиска. Результаты работы алгоритма, представленные в статье, позволяют предположить, какие дополнительные функции необходимо добавить для достижения наилучших результатов. Представленный алгоритм используется в системе управления корпусом «Туган

Тел» и может быть использован как с электронным корпусом татарского языка, так и с электронными корпусами других языков.

Ключевые слова: Распознавание именованных сущностей, NER, система управления корпусом, обработка текста.

Электронные языковые корпусы являются основой для обширных исследований, связанный с изучением языка. Системы управления корпусами помогают решать ряд лингвистических задач, такие как прямой поиск по словоформам и леммам, обратный поиск по морфологическим свойствам, выборка контекстов и n-грамм на основе различных поисковых запросов. Такие простые запросы поддерживаются большинством систем управления корпусами.

Одной из сложных задач поиска в корпусных данных является распознавание именованных сущностей. Эта задача решается десятками исследователей, зачастую с хорошими результатами. Большинство решений, часть из которых описана в Разделе 2 этой статьи, работает с английским, испанским, датским, немецким языками, используя в качестве основы различные методы машинного обучения, регулярные выражения, словари и т. д. В Разделе 4 этой статьи авторами представлен один из возможных алгоритмов для распознавания именованных сущностей, который может быть использован как совместно с электронным корпусом татарского языка, так и с электронными корпусами других языков. Этот алгоритм используется в одном из модулей системы управления корпусом «Туган Тел». Авторы также провели ряд экспериментов, результаты которых представлены в Разделе 4.2 этой статьи.

Система управления корпусом «Туган Тел»

Система управления татарским корпусом (www.corpus.antat.ru) разработана в Институте прикладной семиотики Академии Наук Республики Татарстан. Основной функционал системы управления корпусом включает поиск лексических единиц, морфологический и лексический поиск, поиск синтаксических единиц, поиск n-грамм на основе грамматики и т. д. Основой системы является семантическая модель представления данных. Поиск производится при помощи свободно распространяемого ПО: используется система управления базами данных MariaDB и хранилище данных Redis [1]. Целью является разработка системы управления корпусом с поддержкой электронных корпусов тюркских языков.

Среди известных проектов электронных корпусов для тюркских языков корпусы турецкого и уйгурского языков [2], башкирского, хакасского, казахского (http://til.gov.kz), тувинского языков. Татарский национальный корпус «Туган Тел» – это лингвистический ресурс современного литературного татарского языка. Он содержит более 100 миллионов словоформ по состоянию на ноябрь 2016 года. Корпус содержит тексты различных жанров: художественные тексты, тексты СМИ, официальные документы, учебники, научные статьи и т. д. У каждого документа есть мета-описание [3]: автор, название, информация о публикации, дата создания, жанр и т.д. Тексты, содержащиеся в корпусе, имеют морфологическую разметку, т. е. информацию о части речи и грамматических свойствах словоформы [4]. Морфологическая разметка производится автоматически на основе модуля двухуровневого морфологического анализа татарского языка при помощи приложения PC-KIMMO.

Связанные работы

Системы, основанные на знаниях

Системы NER, основанные на знаниях, используют лексические ресурсы и знания предметной области без необходимости в обучении с аннотированными данными. Такие системы показывают хорошие результаты, когда лексические ресурсы являются полными, но они не работают, например, с примерами из класса drug_n в наборе данных DrugNER [5], так как они не определены в словарях DrugBank. Несмотря на их высокую точность, такие системы показывают низкую отзывчивость из-за специфичных правил языка, предметной области или неполных словарей. Другим недостатком систем NER, основанных на знаниях, является необходимость участия экспертов в разработке и поддержке ресурсов знаний.

Системы с обучением с учителем

Модели машинного обучения с учителем обучают делать прогнозы путем обучения на примерах входов и их ожидаемых выходов и могут быть использованы для замены различных правил. Скрытые модели Маркова (Hidden Markov Models, HMM), Машины опорных векторов (Support Vector Machines, SVM), Условные случайные поля (Conditional Random Fields, CRF) и деревья решений являются распространенными моделями машинного обучения для распознавания именованных сущностей.

Результаты исследований различных моделей машинного обучения от различных авторов представлены в Таблице 1.

Таблица 1. **Исследования различных моделей машинного обучения**

Автор(ы)	Модель машинного обучения	Дополнения	Результаты
Чжоу и Су (2002) [6]	HMM	Включены 11 орфографических свойств, список слов-триггеров для именованных сущностей и список слов из различных справочников.	F-score 96,6% и 94,1% на наборах данных MUC-6 и MUC-7, соответственно.
Андо и Чжан (2005) [7]	Структурное обучение [7]	Выбирался лучший классификатор для каждой вспомогательной задачи, основываясь на его доверии.	F-score 89,31% и 75,27% для английского и немецкого языков, соответственно.
Агерри и Ригау (2016) [8]	Semi-supervised system	Включены орфография, символьные n-граммы, лексиконы, префиксы, суффиксы, биграммы, триграммы, кластеризация из корпусов Брауна и Кларка, кластеризация методом k-средних открытого текста с использованием включения слов.	F-score 84,16%, 85,04%, 91,36%, 76,42% на испанском, датском, английском и немецком наборе данных CoNLL, соответственно.

Извлечение именованных сущностей

Извлечение именованных сущностей из корпусных данных, с одной стороны, позволяет получить необходимые данные по запросу, с другой стороны, проверить корпус на наличие определенной информации и дополнить его документами, содержащими недостающие данные. Алгоритм извлечения именованных сущностей, предложенный в данной статье, позволяет выделить семантические единицы в корпусе, не прибегая к семантической разметке. К тому же, алгоритм не привязан к семантическим типам извлекаемых данных, так как семантический тип определяется ключевым словом в запросе.

Описание алгоритма именованных сущностей

Алгоритм для извлечения именованных сущностей основан на идее сравнения n-грамм. Сравнение производится по всему корпусу, что повышает точность результатов.

Процесс извлечения является итеративным, предельное количество итераций определяется пользователем. На первом шаге производится выборка по начальному поисковому запросу. Начальный поисковый запрос может быть по словоформе, лемме, фразе или по морфологическим свойствам. Список биграмм и их частотность собираются по всей выборке. Биграммы, содержащие результаты, расширяются на одну позицию влево или вправо (задается пользователем). Полученный список сортируется по частотности биграмм в порядке убывания, выполняется срез списка по предопределенному индексу покрытия (например, 95% от всех результатов, этот параметр задается пользователем). Этот результат используется во второй итерации алгоритма, на которой производится фразовый поиск в корпусе по каждой биграмме. Результаты поиска участвуют в формировании списка триграмм, которые расширены на одну позицию влево или вправо, и их частотности. Полученный список триграмм также сортируется по частотности в порядке убывания и выполняется срез по индексу покрытия.

В третьей и последующих итерациях (пока не будет достигнуто предельное количество итераций или ничего не будет найдено на предыдущей итерации) используются списки n-грамм, полученные на предыдущей итерации. Производится поиск n-грамм в режиме фразового поиска по корпусу, чтобы получить список (n + 1)-грамм. Выполняются срез полученного списка по индексу покрытия и сравнение со списком n-грамм, полученном на предыдущей итерации. Точность сравнения P задается пользователем в процентах. Если частотность n-граммы меньше, чем P от количества найденных (n + 1)-грамм, то n-грамма считается найденной именованной сущностью, иначе извлечение продолжается. Таким образом, окончательные результаты будут представлять собой список наиболее устойчивых n-грамм различной длины, содержащих результаты поиска по начальному запросу.

Запрос на извлечение именованных сущностей является расширением кортежа Q и представлен в (1). В дополнение к поисковому запросу, добавлены компоненты, определяющие предельное коли-

чество итераций влево (L) и вправо (R), индекс покрытия (C), и точность сравнения (P).

$$Q = (Q_1, Q_2, L, R, C, P) \quad (1)$$

Эксперименты

Извлечение именованных сущностей при помощи алгоритма, предложенного авторами, требует наличия начального поискового запроса, который должен содержать индикатор определенной именованной сущности. Этот индикатор позволяет классифицировать именованные сущности, поэтому авторы выбрали набор классов schema.org в качестве основы для определения индикаторов. Из этого набора классов авторы выбрали следующие классы для поиска именованных сущностей в корпусе татарского языка: книги, рестораны, фильмы, журналы, компании, аэропорты, корпорации, языки, техникумы, университеты, школы, магазины, музеи, больницы. Министерства и названия улиц также были добавлены в этот список. Ниже представлены некоторые результаты экспериментов, произведенных авторами.

Названия министерств

В рамках задачи поиска именованных сущностей был проведен ряд экспериментов. Одним из наиболее показательных из них стал поиск названий министерств. Начальный поисковый запрос для этого эксперимента представлен в (2).

Q = ((wordform, *ministrlygy,* “”, right, 1, 10, exact), 7, 0, 95, 80) (2)

Результатом этого запроса является список из 50 n-грамм, содержащих словоформу «*ministrlygy*» в последней позиции. Справочный список названий министерств, представленный на вебсайте правительства Республики Татарстан [http://prav.tatarstan.ru/tat/ministries.htm], содержит 17 элементов. 12 из 17 элементов были найдены в корпусе при помощи алгоритма, таким образом, покрытие результатов составляет 70.6%. 5 элементов не было найдено в корпусе по различным причинам, представленным в Таблице 2. Оставшиеся 33 n-грамм – это различные разговорные варианты названий министерств.

Таблица 2. **Список ненайденных названий министерств**

Название	Причина
Urman huҗalygy ministrlygy (тат.) – Министерство лесного хозяйства	Пересечение последовательности словоформ «huҗalygy ministrlygy» (тат.) – министерство хозяйства с последовательностью в другом названии «Transport həm yul huҗalygy ministrlygy» (тат.) – Министерство транспорта и дорожного хозяйства
Yashlər eshləre həm sport ministrlygy (тат.) – Министерство молодежи и спорта	Значения, встречающиеся в корпусе, не соответствуют официальному названию
Transport həm yul huҗalygy ministrlygy (тат.) – Министерство транспорта и дорожного хозяйства	Пересечение последовательности словоформ «huҗalygy ministrlygy» (тат.) – министерство хозяйства с последовательностью в другом названии «Urman huҗalygy ministrlygy» (тат.) – Министерство лесного хозяйства
Hezmət, halykny el belən təemin itү həm social yaklau ministrlygy (тат.) – Министерство труда, занятости и социальной защиты	Значения, встречающиеся в корпусе, не соответствуют официальному названию
Ecologia həm tabigy baylyklar ministrlygy (тат.) – Министерство экологии и природных ресурсов	Значения, встречающиеся в корпусе, не соответствуют официальному названию

Названия улиц

Другой эксперимент был связан с поиском названий улиц. Поисковый запрос для этого эксперимента представлен в (3).

Q = ((wordform, *uramy*, “”, right, 1, 10, exact), 7, 0, 95, 80) (3)

Результатом этого запроса является список из 600 n-грамм, содержащих словоформу «*uramy*» в последней позиции. После ручной проверки данных были получены следующие результаты: 432 (72%) n-грамм являются названиями улиц, 72 (12%) n-грамм также являются названиями улиц, но требуют фильтрации специальных символов, 96 (16%) n-грамм не являются названиями улиц по различным причинам (например, все предложения, содержащие слово «*uramy*»; почтовые адреса и др.).

Названия ресторанов

Другой эксперимент связан с поиском названий ресторанов. Поисковый запрос этого эксперимента представлен в (4).

Q = ((wordform, *restoran*, “POSS_3SG,SG”, right, 1, 10, exact), 7, 0, 95, 80) (4)

Результатом выполнения этого запроса стал список из 285 n-грамм, содержащих лемму “*restoran*” с морфологическими свойствами POSS_3SG и SG в последней позиции, которые в совокупности встречаются в корпусе 359 раз. В данном случае кроме названий ресторанов были получены названия подклассов ресторанов по их географическому положению или по национальным кухням. Так, 107 (37.68%) n-грамм являются корректными названиями ресторанов, их суммарная частотность составила 140 (39%). 37 (13.03%) n-грамм являются названиями подклассов ресторанов, их суммарная частотность – 47 (13.09%). 52 (18.31%) n-грамм содержат названия ресторанов, но требуют очистки от лишних частей, при этом частотность данных n-грамм в корпусе составляет 2 и менее, суммарная частотность – 54 (15.04%). 45 (15.85%) n-грамм содержат названия подклассов ресторанов, но требуют очистки от лишних частей, при этом частотность данных n-грамм в корпусе составляет 2 и менее, суммарная частотность – 48 (13.37%). 43 (15.14%) n-грамм не являются названиями ресторанов, их суммарная частотность составила 65 (18.11%).

Названия корпораций

Следующим экспериментом стал поиск названий корпораций. Поисковый запрос для этого эксперимента представлен в (5).

Q = ((wordform, *korporaciya*, “POSS_3SG,SG”, right, 1, 10, exact), 7, 0, 95, 80) (5)

В результате выполнения этого поискового запроса был получен список из 138 n-грамм, содержащих лемму “*korporaciya*” с морфологическими свойствами POSS_3SG и SG в последней позиции, которые встречаются в корпусе 606 раз. Среди них при ручной проверке было найдено 63 (45.65%) n-грамм, которые являются корректными названиями корпораций, их суммарная частотность – 178 (29.37%). 27 (19.57%) n-грамм содержат названия корпораций, но требуют

дополнительной очистки, суммарная частотность этих n-грамм составила 29 (4.79%). Среди результатов было выделено 15 (10.87%) n-грамм, которые являются неполными названиями корпораций, их суммарная частотность – 58 (9.57%). 30 (21.74%) n-грамм являются названиями подклассов корпораций по принадлежности к отрасли, географическому положению, доли участия государства, такие n-граммы встречаются в корпусе 336 раз (55.45%). 3 (2.17%) n-грамм не являются названиями корпораций, их суммарная частотность составила 5 (0.83%).

Сравнение результатов

Для различных классов именованных сущностей алгоритм показывает различные результаты, показанные в Таблице 3.

Таблица 3. **Результаты экспериментов**

Класс именованных сущностей	Правильные	Требуют фильтрации	Требуют расширения	Правильные названия подклассов	Названия подклассов, требующие фильтрации	Неверные	Всего
Названия министерств	100%	0%	0%	0%	0%	0%	50
Названия улиц	72%	12%	0%	0%	0%	16%	600
Названия ресторанов	37.7%	18.3%	0%	13%	15.9%	15.1%	285
Названия корпораций	45.7%	19.6%	10.9%	21.7%	0%	2.2%	138

Заключение

Алгоритм для извлечения именованных сущностей, предлагаемый авторами в этой статье, показывает различные результаты в зависимости от типа именованных сущностей. Представленные результаты показывают правильность распознавания от 37.7% до 100%.

В дополнение к основной задаче распознавания именованных сущностей, алгоритм применим для решения задачи распознавания названий подклассов именованных сущностей. Эта особенность может быть использована для решения дополнительных задач, таких как классификация текстов, определение тематики текстов и других задач обработки текста.

Анализ результатов, полученных в ходе экспериментов, показывает, что для повышения точности и правильности алгоритма, необходимы тонкая настройка параметров, формирование расширенных словарей для распознавания именованных сущностей, использование дополнительной постобработки результатов.

ЛИТЕРАТУРА

1. Nevzorova O., Mukhamedshin D., Gataullin R. Developing Corpus Management System: Architecture of System and Database //Proceedings of the 2017 International Conference on Information and Knowledge Engineering. – 2017. – С. 108–112.

2. Aibaidula Y., Lua K. T. The development of tagged Uyghur corpus // Proceedings of the 17th Pacific Asia Conference on Language, Information and Computation. – 2003. – С. 228–234.

3. Nevzorova O., Mukhamedshin D., Kurmanbakiev M. Semantic aspects of metadata representation in corpus manager system //Open Semantic Technologies for Intelligent Systems (OSTIS-2016). – 2016. – С. 371–376.

4. Suleymanov D. et al. National corpus of the Tatar language “Tugan Tel”: grammatical annotation and implementation //Procedia-Social and Behavioral Sciences. – 2013. – Т. 95. – С. 68–74.

5. Segura-Bedmar I., Martínez P., Zazo M. H. Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013) //Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013). – 2013. – Т. 2. – С. 341–350.

6. Zhou G. D., Su J. Named entity recognition using an HMM-based chunk tagger //Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. – Association for Computational Linguistics, 2002. – С. 473–480.

7. Ando R. K., Zhang T. A framework for learning predictive structures from multiple tasks and unlabeled data //Journal of Machine Learning Research. – 2005. –№. 6 (Nov). – С. 1817–1853.

8. Agerri R., Rigau G. Robust multilingual named entity recognition with shallow semi-supervised features //Artificial Intelligence. – 2016. – Т. 238. – С. 63–82.

УДК 81'42

SOME FEATURES OF COMMUNICATION STRATEGIES AND TACTICS IN RUSSIAN YOUTUBE VIDEO BLOGGING

A. V. Novozhilov
National Research University "Higher School of Economics", Saint Petersburg
avnovozhilov@edu.hse.ru

The article presents modern practices of video blogging in Russian-speaking segment of YouTube. Also, the article presents the research on communication strategies and tactics, which popular video bloggers use to attract the audience.

Keywords: communication strategies, communication tactics, video blogging, YouTube.

ОСОБЕННОСТИ КОММУНИКАТИВНЫХ СТРАТЕГИЙ И ТАКТИК ВИДЕОБЛОГИНГА В РУССКОМ СЕГМЕНТЕ YOUTUBE

А. В. Новожилов
Национальный Исследовательский Университет «Высшая Школа Экономики», Санкт-Петербург
avnovozhilov@edu.hse.ru

В статье описываются практики видеоблогинга в русскоязычном сегменте YouTube, а именно проводится исследование коммуникативных стратегий и тактик, которые используют популярные видеоблогеры для привлечения аудиотории.

Ключевые слова: коммуникативные стратегии, коммуникативные тактики, видеоблогинг, YouTube.

С появлением и распространением Интернета в 1990–2000 годах культурное пространство для лингвистических исследований значительно расширилось. Появились такие неологизмы, как «лайк», «шер», «репост», возникли новые лингвистические феномены – «олбанский» язык, изменение синтаксиса в чатах и проч.

В 2005 году был создан YouTube – интернет-платформа для публикации видео. Согласно официальным данным, в 2013 году этой платформой пользовались уже 51 миллион россиян [Благовещенский 2018]. Согласно данным Росстата на 1 января 2018 года на-

селение России составляет 146880,4 тыс. [Население…2018]. Таким образом, YouTube пользуются около 34.93 % россиян. Мы приводим все эти цифры для того, чтобы очертить масштаб такого культурно-технологического феномена, как YouTube.

Видеохостинги становятся новым средством коммуникации, в которых формируется особая культура интернет-общения, обладающая своим собственным языком. Особый интерес представляет речь видеоблогеров, ведущих собственные каналы с многомиллионными подписчиками, а также коммуникативные тактики, которые они используют, работая со своей аудиторией и создавая свои личные бренды.

Целью данного исследования является нахождение общих для русскоязычного сегмента YouTube коммуникативных тактик видеоблогинга.

Лингвистические практики интернет-общения в русскоязычном сегменте Интернета в последнее время привлекают все больше исследователей (см., в частности [Боровенков А.Е. Видеоблогинг: сетевые коммуникации и коммуникативные позиции, С. 17–23], [Малюкова А.И., Ахметьянова Н.А. Видеоблогинг как средство интернет коммуникации, 270–271]). Основные теоретические работы, на которые опирается данное исследование – это книги, посвященные исследованию коммуникации и дискурса: монография О. С. Иссерс «Коммуникативные стратегии и тактики устной русской речи» [Иссерс, 2008, 277], книга П. Бейкера и С. Эллис «Ключевые понятия дискурсивного анализа» [Baker, Ellece, 2011, 241], а также статья Дж.Р.Серль «Что такое речевой акт?» [Дж. Р. Серль, Что такое речевой акт, С. 151–169].

1. Теория и методика

Для того чтобы исследовать речевые стратегии и тактики, следует в первую очередь установить несколько априорных положений.

Во-первых, почти у любого сказанного или написанного всегда есть цель. «Своеобразным подтверждением того, что говорящие осознают стратегическое назначение дискурса, могут служить данные американского социолингвиста Сюзан Ервин-Трипп. Автор анализировал различные способы оформления просьб и их восприятие информантами. Эксперимент показал, что говорящие знают о том, что просьбы могут осуществляться неконвенционально, и поэтому используют следующее правило интерпретации: если высказывание может быть истолковано как просьба, то его именно так и следует прежде всего толковать. По сути, мы имеем дело с гиперкоррекци-

ей – превышением иллокутивной силы высказывания. Но это свидетельствует о том, что люди воспринимают речь как способ достичь определенных целей, поэтому и в получаемых ими сообщениях они в первую очередь пытаются обнаружить целевую установку» [Иссерс, 2008: 51]. Здесь нужно оговориться, что О. С. Иссерс работала в основном с медийным дискурсом. В медийном дискурсе цель – повлиять на восприятие мира своего собеседника или предполагаемого слушателя. Однако существуют такие виды речевых актов, которые не преследуют данную цель.

Возвращаясь к нашей теме, стремление повлиять на картину мира слушателя можно назвать «прогнозированием» (как и у О.С. Иссерс), так как речевая стратегия предполагает под собой конечную цель речевого акта, то есть собственно желаемый прогноз, каким образом взаимодействие коммуникантов должно завершиться. Во-вторых, стратегия предполагает коррекцию своего поведения при нежелательных реакциях слушателя или воспринимающего субъекта. Получается, что для коррекции говорящий должен следить за тем, как его речь реализует его установку и его цель.

Таким образом, можно сказать, что в основе коммуникативной стратегии лежит желание коммуниканта дойти до цели, или, как пишет О. С. Иссерс: «речевые стратегии могут быть описаны как совокупность процедур над моделями мира участников ситуации общения» [Иссерс, 2008: 53]. Проще говоря, коммуникативная стратегия – это осознаваемая или подсознательная цель, к которой участник коммуникации хочет прийти путем совершения последовательности речевых актов. И этот самый путь или способ – это речевая (коммуникативная) тактика. Например, целью, стратегией может быть выведывание информации, а тактикой – канюченье *(Ну, скажи, ну тебе сложно что ли, ну, скажи)*. Однако наивно было бы предполагать, что в реальном мире цель коммуниканта может быть достигнута путем одного коммуникативного хода. Чаще всего она (цель) достигается путем построения сложной системы из нескольких коммуникативных ходов, для каждого из которых, или хотя бы для некоторых, выбирается своя тактика.

Для дальнейшего исследования, мы должны углубиться в теорию речевых актов и дать определения понятиям: речевой акт, коммуникативный ход. Также в исследовании мы будем пользоваться понятием «фрейм» и теорией структуры убеждений Р. П. Абельсона.

«Речевые акты обычно производятся при произнесении звуков или написании значков. Какова разница между просто произ-

несением звуков или написанием значков и совершением речевого акта? Одно из различий состоит в том, что о звуках или значках, делающих возможным совершение речевого акта, обычно говорят, что они имеют значение (meaning)… В статье под названием «Значение» Грайс дает следующий анализ одного из осмыслений понятия *meaning:* «Сказать, что *А* что-то имел в виду под *х* (A meant something by *x)* – значит сказать, что *«А* намеревался, употребив выражение *х,* этим своим употреблением оказать определенное воздействие на слушающих посредством того, что слушающие опознают это намерение» [Серль, 1986, 151–169]. Таким образом, речевой акт – это высказывание, наделенное значением и обладающее намерением воздействовать на слушающего. При этом стоит сразу отметить, что коммуникативная стратегия, выраженная в речевом акте, не обязательно должна быть выражена устно, так же как и реакция слушающего. В некоторых коммуникативных ситуациях, например, в нашей, реакция одобрения или неодобрения слушателем может выражаться в количестве просмотров видео, «лайках» или «дизлайках» – видимых индикаторах реакции аудитории. Так же реакция «слушающих» может выражаться в комментариях, оставленных под видео блогера.

Другим важным понятием является «коммуникативный ход», который можно определить, как один из элементов речевого акта, который, основываясь на стратегии, помогает решить коммуникативную задачу. Далее встает вопрос, а чем руководствуются коммуниканты, выбирая тот или иной тип речевой тактики, чтобы достигнуть цели. Для решения этого вопроса обратимся к теории фреймов и теории структуры убеждений Р. П. Абельсона.

Если говорить простым языком, фрейм – это модель поведения, которая хранится у нас в сознании: что нужно сказать, как нужно реагировать на высказывание и так далее. Очень близким к фрейму понятием является ритуальный речевой акт, который реализуется всегда примерно по одному сценарию, то есть опять же фрейму. Примерами таких ритуальных речевых актов может быть поздравление с праздником, надгробная речь, благодарственная речь. Но стоит уточнить, что даже ритуальная речь по имеющемуся в сознании фрейму может быть коммуникативным ходом при реализации коммуникативной тактики. Она способствует реализации речевой стратегии, потому что даже у этого ритуала есть цель: информирование, создание и поддержание образа своего «я» и прочие.

Теории фреймов достаточно близка упомянутая теория Абельсона, однако, в ней нам важна только незначительная ее часть. Так, Р. П. Абельсон, разрабатывая лингвистическую теорию искусственного интеллекта, предложил метафорической представление речи (речевых ходов), как атомов, молекул и планов. Согласно теории Абельсона, у коммуниканта появляется желание (цель). Исходя из этого желания, говорящий представляет в своей голове замысел – то есть «атом замысла». Замысел реализуется в действиях – других атомах. А атомы вместе образуют молекулу. «Таким образом, молекула в модели Абельсона передает идею действия, предпринятого для достижения некоторой желанной для инициатора цели» [Иссерс, 2008, 85]. Молекулы и простые планы всегда являются заготовленными, то есть уже есть в «речевом багаже» коммуниканта. То есть можно сказать, что эти молекулы и планы являются ритуальным речевым актом или фреймом.

По нашему мнению, таким фреймом или молекулой можно назвать приветствие и прощание в видео каждого блогера со своей аудиторией.

Далее рассмотрим культурологические и технологические термины «блогер», «влог», «блог».

Согласно «Новому словарю иностранных слов», блог – это «сетевой дневник, публичный дневник пользователя интернета» [Новый словарь иностранных слов, 2009]. Соответственно блогер – это тот человек, который ведет интернет-дневник, то есть собственно блог.

Так как слово «влог» «просочилось» в русскоязычный Интернет только в качестве слэнга, для его толкования нам придется обратиться к иностранным словарям. «Vlog – a blog (= webpage with information about someone's activities or interests) containing video rather than only text»[1] [Financial and business terms, 2012]. Исходя из данного определения, можем вывести, что видеоблогер – это человек, который публикует свой дневник в формате видео. Основной интернет-платформой для видеоблогеров является YouTube. Теперь обратимся к критериям популярности, по которым мы отбирали видео. Для исследования был отобран видео-контент по следующим критериям: длина, тема, популярность, наличие/отсутствие других участников в кадре. Так, длительность видео ограничена 7. Тема, как уже было сказано выше, очерчена персональным блогом о жиз-

[1] Влог – это блог (веб-страница с информацией о чьих-либо занятиях или интересах), состоящий преимущественно из видео.

ни, то есть влогом, а популярность – количеством просмотров (не менее 1 000 000) и выражений одобрения в виде «лайков» (не менее 50 000). По критерию наличия/отсутствия других участников в кадре будут отбираться видео с одним участником/цей, подпадающих под классический жанр «влога».

«Лайк» калька с английского «like» – «to show that you think something is good on a social networking website by clicking on a symbol or the word 'like'» [Online Cambridge Dictionary]. Таким образом, «лайк» – это кнопка с символом одобрения контента, а «лайкать» – это выражать свое удовлетворение от увиденного через нажатие этой кнопки.

Подписчик, или follower, – это человек, проявляющий интерес и следящий за обновлением контента в блоге или на странице в социальных сетях определенной группы или человека.

Были отобраны шесть видео, подпадающие под обозначенные нами критерии. Так, канал Вангая (Иван Рудской) «EeOneGuy» занимает 10 место по количеству подписчиков в русскоязычном сегменте YouTube, уступая места только музыкальным каналам и каналам с детскими мультфильмами. В табл. 1 приведены данные об используемом в работе видеоматериале.

Таблица 1. **Отобранные блогеры и влоги**

Название канала и автор	Количество подписчиков	Название видео и ссылка на него	Количество просмотров	Длина видео (мин.)
Канал: Ивангай **Автор**: Иван Рудской (под псевдонимом Ивангай)	12 303 116	**Ссылка:** https://www.youtube.com/watch?v=Wbu4gtQ-O7k **Название видео**: «Нарисуй-ка мне…»	22 502 685	5,48
Канал: TheKateClapp **Автор:** Катя Трофимова	5 858 731	**Ссылка:** https://www.youtube.com/watch?v=M8RVoNT4CSA **Название видео**: «МОЕ УТРО! / Завтрак, Повседневный Макияж»	8 781 848	6,56

Продолжение таблицы 1

Канал: МарьянаРо **Автор**: Марьяна Рожкова	5 497 602	**Ссылка:** https://www.youtube.com/watch?v=1BJG3ft0cnY **Название видео**: «ЧТО В МОЕЙ СУМКЕ, МАРЬЯНА РО»	3 019 762	5,27
Автор: Саша Спилберг **Канал**: Спилберг	5 380 823	**Ссылка:** https://www.youtube.com/watch?v=HM-lqDtATPo **Название видео**: «Draw My Life / История Моей Жизни»	9 025 387	5,42
Автор: Алена Венум **Канал**: Алена Венум	2 751 154	**Ссылка:** https://www.youtube.com/watch?v=rv1PFzOir6s **Название видео**: «10 ФАКТОВ О МОЕМ ТЕЛЕ»	4 916 123	4,30
Канал: Anny May	3 717 694	**Ссылка:** https://www.youtube.com/watch?v=19z1a5IYTsY&index=25&list=PLrQ6ALlJ3A8_ODPRAzi f871KWJQxiZJd **Название видео**: «Я БЕРЕМЕННА»	3 102 556	5,00

2. Анализ видеоконтента и его результаты

Как было отмечено ранее, видеоблогинг можно отнести к медиакоммуникации, поэтому основная цель говорящего в такой речевой ситуации – повлиять на картину мира собеседника. Это подтверждается и теорией медийной речи. Массовая коммуникация, и видеоблогинг в том числе – это социально ориентированное общение. «Анализируя этот вид общения, к которому, прежде всего, относится массовая коммуникация – пишет М. Н. Володина, – стоит отметить, что у него двойной субъект. С одной стороны, …непосредственно осуществляющий такое общение.… С другой стороны, коллектив или группа, на которые направлено такого рода воздействие…» [Володина 2008, 137]. Правда стоит заметить, что видеоблогер – это «медиа в себе», то есть он и создает информацию, он

же ее и доносит до своих «подписчиков». Рассмотрим подробнее, что понимается под влиянием на картину мира зрителя/слушателя? Побуждение слушателя к совершению определенных действий, выгодных говорящему. В данном случае к таким действиям можно отнести: подписку на канал блогера, просмотр других его видео, выражение одобрения «лайком» и проч. Таким образом, блогер находит и увеличивает аудиторию, производя видеоконтент, способный заинтересовать своего зрителя, в обмен на одобрение и выражение лояльности (подписка).

Получается, что основная цель, основная стратегия – это привлечение подписчиков, потому что, во-первых, YouTube платит за количество просмотров, во-вторых, стоимость рекламы в видео повышается. Исходя из этой стратегии, определяются пути ее достижения, то есть коммуникативные тактики.

Другой важной стратегией является построение «тайного» круга друзей-слушателей, посвященных в жизнь блогера. Это еще можно назвать удержанием лояльной аудитории и поддержанием сконструированной у этой аудитории картины мира. Сама специфика видеоблога о своей жизни способствует этому.

Таким образом, можно выделить три основных коммуникативных стратегии: влияние на картину мира слушателя/зрителя, привлечение новых подписчиков и поддержание у уже имеющейся аудитории ощущения избранности.

Теперь обратимся к коммуникативным тактикам. Первый важный аспект, на который стоит обратить внимание – это тактика селфбрендинга. С одной стороны, личность, собственный образ – это то, что лежит в основе феномена влога. Люди смотрят на жизнь, личность и образ блогера. Выражаясь экономическими терминами, они «покупают» этот товар в обмен на свое время и лояльность к бренду. Таким образом, можно сказать, что блогер – это медиа, которое производит культурный, развлекательный контент из своей личной жизни.

Получается, что мы оказываемся в области междисциплинарных исследование маркетинга, литературоведения и коммуникации. Проведем первую смелую аналогию. Влог, как новая форма коммуникации, похож на литературное произведение, на драму, в частности. И здесь обратимся к идеям М. М. Бахтина. Н. А. Кафидова в своем обзоре пишет: «сформирован взгляд на произведение как на коммуникативное событие общения автора, героя и читателя, что приводит к осознанию читателя имплицитного и эксплицитного. Так, с одной стороны, возникает представление о читателе как сотворце эстети-

ческого события, носителе «формирующей активности» (М. М. Бахтин)» [Кафидова 2010, 98]. Что имеется в виду, когда говорим об имплицитном читателе в данном контексте? Видеоблогинг в целом, и каждый видеоблогер в частности находят свою целевую аудиторию. Точнее, в нашем случае, наоборот. Целевая аудитория находит видеоблогера. Это как раз и имеется в виду, когда применяем теорию имплицитного читателя Бахтина к видеоблогингу. Видео такой тематики «находят» свою аудиторию, тех людей, которым был бы этот контент интересен. Таким образом, смысл видеоролика так же, как смысл литературного произведения не воспринимается, как готовый, а считывается только в акте восприятия.

Позволим себе также предложить и другую смелую аналогию. Видеоблогинг, по крайней мере, тот его тип, который мы рассматриваем, а именно влог о своей жизни и личности, достаточно схож с литературоведческой концепцией жизнетворчества, получившей свое распространение в начале 20 века. «Жизнетворчество выступает, как сознательное структурирование собственной жизни, процесс ее формообразования и стилизации в заранее выбранном направлении, когда человек предстает автором-героем своего жизненного повествования» [Кулагина 2012, 155]. Автор-герой – вот главное слово, которым можно описать рассмотренные влоги.

Покажем это на примере видео «Draw my life» [URL: https://www.youtube.com/watch?v=HM-lqDtATPo]. «Итак, вот нарисованная история моей жизни или Draw my life. Я родилась 27 ноября 1997 года, была очень маленькой и няшной и не весила почти ничего». Здесь блогер (Саша Спилберг) – автор видео о своей жизни, соответственно, она же его главный герой. Она сама рассказывает историю своей жизни, являясь одновременно повествователем и объектом повествования. Таким способом, создавая свой образ, свой бренд, конструируя публичную мифологию своей жизни, Саша Спилберг выстраивает стратегию самопрезентации. Причем обозначенная цель – влияние на картину мира – обязательна для данной стратегии. Она (автор) как бы сообщает о своем существовании, утверждает его.

Другая особенность тактик, на которую хочется обратить внимание – это сочетание ритуализированной речи, запланированной с импровизацией или видимостью ее.

Начнем с приветствия. Интересная черта приветствия в видеоблогинге – это то, что оно, являясь одновременно ритуалом, в этот же момент создает и поддерживает публичный образ блогера. При-

ветствие каждого видеоблогера в чем-то уникально. Давайте разберем на примерах.

Самое известное приветствие влогера в русскоязычном YouTube – это традиционное приветствие Вангая: «Хаю-хай, с вами Ивангай!». Иван Рудской сделал такое приветствие своим фирменным знаком. А вот приветствие Марьяны Ро: «Привет ребята, меня зовут Марьяна, это мой канал, и добро пожаловать на него». Приветствие Марьяны довольно обычное, ритуализированное, однако с помощью интонации и видеоряда она поддерживает свой образ милой и обеспеченной девушки. Гораздо более ярко в плане интонации приветствие Кати Клэпп: «Всем ппривет, меня зовут Катя Клэпп, и сегодня на моем канале очень гламурное видео, потому что это май морнинг рутин». Мы специально расшифровали *my morning routine* русским транслитом, потому что Катя в этот момент иронично произносит это с русским акцентом. И такая самоподача тоже поддерживает построенный образ «милой домоседки». С помощью, с одной стороны, использования английских слов в своей речи, автор видеоролика показывает условно, конечно же, уровень своего образования и культурного уровня. Получается, что она следит за англоязычным интернетом и западной популярной культурой. Именно такую личностную конструкцию она и транслирует через свои видео. С другой стороны, наигранный русский акцент добавляет к словам Клэпп долю самоиронии.

Таким образом, на этих трех примерах мы проиллюстрировали представленный тезис о сочетании ритуальной речи и маркетинга в приветствиях, которые используют авторы влогов. Они каждый создают свой «ритуал» приветствия напрямую через лексику или интонационно и этот ритуал становится способом поддержания собственного бренда.

Расширяя выделенную нами особенность приветствия на весь создаваемый представленными авторами видеоконтент, можно заметить сочетание в видео четкой драматургии, присущей продуманной речи и импровизации.

Опять же попробуем на примерах доказать этот тезис. Возьмем видео Вангая. Какие можно выделить там смысловые части. Во-первых, приветствие, которое мы уже разбирали выше. Затем завязка: «Шел я значит такой, а тут вдруг: «…Что это?».

Потом по драматургии вводится объяснение, переломная точка, вокруг которой все вертится. Автор подводит нас к тому, что он будет рисовать: «Все указывало на то, что это ГРАФИЧЕСКИЙ

ПЛАНШЕТ! Я же когда-то был художником и даже картины рисовал!». Наконец, начинается саморазвитие действия: «(Вставляет просьбу одного из его поклонников из личных сообщений). Нарисуй себя, когда нет интернета. (Задумывается). Ладно (низким голосом, затем рисует)». Такой элемент сюжета повторяется еще 10 раз.

В конце видео – развязка: «Спасибо всем тем, кто отсылал мне комментарии. Спасибо тебе, что ты посмотрел это видео, и не забудь подписаться на канал и поставить этому ролику «лайк», если ты, конечно, не подписан и уже не поставил лайк. Но и чтоб ты знал, с тобой был Ивангай». Развязка получается тоже ритуализированным речевым актом. Мы потом увидим повторение одной и той же просьбы подписаться и поставить «лайк» во всех видео с той лишь разницей, что эти просьбы могут быть вставлены в разные сюжетные части.

Та же схема драматургии повторяется и в остальных видео. В видео «Draw my life»: приветствие – поэтапная история жизни автора – прощание – приглашение к подписке на канал.

В видео Кати Клэпп «МОЕ УТРО! / Завтрак, Повседневный Макияж» та же структура: приветствие – четкая разбивка на повторяющиеся сюжетные элементы – прощание – приглашение к подписке на канал.

Похожая схема наблюдается и в остальных видео. Немного отличается она в видео Анни Мэй, но в целом видео следует той же структуре: завязка – действие – развязка – приглашение к подписке.

Теперь возвратимся ко второй части тезиса про импровизацию, присущую устной речи. Вот выдержка из расшифровки видео Алены Венум «10 фактов о моем теле»: «Мама себя тоже чувствовала, понимаете, так себе…» [URL: https://www.youtube.com/watch?v=rv1PFzOir6s]. В этот момент она рассказывает об осложнениях при родах ее мамы. В такую личную историю автор вставляет «понимаете» – обращаясь к зрителю, как к реальному, телесному, присущему на момент просмотра видео собеседнику. Получается, что и зритель себя чувствует в такой же ситуации: так, как будто его посвящают в интимную тайну тет-а-тет за дружеским разговором.

Это выдержка из расшифровки видео Анни Мэй «Я БЕРЕМЕННА»: «И вот этому пупсику 14 недель, представляете? Серьезно, ему 14 недель! И что делать дальше? Как вести канал? Как вообще все?.. блин, я не знаю» [URL: https://www.youtube.com/watch?v=19z1a5IYTsY&index=25&list=PLrQ6AL8_lJ3AODPRAzi0f871KWJQxiZJd].

Обилие риторических вопросов, пауз, обращение к зрителю – все это опять же создает несколько эффектов. Во-первых, импровизированной монологической очень интимной речи. Зрителя как будто посвящают в тайны. Создается ощущение уникальности, особенности каждого зрителя для автора ролика, если она рассказывает им такие подробности своей частной жизни. Усиливает этот эффект «избранного круга» следующие слова: «Эта информация, которую вы сейчас услышали, вообще требует подписки, если вы все еще не подписаны, и лайка… Оооо, кошмар… Я даже не знаю, как мне идти домой».

Сильный эффект импровизированной речи создается в видео Марьяны Ро: «Господи... что за это… Ммм, приятно...что такое... Это помада!». Неправильный синтаксис, обилие пауз – все это элементы присущие устной импровизированной речи. В этом же видео создается эффект внутреннего монолога автора с собой: «Я сама немножечко в шоке, потому что я …зачем я ношу носок?» [URL: https://www.youtube.com/watch?v=1BJG3ft0cnY].

Таким образом, наличие четкой драматургии увлекает зрителя, приравнивает видеоролик к озвученному рассказу. В свою очередь содержание этого ролика, то есть разговорная речь, импровизация или эффект импровизации в некоторых местах создают эффект избранности зрителя, эффект особого круга, посвященных в тайну людей.

Говоря об эффекте «тайного» круга друзей, обратимся к третьей стратегии – создание и поддержание этого эффекта. С помощью каких тактик она реализуется?

Во-первых, это использование специфического юмора, иронии и самоиронии. Не секрет, что разное понимание «смешного» дифференцируют людей на группы, поэтому использование юмора, присущего одной из таких групп, может расцениваться, как сигнал: «Я такой же, как вы». Вот самый яркий пример из видео Ивана Рудского: «Нарисует, как гангает инвокер. Что? Вот ты задрот. ЗАДРОООООООООООООООООООООООООТ. Ладно, нарисую. (рисует). Что, вы спросите, где инвокер? На миду стоит. Профессиональный юмор. Это так страшно (смеясь)» [URL: https://www.youtube.com/watch?v=Wbu4gtQ-O7k]. Весь этот непонятный набор слов «гангать», «инвокер», «мид» относится к сленгу группы людей, играющих в популярную интеренет-игру Dota-2. Только игрокам или людям, знакомым со спецификой этой игры, понятен эффект комического в этих словах.

Другая тактика, воплощающая в жизнь ту же стратегию, это тактика разделения секрета между блогером и зрителем. Рассказывая

какую-то интимную подробность своей жизни, какие-то переживания и трудности, выводя в публичное пространство саморефлексию, блогеры как бы делятся секретом. И это тоже знак аудитории: «Теперь у нас с вами есть общий секрет, я с вами им поделилась(лся)». Примеры можно найти в видео 2-6. Возьмем один (видео 3): «Я настолько заигралась в популярность, что только об этом и думала. Я не общалась с семьей, постоянно общалась с одноклассниками, из-за этого меня решили отправить в интернат за границей, и мне снова пришлось расстаться со своими друзьями и со своим парнем». Здесь автор (Саша Спилберг) публично признает свои ошибки – «я заигралась» – и рассказывает, как ей было трудно – «решили отправить в интернат», «пришлось расстаться». Жанр, в котором это происходит, мог бы больше походить на дружескую исповедь, чем на видеоблог. Но это влог, и такого рода откровенность встречается в 5 из 6 роликов, что мы расшифровали, и в 2/3 роликов, которые мы отсмотрели. Следовательно, мы можем говорить в данной ситуации о тенденции и общей коммуникативной тактике видеоблога.

Еще одна тактика, реализующая эту же стратегию – это интерактивное общение со зрителем. Его можно выразить формулой: «Я рассказал(а) о себе – теперь поделитесь своими историями». Такой прием встречается довольно часто. «И делитесь своими комментариями внизу по поводу бодрого утра, как вы его достигаете!» (видео 2) [URL: https://www.youtube.com/watch?v=M8RVoNT4CSA]. Здесь блогер приглашает аудиторию к взаимодействию, показывая, что ему или ей важны и интересны жизни своих подписчиков, что только поддерживает эффект дружеского круга. В этом примере взаимодействие вообще происходит напрямую, в личных аккаунтах блогеров в соц.сетях. «(Показывает комментарий из социальной сети Вконтакте с просьбой и озвучивает ее)» (видео 1) [URL: https://www.youtube.com/watch?v=Wbu4gtQ-O7k].

Заключение

Итак, мы провели краткий анализ коммуникативных стратегий и тактик видеоблогинга в русскоязычном сегменте YouTube, привели несколько литературоведческих аналогий и пришли к следующий выводам:

Авторы видеоконтента рассказывают о себе, конструируют свой публичный образ, таким способом создавая себя как героя массовой культуры.

Из этой «перерожденной идеи жизнетворчества» вытекает форма подачи материала, сочетающая в себе черты четкой, запланированной драматургии и импровизированной устной речи.

Обилие обращений к зрителю, приглашение к подписке на канал, а также темы видео, которые порой затрагивают очень интимные моменты жизни авторов, создают ощущение избранности зрителя, посвященности его в тайны.

Авторы видео создают свои собственные ритуалы. Они берут некоторую стандартную форму, например, приветствие, и наполняют ее своим собственным особенным содержанием, например: «Хаюхай, с вами Ивангай». Таким образом, авторы, создавая и закрепляя такие ритуальные реплики, поддерживают свой бренд.

Блогеры создают публичную мифологию собственной жизни, собственный бренд. Их деятельность направлена на монетизацию и продвижение этого бренда.

С другой стороны, по стратегии действия, по цели коммуникации блогеров можно отнести к категории медиа. Они так же, как и медиа, имеют целью влияние на картину мировидения зрителя/слушателя/читателя. Можно также сказать, что основной стратегией блогера является выработка лояльности зрителя.

Исходя из этой стратегии, можно выделить несколько коммуникативных тактик, реализующих стратегию: вверение тайн, обращение к зрителю, интерактивное общение, например, «напишите необычные факты в комментариях о вашем теле». Таким способом создается эффект заинтересованности, тактика проявления интереса. Автор как бы говорит: «Я рассказал(а) свои секреты и готов(а) выслушать ваши».

Схожей стратегией является побратимство для создания эффекта «избранного круга»: «Я счастлива, что у меня есть вы, потому что я хочу идти по этому пути вместе с вами, я искренне верю, что вместе мы сможем сдвинуть любые препятствия на нашем пути. Навсегда ваш лучший друг, Саша Спилберг» (видео 3) [URL: https://www.youtube.com/watch?v=HM-lqDtATPo].

ЛИТЕРАТУРА

1. Володина Н.М. Язык средств массовой информации: учебное пособие для вузов. М.: Альма Матер, 2008. – 756 с.

2. Иссерс О.С. Коммуникативные стратегии и тактики русской речи. М.: URSS, 2008. – 277 с.

3. Кулагина А.А. Жизнетворческая концепция и принципы создания образа в лирике и драматургии Н.С. Гумилева: дис. канд. фил. наук. М., 2012. – 24 – 155 с.

4. Baker P., Ellece, S. Key terms in discourse analysis. NY., 2011. – 241 p.

5. Кафидова Н. А. Читатель как проблема поэтики / Вестник РГГУ. Серия «История. Филология. Культурология. Востоковедение». – 2010. – №11 (54) – С. 98.

6. Дж. Р. Серль / Что такое речевой акт / Новое в зарубежной лингвистике – М., 1986 – №17 – С. 151–169.

7. Боровенков А.Е. Видеоблогинг: сетевые коммуникации и коммуникативные позиции // Человек. Культура. Образование. 2016. №1 (19). С. 17–23.

8. Малюкова А.И., Ахметьянова Н.А. Видеоблогинг как средство интернет коммуникации. В кн.: Слово и текст в культурном и политическом пространстве. Всероссийская с международным участием очно-заочная научная конференция студентов и аспирантов высших учебных заведений «Слово и текст в культурном и политическом пространстве»: материалы. Сыктывкар. Сыктывкарский государственный университет им. Питирима Сорокина. 2017. 270–271

9. Новый словарь иностранных слов by EdwART [Электронный ресурс] // Академик: [сайт] – URL: https://dic.academic.ru/dic.nsf/dic_fwords/43697/блог (дата обращения: 29.03.2018)

10. Financial and business terms [Электронный ресурс] // Академик: [сайт] – URL: http://business_finance.enacademic.com/26238 (дата обращения: 29.03.2018)

11. Cambridge Dictionary of English Grammar Today [Электронный ресурс] // Online Cambridge Dictionary: [сайт] – URL: https://dictionary.cambridge.org/ru/словарь/английский/like (the date of access: 14.04.2018)

12. Благовещенский А. 51 миллион человек в России смотрят YouTube // Российская газета. – URL: https://rg.ru/2013/04/24/youtube-site.html (дата обращения: 29.03.2018)

13. Карпяк О. Вата с укропом: язык политических мемов // Русская служба BBC – URL: https://www.bbc.com/russian/society/2014/08/140808_ukraine_new_internet_memes.shtml (дата обращения: 29.03.2018)

14. Население (на 1 января 2018г.) [Электронный ресурс] // Росстат [сайт] – URL:http://www.gks.ru/wps/wcm/connect/rosstat_main/rosstat/ru/rates/bfd61f804a41fabfbdc9bf78e6889fb6(дата обращения: 29.03.2018)

15. YouTube для прессы [Электронный ресурс] // YoutTube [сайт] – URL: https://www.youtube.com/intl/ru/yt/about/press/ (дата обращения: 17.03.2018)

УДК 81.322.2; 811.512.132

N-GRAM ANALYZING OF UYGHUR WORDS

M. Orhun
Computer Engineering Department Istanbul Bilgi Unviersity
Istanbul, Turkey
murat.orhun@bilgi.edu.tr

To get statistical result about natural language analysis, it is important to have some linguistic data about language properties. Especially languages which have complicated properties and difficult to define rules, it is even more important to make statistical calculation and apply these results to get some conclusion. N-gram model is a common method to analyze a language at different levels such as, character, word and sentence level. In general, N-gram based modules depends on a corpus and it is the main factor. In this paper, preliminary researches have been done analyzing Uyghur words with N-gram model and results have been discussed with examples.

Keywords: Uyghur N-grams, Uyghur Words, Machine Translation, Uyghur Corpus, Uyghur Sentence.

1. Introduction

In natural language processing it is important to analyze every part of a sentence that consists of phrases, words, helping words, word orders and letters (characters) in each word etc. As a natural language, all cases should be analyzed with respecting to natural properties of a language. In fact, analyzing natural language is very different from analyzing some formal languages. Therefore, there are many problematic cases exist. For some problems, it is possible to define rules and this is the case most scientists prefer. Unfortunately, there are some cases it is difficult or impossible to define a rule. Because such problems not only affected by a current phrase, or sentence, maybe affected by a whole article. This kind of problem may be affected contents of an article, in case it is a problem about word sense disambiguation. Maybe it is a problem about a single character adjusting in a machine translated article. To solve these kinds of problems, if there are some statistical data exist, it might be possible to decide relatively easier.

N-gram module is one of the popular methods to get statistical data about a language. With adjusting gram size, it is possible to collect data at letter (character) level of a word. With these data it might be possible to predicate a character in a word. Because of this reason spell checking

could be done with N-gram modules [1-2]. N-gram modules are also used error correction for OCR processing and error rates have been reduced about 60.2% [3].

Syllabic property of a language is an essential part of a word. With correctly analyzing syllabic structure of a word, it is imaginable to decide whether such a word exist in a particular language or not. In order to decide this, N-gram modules is one of the preferred method to analyze syllabic property of words [4-5]. There are some researches about information extraction, text categorization has been done with using N-grams [6-8].

In Turkic language studies, Turkish is the most researched language comparing to other Turkic languages. In N-gram studies, there are several important researches have been done such as, text normalization, stemming term conflation, information extraction, language identification, spell checking, speech recognition etc. [9-14]. N-gram data usually affected by corpus directly. To get N-gram data about Turkish words some corpus based calculation have been done [15-16].

One of the recent machine translation system about Turkic language is, the machine translation system from Kirghiz to Turkish. In this system statistical machine translation method is used that based on both Kirghiz and Turkish N-gram modules [17]. The translation performance has been evaluated with BLEU method and result measured 0.1 in average [18].

Though N-gram modules is one of the most fundamental approach about natural language processing, but there are not enough resources in Turkic language except than Turkish. The main reason of this is, there is not a corpus exist accessible publicly. In the latest approaches about natural language processing, statistical methods dominant and approaches with hybrid methods show better solutions. Correct N-gram modules and their applications are the parts of hybrid solutions [19].

In this paper, some preliminary N-grams studies about Uyghur words are explained. To get more correct and reliable results, it is essential to have a well implemented corpus. Unfortunately, there is not a corpus available for Uyghur language or even accessible for a research. To solve this problem, a small size of corpus is created from machine readable documents first, after those N-Grams based algorithms are executed on this not tagged corpus. It is also purposing to make a public access corpus for Uyghur language.

This paper is organized as follows: After a short introduction about N-grams in section one, section two explains corpus assembling process for Uyghur language. Section three shows some N-gram data about Uyghur words. Section four evaluates calculated data.

2. Assembling Uyghur Corpus

In computational linguistics statistical approaches include major parts in the recent researches. Unfortunately, building a corpus is a dounting task that consumes a long time and a very expensive cost to ensure a well implemented and user-friendly corpus. This is because a corpus is a collection of all texts in a language. It means that it may include millions of documents in a language those have been published. At the meantime, the quality of the corpus is the significant factor that affects related calculations. Therefore, some languages have compiled corpora around middle of 1900 years and improve them continuously. For example, the first English corpus, Brown corpus, compiled in 1963–1964 with 1 million words and compiled a new version in 1992 with 583 words again [20]. Except this, for English there are other corpora such as British National Corpus (BNC) with 100 million words [21], The Bank of English with 650 million words [22], Corpus of Contemporary American English with 560 million words [23]. Except English there are corpus have been constructed for Russian, Bulgarian, Croatian, Slovenian, Greek, Polish, Spanish, Ukrainian, Turkish, Arabic, Persian, Japanese languages etc. [24]. In Turkic language family Turkish is the only language that has corpus general research [25-26]. Except Turkish, other Turkic language, such as Uyghur have implemented corpora and made some valuable researches. Yet, the corpus is not accessible publicly [27-29]. Some special Uyghur corpus is constructed to make information extraction and used corpus is shared [30].

Because of there is not a general corpus accessible for Uyghur language, in order to make N-gram studies for Uyghur words, an Uyghur corpus is constructed with machine readable documents. To collect documents, it is necessary to process some character normalization operations. In fact, Uyghur people live in three different regions in the world and they use three different scripts mainly. Most Uyghur people live in Xin Jiang Uyghur Autonomous Region in China and they use Arabic scripts officially. The second largest region that Uyghur people are living is the Central Asian republics and they use Cyrillic alphabets. As we know, some of these republics have already decided to use Latin scripts, yet most of the Central Asians still utilize Cyrillic scripts in their daily lives. Except these two regions, in Turkey, Europe, Australia and America, the Latin script Uyghur is used. In order to communicate among these scripts some standardization have been done on Latin scripts for Uyghur language [31-32].

Table I. **The elements of the corpus**

No.	Sources	Words	No.	Sources	Words
1	www.wetinim.com	181,260	10	http://www.xjsport.cn/	12,417
2	http://www.trt.net.tr/uyghur/	19,245	11	http://www.hawar.cn/index.shtml	37,359
3	http://www.akademiye.org/	120,629	12	http://kagsay.com/	15,780
4	http://uyghur.people.com.cn/	32,519	13	http://www.axpaz.com/	32,394
5	http://www.uycnr.com/	30,189	14	Entity corpus [30]	143,983
6	http://uy.ts.cn/	32,081	15	http://www.uycnr.com/(2014)	140,559
7	http://www.nur.cn/index.shtml	20,531	16	Poems	131,227
8	http://uyghur.xjdaily.com/	73,219	17	Stories	306,064
9	http://uyghurpedia.com	44,373	18	Novels	142,741
	Total	1,516,570			

To construct the Uyghur corpus, Uyghur Latin script is used and N-gram algorithms is implemented according to this alphabet. In Uyghur Latin script there are 32 letters.

Vowels: a, e, é, i, o, ö, u, ü

Constants: b, p, t, j, ch, x, d, r, z, zh, s, sh, gh, f, q, k, g, ng, l, m, n, h, w, y.

In this alphabet, there are five characters that consist of double letters such as /zh/, /sh/, /gh/, /ch/, /ng/. To represent these characters with single character, /zh/ is represented with upper case /Z/, /sh/ is represented with upper case /S/, /gh/ is represented with uppercase /G/, /ch/ is represented with upper case /C/ and /ng/ is represented with upper case /N/.

In this case, the Uyghur corpus texts in context consists of following alphabet with small modification of Uyghur Latin script while vowels are not changed.

Vowels: a, e, é, i, o, ö, u, ü

Constants: b, p, t, j, C, x, d, r, z, Z, s, S, G, f, q, k, g, N, l, m, n, h, w, y.

The quality and size of the corpus are main factor that affects calculations result apparently. Because of this reason, corpus texts are selected from official web sites, officially printed electronic documents and some corpora that used in previously researches that available publicly [28]. As a result, the Uyghur people live in three different regions, they

use some local dialects in their publications. Nevertheless, in this corpus, such local dialects are converted into central dialects that used in Xin Jiang Uyghur Autonomous Region.

After the alphabets and characters conversion, the Uyghur corpus contains not tagged plain texts only. The main purpose of this corpus is to make some N-gram calculations about Uyghur words, therefore it is not a balanced text corpus. File length of some elements of the corpus different from other files. Hence, numbers of words are different in each file as shown in Table I.

There are 13 different Uyghur websites are used to collect texts that published articles about culture, politics, sports, economy, technology, literature, history, weather, food, and everyday general news etc. There are five more elements included into to the corpus about poems, stories, novels and two more small corpus that used previous researches [30]. Total there are 1,516,570 words were included from 18 different sources. All sentences are tagged with starting "<S>" and ending "</S>" tags only. Actually, this is not a tagged text corpus except starting and ending tags.

3. Analyzing of words

Table II. **Frequencies of Uyghur Characters**

No	Characters (Latin)	Characters (Arabic Script)	Frequencies (%)	No	Characters (Latin)	Characters (Arabic Script)	Frequencies (%)
1	i	ىئ	15,647	17	p	پ	1,986
2	a	ائ	8,735	18	o	ۇئ	1,860
3	e	ەئ	6,809	19	é	ېئ	1,490
4	l	ل	6,043	20	N	ڭ	1,411
5	n	ن	6,033	21	G	غ	1,382
6	S	ش	5,514	22	z	ز	1,359
7	r	ر	5,060	23	ü	ۈئ	1,359
8	t	ت	4,537	24	C	چ	1,219
9	u	ۇئ	3,885	25	g	گ	0,940
10	m	م	3,581	26	h	ھ	0,934
11	q	ق	3,541	27	w	ۋ	0,900
12	d	د	3,322	28	ö	ۆئ	0,765

13	y	ي	2,940	29	x	خ	0,642
14	k	ك	2,920	30	j	ج	0,474
15	s	س	2,358	31	f	ف	0,050
16	b	ب	2,285	32	Z	ژ	0,018

In this section, different properties of Uyghur words are analyzed with N-gram modules.

• Frequencies of characters: To calculated character frequencies of Uyghur language, with unigram module all characters are counted in the corpus and their distribution is given in Table II. As shown in Table II, the /i/ character is the most frequent among other characters. The least frequent character is /Z/ that represents the “zh” character in Uyghur language. The “zh” character is used for loaded words in Uyghur and most of the words that contain the “zh” character which related to Chinese or Russian sources. Another least used character is the /f/ character. This character is also used to express borrowed words or foreign origin words, such as English and Persian. As a result, frequencies of vowels are higher than constants. The reason is, In Uyghur language, all the words must include vowels and the vowel /i/ is the most used character among them.

• Length of words: In Natural language processing, it is important to know length of the words. Most of algorithms, such as N-gram modules, uses word length as a parameter. Because of Uyghur language is an agglutinative language, it is possible to create a word with unlimited characters theoretically. In this article, words that consist of one to 20 characters are analyzed. As shown in Table III, the most used Uyghur words consist of 5 characters. In general, the majority of word length in Uyghur is the sets of words that consist characters between 2 to 10. In this statistic, multi-words which are written with hyphen are not- included. Other cases the word length may exceed more than 20 characters. There are some examples words that are given in the Table III. According to the statistic, there is one word, “ u” (she/he) detected only. One of the longest word that find in this corpus is “dawamlaSturiSiNlarni”, (hope to you continue). The word length and their percentages are shown in the Figure I. As shown in Figure I, in Uyghur literature words that consist of more than 14 are less used.

• Bigrams of Uyghur characters: spell checking is one of the important tasks in Natural language processing that could be solve with N-gram modules. With statistical calculation of previous characters, probability

of next characters could be solved [1-3]. In this article bigrams counted in this corpus. Some bigrams occurred with high amount of numbers such as, “yé” counted 15860, “ti” counted 122392, “iC” counted 16363 etc. In general, bigrams with /i/ character counted more than other bigrams and this result matches character frequencies that calculated in Table II. There are some bigrams counted less than 5 times. For example, the bigram “jk” counted one time and “jS” counted two times. It means that these characters maybe occurred in adapted words or may be not written correctly in a file.

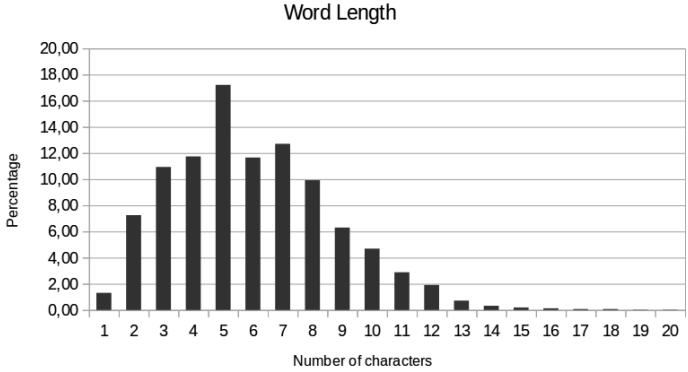


Figure I. Word length of Uyghur words

• Bigram probability of Uyghur words: To estimate the probability of a sentence, bigram module could be used. In order to make this calculation, all words counted in the corpus according to bigram and unigram concept. Bigram probabilities are tested with following Uyghur sentence is tested with those data.

Table III. **Frequencies of Uyghur Characters**

N. char	Word Frequencies (%)	Example	N. char	Word Frequencies (%)	Example
1	1,307	u	11	2,880	yalghanCiniN
2	7,250	bu, Su, eN, su	12	1,912	ateSpereslik
3	10,929	men, sen	13	0,711	ziyaritimizni
4	11,732	ömür, Sair	14	0,320	oqutquCilarniN
5	17,208	deme, budur	15	0,189	tonuSturuluSiCe

6	11,653	jennet, perhat	16	0,125	keSpiyatCilarniN
7	12,703	mesilen, tarixta	17	0,065	dawamliSiSCanliqi
8	9,921	CéCeksiz	18	0,068	milletperwerlikniN
9	6,298	yollanGan	19	0,024	mukemmelleStürül-gen
10	4,685	aCCiqliniS	20	0,023	dawamlaSturiSiN-larni

“kiCik CaGlirim hélimu isimde” (Still, I remember my childhood). The unigram data for each word in the sentence is given in Table IV.

Table IV. **Unigram data for sentence words**

kiCik	CaGlirim	hélimu	isimde
10	2	85	13

The Table IV gives total number of words that counted in the corpus. Because of the size of corpus is small, therefore occurred numbers appear low.

Table V. **Bigram probabilities of Uyghur words**

	kiCik	CaGlirim	hélimu	isimde
kiCik	0,1	0,1	0	0
CaGlirim	0	0	0,5	0
hélimu	0	0	0	0,04
isimde	0,08	0	0	0

In Table V, bigram probabilities of the words are given. As shown in Table V, the word «kiCik» may followed the «kiCik» with the probability of 0,1. And the word «CaGlirim» may follow the «kiCik» with the probability of 0,1. But both of «hélimu» and «isimde» don’t follow the «kiCik».

At the second row, probabilities are calculated of the words that may follow the “Caglirim”. As shown in the Table V, only one word “hélimu” follows with the probability of 0,5. Besides, probabilities for “hélimu” and “isimde” are calculated according to corpus data. The “0” probabilities shows no words follows the current word given in the first column.

4. Conclusion

In this article, to analyze Uyghur words statistically, a corpus is assembled from 18 different resources. Because of Uyghur people live in different regions and use different scripts to publish articles, all published articles are converted into Latin scripts. Also, some text files that used in previous researches have been included [30].

At this level, a small corpus is created approximately with 1.5 million words. At first character frequencies is calculated and found out vowels have more frequencies than constants. The bigram result also shows and proofs that vowels have more usage in Uyghur words. Because vowels are core unit of all Uyghur words. And calculated word length and find out most words consists of 5 characters.

To estimate bigram probabilities of Uyghur sentences, bigram probabilities are calculated of Uyghur words. Indeed, some words give zero probability even in literature some words pair appear together.

To conclude, the size of the corpus is main drawback of this research and size of the corpus is increasing updated files. Apart this, another problem about collecting files is, spelling rules about Uyghur language. The spelling rules have been changed five times over the past 3 decades results that all files cannot be used in the corpus directly.

Since, the spelling rules have been changed five times over the past 3 decades results that all the accessible files in Uyghur cannot be used in the corpus directly.

REFERENCES

1. Barari, L., QasemiZadeh, B.: CloniZER spell checker adaptive language independent spell checker. In: AIML 2005 Conference CICC, Cairo, Egypt, pp. 19–21 (2005)

2. Deorowicz, S., Ciura, M.G.: Correcting spelling errors by modelling their causes.International Journal of Applied Mathematics and Computer Science 15(2), 275–285 (2005)

3. Tong, X., Evans, D.A.: A statistical approach to automatic OCR error correction in context. In: Proceedings of the Fourth Workshop on Very Large Corpora, Copenhagen. Denmark, pp. 88–100 (1996).

4. Kang, S.S., Woo, C.W.: Automatic segmentation of words using syllable bigram statistics. In: Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, Tokyo. Japan, November 27–30, pp. 729–732 (2001).

5. Kyle, Kristopher et al. “Native Language Identification: A Key N-gram Category Approach.” BEA@NAACL-HLT (2013).

6. Brown, P. F., Della Pietra, V. J., deSouza, P.V., Lai “Class-based n-gram models of Natural Language”, Computational Linguistics, vol. 18, pp. 467–479, 1992.

7. W.B. Cavnar, J.M. Trenkle, ”N-gram-based text categorization”, Proceedings of SDAIR-94, 3rd Annua Symposium on Document Analys is and Information Retrieval, Las Vegas, US 161–175, 1994.

8. I. Suzuki, Y. Mikami, A. Ohsato, Y. Chubachi, “A language and character set determination method based on N-gram statistics”, ACM Transactions on Asian Language Information Processing(TALIP), Volume 1 Issue 3, Pages 269–278, 2002.

9. Yıldırım, S., Yıldız, T., An Unsupervised Text Normalization Architecture for Turkish Language, Research in Computing Science 90 (2015)

10. Ekmekcioglu, F. C., Lynch, M. F. and Willett, P. (1996): Stemming and N-gram Matching for Term Conflation In Turkish Texts. Inf. Research, Vol. 2, No. 2.

11. Güven, A., Bozkurt, Ö. Ö., Kalıpsız, Oç, Advanced Information Extraction with n-gram based LSI, World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering Vol:2, No:5, 2008.

12. Bayrak, Ş., Takçı, H., Eminli, M., Makine Öğrenme Yöntemleriyle N-Gram Tabanlı Dil Tanıma, ELECO ‘2012 Elektrik- Elektronik ve Bilgisayar Mühendisliği Sempozyumu, 29 Kasım – 01 Aralık 2012, Bursa.

13. Aşliyan R., Günel K., Yakhno T. (2007) Detecting Misspelled Words in Turkish Text Using Syllable n-gram Frequencies. In: Ghosh A., De R.K., Pal S.K. (eds) Pattern Recognition and Machine Intelligence. PReMI 2007. Lecture Notes in Computer Science, vol 4815. Springer, Berlin, Heidelberg

14. B. Roark, M. Saraclar, and M. Collins, “Discriminative n-gram language modeling,” Computer Speech and Language, vol. 21, no. 2, pp. 373–392, 2007.

15. Çebi, Y., Dalkılıç, G.: Turkish Word N-gram Analyzing Algorithms for a Large Scale Turkish Corpus – TurCo. In: ITCC 2004, IEEE International Conference on Information Technology (2004)

16. Dalkılıç, G., Çebi, Y.: A 300MB Turkish Corpus and Word Analysis. In: Yakhno, T. (ed.) ADVIS 2002. LNCS, vol. 2457, pp. 205–212. Springer, Heidelberg (2002)

17. Tayirova, N., Tekerek, M., Brimkuluv, U., Kırgız ve Türkiye Türkçeleri arasında istatistiksel bilgisayarlı çeviri uygulaması ve başarım testi, Manas Journal of Engineering, V3(issue 2), 2015, pages 59–68.

18. Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). BLEU: a method for automatic evaluation of machine translation. In Proceedings of

the 40th annual meeting on association for computational linguistics (pp. 311–318). Association for Computational Linguistics.

19. Pereira F.C., Singer Y., Tishby N. (1999) Beyond Word N-Grams. In: Armstrong S., Church K., Isabelle P., Manzi S., Tzoukermann E., Yarowsky D. (eds) Natural Language Processing Using Very Large Corpora. Text, Speech and Language Technology, vol 11. Springer, Dordrecht

20. Jurafsky, D., Martin, J.H., Speech and Language Processing, Prentice Hall, 2000, pp. 193–199.

21. British National Corpus (BNC), http://www.natcorp.ox.ac.uk/using/index.xml, May 2018.

22. The Bank of English, https://en.wikipedia.org/wiki/Bank_of_English, May 2003.

23. Corpus of Contemporary American English (COCA), https://en.wikipedia.org/wiki/Corpus_of_Contemporary_American_English, May 2018.

24. Corpus List, https://en.wikipedia.org/wiki/List_of_text_corpora, May 2018.

25. Turkish National Corpus (TNC), https://www.tnc.org.tr/, May 2018,

26. Atalay, N.B., Oflazer, K., Say, B.: The annotation process in the Turkish treebank. In: 4th, International Workshop on Linguistically Interpreted Corpora (2003).

27. Yusup Aibaidula, Kim-Teng Lua,The Development of Tagged Uyghur Corpus. Proceedings of PACLIC17, 1–3 October 2003, Sentosa, Singapore (pp. 228–234). Singapore: COLIPS publications.

28. Mamitimin, S., Ibrahim, T., & Eli, M. (2013). The Annotation Scheme for Uyghur Dependency Treebank. 2013 International Conference on Asian Language Processing, 185 – 188

29. Mamitimin,S., Dawut, U., Chinese-Uyghur Parallel Corpus Construction and its Application, Conference, Using corpora in contrastive and translation studies; 2008; Hangzhou, China, in Using corpora in contrastive and translation studies; 281–295, Cambridge Scholars , Newcastle upon Tyne; 2010

30. Abiderexiti, K., M. Maimaiti, T. Yibulayin and A. Wumaier (2016). Annotation Schemes for Constructing Uyghur Named Entity Relation Corpus. The 20th International Conference on Asian Language Processing (IALP 2016).

31. Duval J.R., Janbaz, W. A., An Introduction to Latin Script Uyghur, Middle East & Central Asia Politics, Economics, and Society Conference. Sept 7–9, University of Utah, 2006, Salt Lake City, USA.

32. Uyghur Alphabets, https://en.wikipedia.org/wiki/Uyghur_Latin_alphabet, May 2018.

УДК 517:811

TOPOLOGICAL APPROACH TO THE ANALYSIS OF COGNITIVE AND LANGUAGE SYSTEMS

P. S. Pankov, S. J. Karabaeva
Institute of Mathematics of the National Academy of Sciences of the Kyrgyz Republic, Bishkek
pps5050@mail.ru, sonun2008@mail.ru

The article describes the generating and analytical models of cognitive activity, i.e. the language of topology is considered, the set of descriptions and the set of schemes having one-to-one correspondence are obtained. On this basis, a linguistic experiment and team competition of students are conducted.

Keywords: model, set, topology, cognitive activity, linguistic experiment, theorem.

ТОПОЛОГИЧЕСКИЙ ПОДХОД К АНАЛИЗУ КОГНИТИВНЫХ И ЯЗЫКОВЫХ СИСТЕМ

П. С. Панков, С. Ж. Карабаева
Институт математики НАН КР, Бишкек
pps5050@mail.ru, sonun2008@mail.ru

В статье описывается порождающая и аналитическая модели когнитивной деятельности, т. е. рассматривается язык топологии, получены, множество описаний и множество схем имеющих взаимно-однозначное соответствие. На этой основе проведены лингвистический эксперимент и командные соревнования студентов.

Ключевые слова: модель, множество, топология, когнитивная деятельность, лингвистический эксперимент, теорема.

Введение

Пространственное расположение и движение объектов можно задавать как языковыми, так и математическими средствами, различными способами. При этом нужно учитывать естественную неясность языковых средств, в том числе топологическую, выявленную Л. Заде [1] в основном для прилагательных. При помощи экспериментов нами эта неясность выявлена и для других частей речи. Также нами обнаружена другая неясность, проистекающая от различного подсознательного использования логических кванторов существования и всеобщности.

Абстрактные понятия могут быть заданы с помощью символов, а названия реальных предметов и объектов не могут быть заданы символами, так как в сознании людей они связаны мысленными ассоциациями, т. е. изображения абстрактных слов могут быть восприняты сознанием как символы, или иначе сказать, осознаны как визуальные образы.

Представления о пространстве начинают формироваться, основываясь на примитивных ощущениях и подсознательных реакциях человеческого (детского) мышления на окружающую среду в и находят свое отражение в языке.

Согласно нашей точке зрения, процесс образования представлений человека об окружающем пространстве представляет собой процесс создания ментальной репрезентации о пространственной организации окружающей среды в лингвистической топологии и имеет пространственные координаты: верх–низ (в связи с гравитацией), право–лево (здесь уже сложнее – либо по отношению к наблюдателю, либо по отношению к наблюдаемому объекту, либо по отношению к движению) или близко–далеко, она определяет местоположение отдельных воспринимаемых предметов.

Также мы считаем, что смысл и применение отдельных терминов основано на нескольких (немногих) услышанных человеком (ребенком) примерах и человеческом свойстве экстраполяции, которая производится неформально и поэтому не может охватить всех возможных случаев.

Также отметим, что в юридической лингвистике в основном рассматриваются высказывания с точки зрения их законности, «оскорбительности», эмоциональной окраски, см. например [3]. Нам неизвестны работы, где изучаются тексты, описывающие топологический и геометрический образ реальной ситуации, как юридические документы. Примеры, в том числе найденные нами, показывают, что это – нетривиальный вопрос. Различное понимание пространственных терминов в протоколе описания места происшествия, свидетельских показаниях может привести к юридической ошибке.

Также имеются ошибки в переводе пространственных понятий с одного языка на другой. Нами найдены примеры отсутствия адекватного перевода (именно перевода, а не объяснения) с одного языка на другой.

В статье рассмотрены примеры, показывающие недостаточность обычных терминов, изложены результаты наших экспериментов и выявлена специфика пространственных понятий в кыргызском языке.

1. Примеры использования понятий, связанных с направлением

П р и м е р 1. Из-за неопределенности привязки терминов право–лево к направлению иногда добавляется (здесь уже сложнее – либо по отношению к наблюдателю, либо по отношению к наблюдаемому объекту, либо по отношению к движению) слова «справа (слева) по направлению движения».

П р и м е р 2. В языке имеются представления о направлении и движении объекта, как целого, но нет термина, показывающего изменение направления составляющих объекта. Поэтому в Строевых уставах (см. например [2]) специально оговаривается: «Фланг – правая (левая) оконечность строя. При поворотах строя названия флангов не изменяются». (Имеется в виду поворот каждого солдата на своем месте).

2. Представление пространства в языке

Изменение представлений о пространстве и расширение знаний о нем в той или иной науке часто находит свое отражение в изучении пространства в лингвистике.

Пространство – одна из основных форм существования материи, и совершенно естественно, что интерес к его изучению усиливается с изучением искусственного интеллекта.

Последовательность взаимодействия объектов и разновидности их локализации образуют структуру пространства. Каждый из видов пространства находит свое отражение в языке в той или иной форме.

Так, например, эволюционные изменения в понимании пространства в математике (от евклидова пространства к многомерному пространству и топологии [4], а затем и к теории относительности) нашли свое отражение в лингвистических исследованиях, в частности в изменении подходов к трактовке содержания пространственных концептов (геометрический, топологический, функциональный подходы) [5].

Очевидно, что одним из наиболее важных аспектов понимания пространства является представление об отношениях между объектами окружающего мира. Среди разнообразных языковых средств реализации пространственных отношений в русском языке особое место занимает наречие, как обстоятельство места действия. Именно эти лексические единицы реализуют реляционные пространственные концепты и осуществляют лингвистические действия в пространстве.

Нами установлено, что, в отличие от некоторых других языков, в том числе от русского, в кыргызском языке пространственные понятия представляются преимущественно не в виде «отношений» объектов, а в создании виртуальных пространственных областей, основанных на объекте (объектах) в связи как с гравитацией, так и с взаимным расположением и движением объектов и субъекта. Эти области обозначаются существительными, используемыми по общим правилам.

П р и м е р 3. Фраза «Көпөлөк үстөлдүн үстүндө учуп жатат» (дословно – «Бабочка летает в верхнем пространстве стола») допускает адекватный перевод на русский язык: «Бабочка летает над столом». Но фраза «Көпөлөк үстөлдүн үстүнөн учуп кетти» (дословно – «Бабочка улетела из верхнего пространства стола») такого перевода не допускает.

Переводчики в таких случаях делают приблизительный перевод, например «Бабочка улетела от стола». Но в юридическом или техническом тексте такое недопустимо.

3. Виды моделей

Имеет смысл различать модели, которые порождают текст, и модели, которые производят восприятие текста. Первую модель можно назвать порождающей, вторую модель – анализирующей, или аналитической. Оба эти типа моделей можно объединить в один функциональный класс моделей когнитивной или можно сказать, речевой деятельности.

Порождение текста – это создание в тексте некоторого смысла, т. е. переход «рисунок (схема) –> текст». Анализ, или восприятие, текста– это извлечение определенного смысла из данного описания, или переход «текст –> рисунок (схема)». Соответственно модели когнитивной деятельности должны устанавливать соотношение «рисунок (схема) < – > текст». Порождающая и аналитическая модели воспроизводят каждая одну из сторон, отражают когнитивную деятельность в целом.

Модели когнитивной деятельности – не единственно возможный тип лингвистических моделей.

Модель когнитивной деятельности – это один из типов лингвистических моделей в языковой системе.

Чтобы осуществлялась речевая когнитивная деятельность, необходимо наличие языковой системы, т. е. необходимо ее моделирование. При порождении модели необходимый смысл «кодируется»

с использованием элементов и правил, которыми располагает языковая система.

Поэтому моделирование системы языка отражает переход «текст –» языковая система». Этот тип модели можно назвать исследовательским, так как здесь отображается прежде всего деятельность исследователя-лингвиста по выяснению системы языка (см. ниже).

4. Проведение экспериментов

Э к с п е р и м е н т 1. Область применимости глаголов. Рассмотрим дугу окружности на плоскости, изображающую «двор» (по кыргызски – «короо»). При центральном угле, равном 270°, все носители языка согласились с командой «Короого кир!» – «Войди во двор!». При центральном угле, равном 60°, все носители языка не согласились с такой командой, а сказали, что нужна команда « ?га кел!» – «Подойди к (уже не к двору, а к другому объекту, например, к изогнутой стене)».

При промежуточных значениях центрального угла мнения носителей языка расходились, с вариациями, согласно [1].

Э к с п е р и м е н т 2. Требовалось показать точки, соответствующие высказыванию «точка под объектом». Когда объект был прямоугольником, ответы были близкими, с вариациями согласно [1].

Но когда объект был отрезком, наклоненным под 45°, были два принципиально различных вида ответов, соответствующих высказываниям «точка под всеми точками объекта», «точка под какой-нибудь точкой объекта».

При этом понятие «точка под точкой» тоже варьировалось в виде величины угла, направленного вниз от верхней точки, содержащего, по мнению участника эксперимента, все точки «ниже» ее.

Э к с п е р и м е н т 3. Командное соревнование «с дистанционно разделенными членами команды».

Мы провели эксперимент с носителями языка по описанию изображения с геометрическими фигурами с помощью пространственных элементов. Участников разбили по парам. Первый участник пытался описать геометрическое изображение словами, а второй пытался декодировать описание первого участника и создать геометрическую модель. Соревнование показало, как организаторам, так и самим участникам, сложности в адекватном соответствии текста и геометрического изображения.

Заключение

Некоторое из вышеизложенного опубликовано в [6]-[10], другое публикуется впервые. Авторы благодарят студентов, принимавших участие в командных соревнованиях и других экспериментах.

ЛИТЕРАТУРА

1. Zadeh L.A. The concept of a linguistic variable and its application to approximate reasoning // Information Sciences, 1975, Vol. 8, pp. 199–249, 301–357; Vol. 9, pp. 43–80.

2. Голев Н.Д. Об объективности и легитимности источников лингвистической экспертизы // Юрислингвистика – 3: Проблемы юрислингвистической экспертизы: Межвуз. сборник научных трудов. Барнаул, 2002.

3. Строевой устав Вооруженных Сил Российской Федерации, от 11.03.2006.

4. Борубаев А.А., Панков П.С. Компьютерное представление кинематических топологических пространств. – Бишкек: Кыргызский государственный национальный университет, 1999.

5. Бороздина И.С. Категоризация, концептуализация и вербализация пространственных отношений и объектов. – Курск: Курск. гос. ун-т, 2009. – 197 с.

6. Karabaeva S., Dolmatova P. Mathematical and computer models of spatial re-lations in Kyrgyz language // Proceedings of V Congress of the Turkic World Ma-thematicians / Ed. A.Borubaev. – Bishkek: Kyrgyz Mathematical Society, 2014.

7. Karabaeva S. Peculiarities of spatial relations in Kyrgyz language // Abstracts of the Issyk-Kul International Mathematical Forum / Ed. A.Borubaev. – Bishkek: Kyrgyz Mathematical Society, 2015.

8. Pankov P.S., Karabaeva S.J. Mathematical and computer models of spatial concepts in Kyrgyz language // Интернет-журнал ВАК КР, 2016, № 3.

9. Карабаева С. Единый алгоритм словоизменения и представление простран-ства в кыргызском языке. – Saarbrűcken, Deutschland: Lap-Lambert Academic Publishing, 2016.

10. Karabaeva S.J., Pankov P.S. Independent computer presentation of spatial notions in Turkic languages // V Междунар.конф. покомпьютерной обработке тюркских языков «TurkLang 2017». Том 1. – Казань: Издательство АНРТ, 2017.

УДК 378.147 + 519.767.6

EXPERIMENTAL VERIFICATION OF THE E-ASSESSMENT ALGORITHM FOR ANALYSIS OF NATURAL LANGUAGE QA-TEXTS WITHIN THE E-LEARNING SYSTEM

N. A. Prokopyev
Kazan Federal University, Kazan
nikolai.prokopyev@gmail.com

Automated knowledge assessment is an important component of e-Learning systems to ensure effective management of the learning process. As a rule, in modern e-learning systems, the evaluation of the learner's knowledge is carried out through testing - through choosing the right variant from the set of presented answers. It is obvious, that multiple-choice tests can't provide full evaluation of learner's knowledge, thereafter, they do not provide a flexible and qualitative management of training. Thus, it is increasingly important problem to implement an e-Assessment system featuring support of answers that are freely formulated in natural language. An analysis of the most recent publications indicates that automated assessment of natural language QA-texts remains weakly developed in modern e-Learning systems. In this paper an experimental prototype of e-Assessment system featuring automated evaluation of natural language answers is presented. The theoretical model presented by D. Suleymanov in his doctoral thesis is used for develop of answer processor component. An experiment on implemented prototype of the system was conducted to acquire the data needed for further development of algorithm for analysis of natural language QA-texts within the e-Learning system. Experiment results and conclusions are presented.

Keywords: Natural Language processing, NLP, e-Learning, Question-Answer text, QA-text, e-Assessment, automated assessment.

ЭКСПЕРИМЕНТАЛЬНАЯ ПРОВЕРКА АЛГОРИТМА АНАЛИЗА ЕСТЕСТВЕННО-ЯЗЫКОВЫХ ВОПРОСНО-ОТВЕТНЫХ ТЕКСТОВ В СИСТЕМЕ ЭЛЕКТРОННОГО ТЕСТИРОВАНИЯ

Н. А. Прокопьев
Казанский федеральный университет, Казань
nikolai.prokopyev@gmail.com

Автоматизированный контроль знаний является важной составляющей в системах электронного обучения (e-Learning), обеспечивающей эффективное усвоение материала обучаемым. Как правило, в существующих об-

разовательных технологиях при контроле ответа используется тестовый подход, когда испытуемому задаются вопросы с заготовленными вариантами ответа. Очевидно, тесты с вариантами ответа не дают полной оценки знаний обучающегося, соответственно, не способствуют гибкому управлению качеством образования. Поэтому актуальной является задача создания системы контроля ответа, свободно формулируемого самим обучаемым на ЕЯ. Как показывает анализ публикаций, в современных системах электронного обучения (e-Learning) практически не реализована возможность автоматизированного контроля вопросно-ответных текстов на ЕЯ. В статье представлен экспериментальный прототип системы электронного контроля (e-Assessment), предусматривающей автоматическую проверку естественно-языковых ответов обучаемого. В качестве основной модели при реализации модуля проверки ответов используется теоретическая модель, описанная в докторской диссертации Сулейманова Д.Ш. Проведен эксперимент над реализованным прототипом системы с целью получения данных, необходимых для дальнейшего развития алгоритма анализа естественно-языковых ответов в рамках системы контроля знаний. В статье представлены результаты и выводы из данного эксперимента.

Ключевые слова: обработка естественного языка, электронное образование, вопросно-ответные системы, контроль ответа.

Введение

Контроль знаний обучающихся является важным прикладным направлением в образовании в информационную эру, которая характеризуется возрастающим объемом создаваемой человеком информации и активными научными исследованиями во множестве новых междисциплинарных направлений. Как следствие, возникает необходимость компьютеризации процесса обучения и контроля знаний в виде систем электронного образования.

Наиболее актуальное направление компьютеризации таких систем – внедрение алгоритма проверки ответов на вопросы контроля знаний, представленных в виде полноценных предложений на естественном языке. В качестве основного метода анализа естественно-языковых ответов в данной работе используется подход, предложенный Сулеймановым Д. Ш. [1].

Основным методологическим принципом этого подхода является утверждение о том, что заданный вопрос естественным образом ограничивает контекст ответа, как по множеству вариантов ответа, так и по структуре. Из этого следуют принципы реализации, заключающиеся в возможности выделить из ответа семантические структурные единицы, называемые концептулами, и задать контекстную

грамматику правильного ответа как цепочку концептул, чем, соответственно, свести задачу семантического анализа к классической задаче синтаксического разбора. Таким образом, реализуется прагматически-ориентированный подход к анализу естественного языка, то есть реализация алгоритма разбора естественного языка не универсального, а направленного на решение конкретной задачи, ограниченной рамками вопросно-ответного контекста.

1. Постановка задачи и обоснование методологии проведения эксперимента

Для исследования возможностей применения вышеописанного алгоритма естественно-языкового анализа в контексте компьютеризации образования была поставлена задача разработки экспериментального прототипа системы контроля знаний, реализующего этот алгоритм. Данный прототип должен обладать достаточным функционалом для реализации экспериментов над алгоритмом, заключающихся в проведении аутентичного тематического электронного контроля знаний.

Группа участников такого контроля знаний – прежде всего это студенты высших учебных заведений – дает ответы на вопросы по определенной теме, не зная, что они участвуют в эксперименте. Предполагается, что полученные таким образом ответы максимально соответствуют ответам, которые алгоритм должен корректно обрабатывать в полноценно интегрированных в образовательный процесс системах электронного контроля знаний.

Полученные ответы поступают на вход анализатору ответов, а результаты проверки становятся выходными данными эксперимента. В результате такого эксперимента будут собраны данные о том, как студенты отвечают на вопросы, и насколько корректно алгоритм обрабатывает их ответы. Впоследствии, основываясь на результатах таких экспериментов, предполагается сформулировать направление дальнейшего развития алгоритма анализа ответов для внесения в него необходимых изменений. В данной статье описан первый такой эксперимент над первой версией прототипа системы.

Для выработки методологии проведения эксперимента были изучены существующие подходы к экспериментальной проверке эффективности алгоритмов естественно-языкового анализа вопросно-ответных текстов в области электронного образования. Для этого были рассмотрены похожие проекты и находящиеся в разработке си-

стемы по реализации естественно-языкового контроля знаний, для которых была проведена некоторая экспериментальная проверка.

Статья [2] представляет собой отчет о разработке системы автоматической оценки ответов на вопросы из теста GRE (Graduate Record Examinations) по предметам «Биология» и «Психология». Данный тест предназначен для поступающих в магистратуру и аспирантуру в вузах США. Вопросы теста подразумевают ответ, состоящий из 1-3 предложений, который автоматически оценивается по предлагаемому авторами подходу «c-rater». Для проведения эксперимента, оценивающего работу системы, были адаптированы под работу алгоритма 15 вопросов по биологии и 20 вопросов по психологии. Были приглашены 10670 студентов, но только 971 приняли участие. Их работы были не только оценены алгоритмом, но и отдельно каждая работа была проверена двумя экспертами. Авторы утверждают, что, согласно их исследованию, процентное соотношение участников эксперимента по полу и возрасту соответствует общему соотношению при GRE тестировании, однако средняя оценка за тест из эксперимента выше, чем общая средняя оценка. Для исследования работы алгоритма авторы измеряли численное выражение согласия между экспертами, а также согласия каждого из экспертов с оценкой, данной алгоритмом.

Система оценки вопросов с полнотекстовым ответом, описанная в статье [3], использует так называемый авторами статьи подход «нечеткой близости», измеряющий семантическую близость ответа обучающегося и модели эталонного ответа на основе матрицы знаний. Подход сводится к выделению определенных слов из текста ответа, представляющих: логические связи, количественные отношения, утверждения и ключевые слова ответа. На основе сравнения этих наборов слов с моделью эталонного ответа измеряется семантическая близость ответа обучающегося к эталонному, и вычисляется оценка. Для исследования работы алгоритма в системе были созданы три вопроса. Каждый вопрос был представлен отдельной группе студентов, и их ответы были оценены алгоритмом. Кроме того ответы студентов были оценены двумя экспертами, которые не были уведомлены о проведении эксперимента. Далее оценка, данная алгоритмом, сравнивалась авторами исследования со средней оценкой экспертов.

Авторы статьи [4] предлагают систему автоматизированной оценки ответов на вопросы типа «эссе». Ими описан подход, заключающийся в использовании информационного поиска в энциклопе-

дических и вопросно-ответных системах для генерации вопросов и оценки ответов. Оценка происходит путем разделения ответа на предложения, токенизации, автоматической морфологической разметки при помощи пакета OpenNLP и дальнейшего анализа при помощи семантической сети WordNet. Для исследования работы алгоритма был проведен эксперимент в виде тестирования на уроке, в котором приняли участие 7 студентов, давших ответ в виде эссе на 5 вопросов. Для численной оценки использовался коэффициент Пирсона, показывающий корреляцию между оценкой, данной алгоритмом, и оценкой, данной преподавателем.

В статье [5] описана система оценки ответов на полнотекстовые вопросы, сочетающая алгоритмы измерения семантической близости ответа обучающегося и модели ответа с алгоритмом сопоставления графа зависимости слов для учета структуры ответа. Такой подход позволяет авторам применить методы машинного обучения на основе опорных векторов для оценки ответа. Для проверки работы такого подхода были созданы 80 вопросов по предметной области «Информатика», проведен эксперимент, в котором участвовали 30 студентов. Их ответы были независимо оценены двумя экспертами, и часть ответов была использована в качестве обучающего множества для системы. Далее при исследовании работы алгоритма использовались классические методы оценки систем машинного обучения.

В работе [6] дано описание метода оценки вопросов с полнотекстовым ответом. Авторы предлагают свой метод оценки семантической близости, основанный на сравнении ответа обучающегося с эталонным ответом при помощи разбиения ответа на n-граммы и учета синонимии. Всего в рамках исследования было проведено 3 эксперимента, получено 1929 ответов. Для исследования алгоритма использовался коэффициент корреляции Пирсона. Авторы исследовали также работу алгоритма с применением стеммера Портера и провели сравнительный анализ с результатами других алгоритмов, основанных на выделении ключевых слов и на опорных векторах.

2. Реализация прототипа системы контроля знаний и алгоритма анализа ответов

Спецификация разработанного экспериментального прототипа системы наиболее подробно дана в статье [7]. Далее приведено краткое содержание этой спецификации в виде описания компонентов системы и особенности реализации версии прототипа на момент

проведения эксперимента, на рисунке 1 представлена общая диаграмма потоков данных в системе.

Редактор онтологии. Данный компонент предназначен для создания баз знаний предметных областей (отдельных предметов, дисциплин, иных ограниченных связных наборов знаний). Предоставляет инструменты для их заполнения и редактирования. Кроме того, к нему относится генератор вопросов, который позволяет по заполненной базе знаний сгенерировать всевозможные вопросы по встроенным в программу шаблонам и при необходимости скорректировать результаты генерации.

Конструктор тестов. Компонент предназначен для удобного визуального создания ветвящихся тестов для контроля знаний с вероятностным или адаптивным переходом по ветвям дерева теста. Под адаптивным переходом понимается выбор следующего вопроса в зависимости от правильности ответа на текущий вопрос.

Экзаменатор. Основной компонент для прохождения контроля знаний. Реализует алгоритм анализа ответов, переход по дереву теста, сбор результатов, алгоритм адаптации теста. Результаты доступны администратору сервера в виде файлов с логами.

Рис. 1. Диаграмма потоков данных системы

Роли пользователей в системе:

• Администратор – имеет полный доступ к системе и ее базе данных;

• Эксперт – имеет доступ ко всем компонентам;

• Преподаватель – имеет доступ к Конструктору тестов и Экзаменатору;

• Студент – имеет доступ только к Экзаменатору.

После регистрации пользователь не имеет ни одной роли, их назначает администратор.

Алгоритм анализа ответов представлен схемой на рисунке 2. **Лексический процессор** получает на вход ответ, выраженный на естественном языке, и модель соответствия для заданного вопроса. Модель соответствия представляет собой словарь соответствий вида «Концептула (согласно определению из [1])» – «Список лексем, соответствующих концептуле». Лексемы из модели соответствия предварительно приведены к некоторой базовой форме для того, чтобы алгоритм мог обрабатывать различные формы одного и того же слова. В текущей реализации для этого используется стемминг по алгоритму стеммера Портера [8]. Пример модели соответствия приведен далее в главе, посвященной эксперименту. На выходе лексический процессор формирует канонизированное представление ответа, состоящее из цепочки концептул, соответствующих естественно-языковой форме ответа, частичный вектор ситуации, массив запрещенных лексем и массив неопределенных лексем из ответа. Для представления ответа в виде цепочки концептул, лексический процессор производит разбиение ответа на лексемы и стемминг полученных лексем. Пример канонизированного представления приведен далее в главе, посвященной эксперименту. Частичный вектор ситуации – это числовой вектор, содержащий данные о соотношении количества лексем в ответе к количеству лексем, предусмотренных моделью соответствия, о количестве запрещенных лексем и о количестве неопределенных лексем.

На вход **семантического интерпретатора** поступает канонизированное представление ответа и индивидуальная концептуальная грамматика (ИКГ, согласно определению из [1]) для конкретного типа вопроса. Интерпретатор проверяет соответствие цепочки концептул и синтаксического дерева ИКГ и на выходе формирует полный вектор ситуации. Проверка соответствия ответа и ИКГ происходит путем попытки обхода синтаксического дерева ИКГ по узлам согласно канонизированному представлению ответа. Если обход завершается на конечном узле дерева (листе), то считается, что ответ соответствует ИКГ. При обходе пропускаются концептулы LNE,

соответствующие неопределенным лексемам. На выходе семантический интерпретатор формирует полный вектор ситуации – числовой вектор, содержащий, в дополнение к данным частичного вектора ситуации, данные о соответствии канонического представления и ИКГ и о полноте этого соответствия. Вектор ситуации предполагается использовать в дальнейшем для оценки ответа.

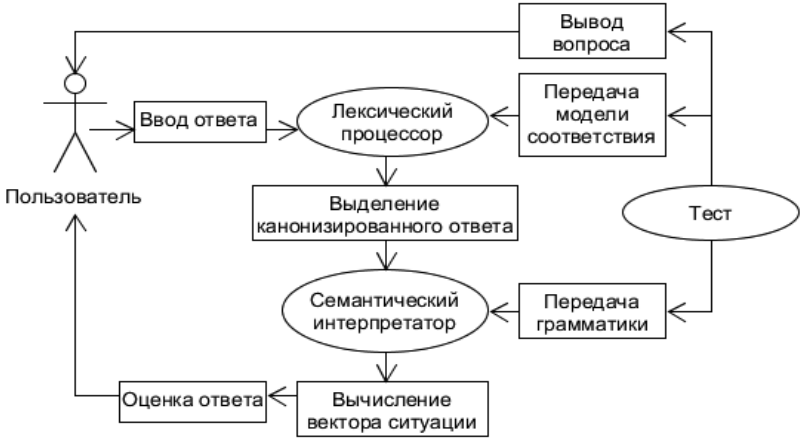


Рис. 2. Алгоритм анализа ответов

3. Эксперимент

Входные данные. Для эксперимента была реализована ИКГ вопросов класса ФУНКЦИЯ, которая в текущей реализации может быть описана в форме Бэкуса-Наура следующим образом:

ИКГ ФУНКЦИЯ ::= [SS* →] ((RA* → (GPP → SP* → SA* | SA* → GPP → SP*) | RP* → (GPA → SA* → SP* | SP* → GPA → SA*))| ((GPP → SP* → RA* → SA* | SA* → RA* → GPP → SP*) | (GPA → SA* → RP* → SP* | SP* → RP* → GPA → SA*)))

Был создан набор вопросов для эксперимента по предметной области «Базы данных», состоящий из 10 вопросов с соответствующими им моделями соответствия. Пример вопроса из данного набора приведен в Таблице 1.Следует отметить, что в вопросах не использовалась часть ИКГ, относящаяся к концептуле SP «понятие-результат».

Таблица 1. **Пример вопроса и его модели соответствия**

Вопрос	Какие функции выполняет СУБД?
SS	субд, систем управлен баз дан
SA	дан во внешн памят, внешн дан, дан в оперативн памят, язык бд, язык баз да, язык sql
RA	управля, работа
GPA	с

Группа участников. В эксперименте приняли участие студенты 3 курса бакалавриата ИВМиИТ КФУ по направлению «Программная инженерия», всего участников было 13. В целях соблюдения аутентичности эксперимента студенты не были проинформированы о нем и должны были проходить вопросно-ответный контроль в полном соответствии с обычными правилами оценки правильности ответов преподавателем. Каждому студенту давалось неограниченное количество попыток на прохождение контроля.

Выходные данные. По итогу эксперимента было получено 228 ответов, в среднем по 23 ответа на вопрос. Большинство ответов не были корректно разобраны алгоритмом анализа ответа. Общий вид выходных данных представлен на примере в Таблице 2.

Таблица 2. **Пример ответов и данных об их разборе алгоритмом**

Какие функции выполняет СУБД?	
Ответ	**Канонизированное представление**
управляет внешней памятью, управляет буферами оперативной памяти, поддержка языков БД	RA: управля, LNE: внешн, LNE: памят, RA: управля, LNE: буфер, LNE: оперативн, LNE: памят, LNE: поддержк, SA: язык бд
Основные функции СУБД управление данными во внешней памяти	LNE: основн, LNE: функц, SS: субд, LNE: управлен, SA: дан во внешн памят
Управляет базой данных.	Ra: управля, LNE: баз, LNE: дан

4. Анализ результатов эксперимента

Большая часть ответов не была корректно разобрана алгоритмом, который, в основном, не распознавал лексемы и на выходе лексиче-

ского анализатора формировал концептулы LNE в тех случаях, когда должен был согласно ожиданиям распознать лексему. По итогам эксперимента выявлены следующие группы ситуаций некорректного функционирования алгоритма и пути их разрешения.

1. Использование отглагольных существительных вместо глаголов и в целом существительных обозначающих отношение типа ДЕЙСТВИЕ. Примеры:

- «администрирование» вместо ожидаемого системой «администрирует»,
- «управление» вместо «управляет»,
- «отмена» вместо «отменяет»,
- «удаление» вместо «удаляет».

Некорректная работа алгоритма в данной группе ситуаций, согласно анализу выходных данных эксперимента, является основным источником нераспознанных лексем в случаях, когда ожидались распознанные лексемы. Вариант разрешения данной группы ситуаций: расширение модели соответствия для включения в нее всех существительных, обозначающих отношение типа ДЕЙСТВИЕ. С этой целью необходимо применить иной подход к стеммингу взамен стеммера Портера, позволяющий более гибко описывать разные словоформы одной и той же семантикой без необходимости перечислять все варианты слова. Такими подходами могут быть морфологическая нормализация либо подход, предложенный в работе [1] для описания модели ответа.

2. Формулировка ответа с использованием вспомогательных слов в сочетании с глаголом либо отглагольным существительным, обозначающим отношение типа ДЕЙСТВИЕ. Примеры:

- «дает возможность удалять»,
- «позволяет сделать выборку»,
- «с помощью этого оператора происходит заполнение»,
- «обеспечивает возможность изменять структуру существующей таблицы».

Разрешение данной группы ситуаций частично связано с подходами к разрешению предыдущей группы, поскольку наличие вспомогательных слов само по себе не приводит к ошибкам в работе алгоритма, к ним приводят нераспознаваемые словоформы глаголов и отглагольных существительных, обозначающих основной предмет вопроса.

3. Особенности пунктуации при ответе при группировке нескольких понятий-аргументов. Примеры:

• «возвращающий набор данных (выборку)»,

• «Обновляет данные/значения»,

• «создание сущности в бд(таблица, функция, схема, пользователь, ...)».

Вариантом разрешения этой группы ситуаций может быть реализация удаления из ответа пунктуационных символов перед его обработкой лексическим процессором, замены их на символ пробела. Однако, стоит заметить, что при этом может нарушаться структура ответа, вследствие чего он может перестать соответствовать ИКГ.

4. Множественный ответ на вопрос, введенный в единственное поле ответа. Примеры:

• «Управление данными, резервное копирование и восстановление»,

• «Хранение/модификация данных, управление транзакциями, поддержка языка SQL».

Причина некорректной работы алгоритма в данной группе ситуаций однозначно заключается в особенностях текущего варианта реализации, которая не учитывает, что ИКГ может быть рекурсивная или комбинированная для учета перечисления ответов. Таким образом, для разрешения этой группы ситуаций требуется доработка системы в соответствии со спецификацией из работы [1].

5. Формулировка ответа в виде определения для объекта вопроса, а не в виде перечисления его функций. Примеры:

• «оператор заполнения таблиц заданными вами значениями»,

• «Оператор DROP является оператором удаления объектов базы данных»,

• «Оператор языка SQL, который применяется для того, чтобы: отменить все изменения, внесенные начиная с момента начала транзакции или с какой-то точки сохранения».

Причина некорректной работы алгоритма в данной группе ситуаций, как и в предыдущей, также заключается в текущем варианте реализации, а именно: не реализованы ИКГ для иных классов вопросов. Согласно [1] ответы в виде определения объекта вопроса соответствуют ИКГ класса ОПИСАНИЕ, который является комбинацией ИКГ класса ФУНКЦИЯ, ИКГ класса ВРЕМЕННОЕ ОТНОШЕНИЕ и ИКГ класса ПРОСТРАНСТВЕННОЕ ОТНОШЕНИЕ. Отсюда следует, что для разрешения этой группы ситуаций требуется реализация данных ИКГ.

6. Использование жаргонной лексики в ответе. Примеры:

• «фиксирует изменения сделанные insert-ом, delete-ом и update-ом»,

• «Добавление, удаление или изменение колонок в уже существующей таблице».

Разрешение данной группы ситуаций заключается в расширении модели соответствия дополнительными лексемами, однако предварительно необходимо рассмотреть вопрос включения жаргонной лексики в качестве корректной при ответе. Возможно, есть необходимость вносить данные об использовании такой лексики в вектор ситуации, как необходимую информацию для оценки ответа.

7. Лишняя информация в ответе. Примеры:

• «Ответ кроется в самом названии. Управляет базой данных»,

• «Возвращает выборку (относится к DML)».

Разрешение данной группы ситуаций некоторым образом пересекается с разрешением группы ситуаций 7: с реализацией удаления лишней информации из ответа перед его обработкой лексическим процессором. Однако это так же может исказить структуру ответа.

8. Ответ содержит орфографические ошибки либо опечатки. Примеры:

• «Управление транакциями»,

• «встака данных в таблицу».

Предлагаемый вариант разрешения данной группы ситуаций: использование автоматического подхода к коррекции грамматических ошибок. В свою очередь, возможно, следует включать данные о наличии орфографических ошибок в вектор ситуации.

Заключение

Дальнейшая разработка алгоритма анализа ответа направлена на решение обозначенных проблем. Полученные в результате эксперимента ответы участников будут повторно использованы для проведения дальнейших экспериментов над алгоритмом в процессе его доработки. Некоторые из предложенных в статье изменений уже применены к прототипу системы: реализован морфологический подход к выделению основы при создании модели соответствия и выделении лексем из ответа, доработана генерация вопросов с учетом этих изменений. На момент публикации готовится новый эксперимент, предназначенный для проверки предположения о том, что данные изменения решают некоторые из обозначенных проблем.

ЛИТЕРАТУРА

1. *Сулейманов Д.Ш.* Системы и информационные технологии обработки естественно-языковых текстов на основе прагматически-ориентированных лингвистических моделей: дис. … д-ра техн. наук. Казань, 2000.

2. *Attali Y., Powers D., Freedman M., Harrison M. Obetz S.* Automated Scoring of Short-Answer Open-Ended GRE Subject Test Items // GRE Board Research Report. 2008. No. GRE-04-02. doi:10.1002/j.2333-8504.2008.tb02106.x

3. *Chakraborty U. Kr., Roy S., Choudhury S.* A Fuzzy Indiscernibility Based Measure of Distance between Semantic Spaces Towards Automatic Evaluation of Free Text Answers // International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems. 2017. Vol. 25. P. 987-1004. doi:10.1142/S021848851750043X.

4. *Dumal P. A. A., Shanika W. K. D., Pathinayake S. A. D., Sandanayake T. C.* Adaptive and automated online assessment evaluation system // 2017 11th International Conference on Software, Knowledge, Information Management and Applications (SKIMA). Malabe, Sri Lanka, 2017. doi: 10.1109/SKIMA.2017.8294135.

5. *Mohler M., Bunescu R., Mihalcea R.* Learning to Grade Short Answer Questions Using Semantic Similarity Measures and Dependency Graph Alignments // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. Vol. 1. Portland, Oregon, 2011. P. 752–762.

6. *Selvi P., Bnerjee A. K.* Automatic Short Answer Grading System (ASAGS) // Arxiv Computing Research Repository. Vol. abs/1011.1742. 2010.

7. *Прокопьев Н.А., Сулейманов Д.Ш.* Автоматизированный анализ естественно-языковых вопросно-ответных текстов в системе электронного тестирования // Пятая Международная конференция по компьютерной обработке тюркских языков «TurkLang 2017» – Труды конференции. Т. 1. Казань, 2017. С. 92–98.

8. *Russian stemming algorithm* [Электронный ресурс] URL: http://snowball.tartarus.org/algorithms/russian/stemmer.html (дата обращения: 29.10.2018).

УДК 351.854

MODERN APPROACHES OF FORMALIZATION OF THE CONCEPT OF ETHICS IN ARTIFICIAL INTELLIGENCE

G. V. Royzenson
Institute for Systems Analysis of Federal Research Center «Computer Science and Control» of RAS, Moscow
Moscow Institute of Physics and Technology (State University) (MIPT), Dolgoprudny, Moscow Region
National Research University «Moscow Power Engineering Institute», Moscow
rgv@isa.ru

In 2016, IEEE launched a global initiative in the field of ethics of artificial intelligence (ethically aligned design). The paper reviewed the draft standards of IEEE ethics AI. The main problems of the formalization of the concept of ethics in AI are analyzed. The main types of risks associated with the introduction of modern technologies using AI into the daily life are considered. A critical analysis of various mathematical tools to formalize the concept of ethics in AI has been carried out. The proposed mechanisms for the successful solution of the problem. Examples of solving practical problems are given.

Keywords: artificial intelligence, ethics, IEEE standards, verbal decision analysis.

СОВРЕМЕННЫЕ ПОДХОДЫ ФОРМАЛИЗАЦИИ ПОНЯТИЯ ЭТИКИ В ИСКУССТВЕННОМ ИНТЕЛЛЕКТЕ

Г. В. Ройзензон
Институт системного анализа ФИЦ ИУ РАН, г. Москва
МФТИ, г. Долгопрудный, Московская область
МЭИ, г. Москва
rgv@isa.ru

В 2016 IEEE выступила с глобальной инициативой в области этики искусственного интеллекта (этически обусловленное проектирование). В работе рассмотрены проекты стандартов IEEE этики ИИ. Проанализированы основные проблемы формализации понятия этики в ИИ. Рассмотрены основные типы рисков, связанные с внедрением в повседневную жизнь современных технологий, использующих ИИ. Проведен критический анализ

различного математического инструментария, позволяющего формализовать понятие этики в ИИ. Предложены механизмы для успешного решения поставленной задачи. Приведены примеры решения практических задач.

Ключевые слова: искусственный интеллект, этика, стандарты IEEE, вербальный анализ решений.

Введение

Бурные темпы цифровизации самых разных сфер деятельности человека во многом обусловлены широким внедрением технологий, использующих искусственный интеллект (ИИ). К подобным технологиям можно отнести интеллектуальные биржевые роботы, системы оценки кредитного риска [20], беспилотные автомобили и летательные аппараты [1], различные технологии биометрической идентификации (распознавание лиц, отпечатков пальцев и др.) и т. п. Подобное развитие и внедрение технологий ИИ неразрывно связано с появлением совершенно новых типов рисков.

К концу следующего десятилетия можно ожидать лавинообразный рост (несколько десятков миллиардов единиц) числа разнообразных интеллектуальных устройств (ИУ) разного масштаба (интеллектуальных роботов (ИР), умных машин, умных предприятий, умных городов и т. п.) [1]. Соответственно, подобное развитие событий приведет к необходимости обработки огромных массивов информации, поступающих от ИУ. В этой связи исследователи столкнутся с проблемой обработки больших данных (big data) совершенно другого масштаба, даже по сравнению с настоящим моментом. Под большими данными понимаются такие данные, объем которых превосходит текущие возможности оперирования ими в обозримый период. Важным направлением исследований является использование больших данных в рамках интерактивных компьютерных систем [17]. Итак, можно предложить альтернативное определение больших данных, формулируемое следующим образом: под большими данными понимаются массивы разнородной (структурируемой, слабо структурируемой и неструктурируемой) информации, которые не могут быть непосредственно использованы в человеко-машинных процедурах многокритериального принятия решений. С учетом вышеизложенного, рост числа ИУ требует разработки совершенно новых подходов к оценке технологических рисков [11, 36], в которых будут гармонично сочетаться возможности интеллектуальных устройств самостоятельно предотвращать какие-либо нежела-

тельные последствия (аварии) и возможности человека вмешиваться в такие процессы. Очевидно, что роль человеческого фактора при оценке технологических рисков (например, статистические и экспертные подходы) будет постепенно сокращаться (человек просто физически не успеет среагировать на различные опасные ситуации, которые могут возникать при использовании ИУ). Подобное развитие событий потребует проектирования ситуационных центров (далее СЦ) нового поколения для отслеживания состояния огромного числа ИУ для возможности оперативного вмешательства в их работу [5]. Таким образом, ускорение темпа развития информационных технологий ставит новые задачи и предоставляет новые возможности использования СЦ для решения самого широкого круга вопросов мониторинга различных сфер деятельности.

С учетом вышеизложенного в 2016 году структура IEEE (Institute of Electrical and Electronics Engineers – Институт инженеров электротехники и электроники), параллельно с еще несколькими организациями (например, ЮНЕСКО), выступила с глобальной инициативой в области этики искусственного интеллекта (далее, ИИ) [13, 26, 27, 40]. Важность предпринятых IEEE усилий определяется ее фактически ведущей ролью в сообществе ученых и инженеров в электротехнике, электронике, информационных технологиях, телекоммуникации и т. д., что обязательно окажет самое непосредственное влияние на разработчиков технологий ИИ. В результате усилий IEEE был разработан документ «Ethically Aligned Design» («Этически обусловленное проектирование») [40]. В документе IEEE отражены основные актуальные угрозы и риски [4], связанные с внедрением автономных систем на базе ИИ.

В соответствии с работами [4, 16] под измерением риска понимают определение опасности от того или иного источника (вида деятельности) для индивидуума или группы. Отметим основные четыре подхода, которые используются для измерения риска. Первый подход хорошо известен как инженерный. В рамках данного подхода основные усилия направлены на сбор статистических данных о поломках, авариях, связанных с утечкой вредных веществ в окружающую среду [11]. Инженерный подход ориентирован на количественный расчет вероятности поломок, отказов и других нежелательных событий. Второй подход принято называть модельным. Данный подход предполагает моделирование процессов, которые могут спровоцировать различные нежелательные последствия (аварии и т.п.) [11]. Третий подход к измерению риска известен как экс-

пертный [4, 16, 20]. Такой подход к оценке риска применяют когда возникают определенные сложности при использовании инженерного и модельного подходов. Четвертый подход измерения риска, известен как социологический [3, 16, 32]. В рамках данного подхода предполагается измерить восприятие населением и его отдельными группами того или иного риска.

Рассмотрим несколько небольших примеров, связанных с оценкой риска при внедрении в повседневную жизнь различных технологий ИИ. По причине перераспределения рабочей силы на постсоветском пространстве после распада СССР в Москве сосредоточено несколько сотен тысяч водителей автотранспорта (таксисты, водители маршруток, троллейбусов, автобусов и т. п.). Соответственно, если использование беспилотных автомобилей будет набирать современные темпы, то буквально через несколько лет значительная часть водителей – выходцев из стран ближнего зарубежья, останется без работы. Большинство этих людей другими профессиями быстро овладеть не смогут. Можно предположить, что для такого мегаполиса как Москва (как, впрочем, и для любого другого) такое развитие ситуации может легко привести к социальному взрыву, если предварительно не проделать глубокий анализ всех возможных последствий от внедрения тех или иных технологий ИИ (фактически нужно оценить социально-экономические риски).

Рассмотрим еще пример с моногородом (а это город, как хорошо известно, который сформировался и развился вокруг какого-то крупного промышленного предприятия). Предположим, на этом предприятии работает 20 000 тыс. человек. При этом используется технология, которая связана с достаточно низкой (по современным меркам) производительностью труда, относительно высоким процентом брака готовой продукции и наносит достаточно существенный экологический ущерб. Есть новая технология, использующая ИИ, которая сразу решает большинство перечисленных проблем (рост производительности труда, уменьшение экологического ущерба и т. п.) и внедрение которой может снова сделать этот промышленный гигант конкурентным на мировом рынке. Но внедрение этой новой технологии предполагает увольнение с предприятия 18 000 тыс. человек, что также грозит моногороду социальными потрясениями, т. к. других предприятий, в которые можно было бы устроить работников по специальности, в городе попросту нет. Что делать в этой ситуации? Если, вообще ничего не менять, то предприятие разорится, т. к. станет совершенно не конкурентоспособным и вообще

все 20 000 тыс. людей останутся без работы. Если внедрить самую передовую технологию, использующую ИИ, то тоже значительную часть сотрудников нужно будет уволить, но предприятие сохранится и в городе можно будет, например, улучшить экологию, а на вырученные от поступления в бюджет региона налоговые средства организовать новую программу по переподготовке кадров и т. п.

Поэтому, важность внедрения стандартов этики ИИ IEEE состоит в том, что очевидно эта проблема является междисциплинарной. При ее решении напрямую затрагиваются не только различные чисто этические проблемы, но и стоит задуматься о различных социально – экономических последствиях (рисках) от внедрения в нашу повседневную жизнь различных технологий, использующих ИИ [28].

В июле 2017 года была образована Российская рабочая группа IEEE по вопросам этики ИИ в составе: Готовцев П. М. (руководитель), Карпов В. Э., Овсянникова Е. Е. (секретарь) и Ройзензон Г. В. [27]. Основные цели группы:

- представлять мнения и предложения российских ученых, участвующих в работах над документами IEEE: Ethically Alligned Design;
- информировать российских ученых о результатах деятельности рабочей группы IEEE по созданию документа IEEE: Ethically Alligned Design;
- привлекать российских ученых к исследованиям по тематике «Этика систем искусственного интеллекта».

В рамках инициативы IEEE предполагается разработка серии стандартов (см. табл. 1), применение которых, скорее всего, станет обязательным для всех специалистов и организаций, занятых в создании различных продуктов в той или иной степени использующих технологии ИИ. Кроме того, со своей стороны, организации и специалисты, использующие технологии ИИ, должны в самое ближайшее время сформулировать и внести свои предложения по выработке дополнительных условий соответствия упомянутым стандартам. Таким образом, в процедуре выработки стандартов должны принять участие несколько сторон (например, IEEE, разработчики и научное сообщество).

Одна из целей данной работы – обзор существующего математического инструментария, который может быть использован для формализации понятия этики в ИИ. В частности, особый интерес представляют подходы, позволяющие оценивать те или иные техно-

логии, использующие ИИ, на соответствие определенным требованиям (этическим нормам, критериям, стандартам и т. п.).

Прежде чем рассматривать конкретные математические подходы для решения поставленной задачи, важно проанализировать сформулированные к настоящему моменты основные определения для используемых терминов.

Этика – философская дисциплина, исследующая вопросы морали и нравственности.

Относительно определения научного направления ИИ дело обстоит несколько сложнее. Нужно признать, что какого-то одного устоявшегося и единодушно принятого научным сообществом определения к настоящему моменту не выработано. Разработано огромное количество различных определений ИИ, сравнение и анализ которых явно выходит за рамки представленной работы. По мнению автора настоящей статьи под ИИ понимается группа методов и подходов, которые ориентированы на решение слабоструктурированных задач. Разумеется, представленное определение не претендует на истину в последней инстанции.

Прежде всего, нужно упомянуть классические фундаментальные работы по этике как отечественных, так и зарубежных специалистов, которые оказали существенное влияние на современные представления. В частности, можно отметить работы: Апресяна Р. Г. [2]; Гусейнова А. А. [6]; Дробницкого О. Г. [9]; Кропоткина П. А. [14]; Фролова И. Т. и Юдина Б. Г. [34]; Швейцера А. [37]; Шпемана Р. [38] и ряда других.

В истории науки можно привести несколько ярких примеров того как развитие новых технологий и вопросы этики приводили к весьма существенным противоречиям и столкновениям мнений политиков, общественных организаций, ведущих мировых ученых и т. п. При этом важно обратить внимание на то, что в вопросах этики в рамках научно-технического прогресса инициатива исходила зачастую именно от самих ученых.

В 1955 году Б. Рассел озвучил одну из первых таких инициатив (знаменитый манифест Рассела-Эйнштейна). Инициатива Б. Рассела положила начало широко известному теперь Пагуошскому движению за мир и разоружение [34], поскольку к середине 50-х годов угроза всеобщей ядерной катастрофы стала очевидной и появилась необходимость мобилизовать авторитетнейших ученых (Ф. Жолио-Кюри, Л. Полинг, А. Эйнштейн и др.) для преодоления сложившейся критической ситуации. В результате, правда не сразу, были пре-

кращены воздушные, наземные и подводные испытания ядерного оружия [7].

Еще одним таким ярким примером может служить инициатива американского генетика П. Берга, который в 1974 году предложил наложить мораторий на эксперименты с рекомбинантной ДНК для того чтобы иметь возможность оценить все риски и последствия использования этой новой технологии [34].

Основные выводы, которые можно сделать из подобных инициатив ведущих мировых ученых, заключаются в том, что развитие и неразумное использование тех или иных технологий может угрожать существованию человечества (например, ядерное оружие) [7]. Безусловно, повсеместное внедрение различных технологий ИИ также сопряжено с определенными опасностями и рисками [4], что во многом и стало причиной инициативы по разработке проектов специальных стандартов в области этики ИИ со стороны IEEE (см. Табл. 1).

Еще один важный вывод, который можно здесь сделать, заключается в том, что, с одной стороны, развитие тех или иных опасных технологий без учета вопросов этики может привести человечество к совершенно катастрофическому результату. С другой стороны, если «загнать» развитие современных технологий в слишком жесткие рамки, это станет причиной замедления темпов научно-технического прогресса (очень характерный пример с уже упомянутыми технологиями генной инженерии [34]).

Кроме того, есть еще один важный аспект, связанный с инициативой IEEE по разработке стандартов. К настоящему моменту на рынке появилось огромное количество продуктов, которые претендуют на то, что в них, в той или иной степени, используются технологии ИИ. В действительности при тщательном изучении оказывается, что представленные продукты вообще не имеют никакого отношения к рассматриваемой предметной области. Таким образом, предстоящая возможная достаточно жесткая сертификация продуктов, использующих технологии ИИ, заставит многих недобросовестных производителей задуматься, прежде чем осуществлять различные маркетинговые действия (иными словами голословно заявлять, что в их продукции используются технологии ИИ). Соответственно разработка рассматриваемых стандартов может послужить определенным фильтром, т. к. многие производители продукции много раз подумают, стоит ли им заявлять, что их продукция использует технологии ИИ, зная, что им предстоит обязательная непростая процедура сертификации.

Таблица 1. **Проекты стандартов IEEE**

Код проекта стандарта IEEE	Оригинальное название	Перевод
P7000	Model Process for Addressing Ethical Concerns During System Design	Проект стандарта модельного процесса для решения этических проблем при проектировании систем
P7001	Transparency of Autonomous Systems	Проект стандарта прозрачности автономных систем (АС)
P7002	Data Privacy Process	Проект стандарта обеспечения конфиденциальности данных
P7003	Algorithmic Bias Considerations	Проект стандарта учета необъективности алгоритма
P7004	Standard for Child and Student Data Governance	Проект стандарта управления данными детей и студентов
P7005	Standard for Transparent Employer Data Governance	Проект стандарта для прозрачного управления данными работодателя
P7006	Standard for Personal Data Artificial Intelligence (AI) Agent Standard for Personal Data Artificial Intelligence (AI) Agent	Проект стандарта для интеллектуального агента управления персональными данными
P7007	Ontological Standard for Ethically Driven Robotics and Automation Systems	Проект онтологического стандарта робототехники и систем автоматизации, управляемых на основе этики
P7008	Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems	Проект стандарта для учета этических принципов в робототехнике, интеллектуальных и автоматизированных системах
P7009	Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems	Проект стандарта проектирования безотказных автономных и полуавтономных систем
P7010	Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems	Проект стандарта метрик благосостояния для систем искусственного интеллекта и автономных систем, действующих на основе этических принципов

Продолжение таблицы 1

P7011	Standard for the Process of Identifying and Rating the Trustworthiness of News Sources	Проект стандарта определения и оценки надежности новостных источников
P7012	Standard for Machine Readable Personal Privacy Terms	Проект стандарта для машиночитаемых правил конфиденциальности личных данных
P7013	Inclusion and Application Standards for Automated Facial Analysis Technology	Проект стандарта включения и применения технологии автоматизированного анализа состояния лица

Анализ представленных классических работ по этике, а также некоторые выводы авторов работы [13] позволяют констатировать, что вопросы соотношения этики и ИИ коренным образом отличаются от того, что понимается, например, под этическими проблемами генных технологий, информатики, естествознания и т. п. Это отличие определяется тем, что в ИИ этические вопросы ближе к пониманию этики в философском или социогуманитарном смысле, и связаны эти этические аспекты прежде всего с тем, что они касаются вопросов поведения и принятия решений. Соответственно данный аспект существенно влияет на выбор математического инструментария для формализации понятия этики в ИИ.

1. Краткое описание проектов стандартов ИИ

Приведем более подробное описание целей проектов стандартов этики ИИ, представленных в табл. 1.

P7000. Инженеры, технологи и другие заинтересованные стороны проекта нуждаются в методологии для выявления, анализа и согласования этических проблем конечных пользователей с самого начала жизненного цикла систем и программного обеспечения. Цель этого стандарта – обеспечить практическое применение такого типа методологии проектирования на основе ценностей, которая демонстрирует, что концептуальный анализ ценностей и обширный анализ возможности их соблюдения могут помочь в улучшении требований этической составляющей в системах и жизненных циклах программного обеспечения. Этот стандарт предоставит инженерам и технологам полезный инструмент, выстраивающий процессы управления инновациями, подходы к разработке ИС и методы разра-

ботки программного обеспечения с целью минимизации этического риска для организаций, заинтересованных сторон и конечных пользователей.

P7001. Ключевой проблемой автономных систем (АС) является то, что их работа должна быть прозрачной для широкого круга заинтересованных сторон по ряду причин. (i) Прозрачность важна для пользователей, поскольку она формирует доверие к системе, предоставляя простой способ понять, что и почему делает система. Если взять в качестве примера робота, ухаживающего за пожилыми и больными людьми, то прозрачность означает, что пользователь может быстро понять, что может делать робот в разных обстоятельствах или, если робот должен сделать что-то неожиданное, пользователь должен иметь возможность спросить робота «почему это только что было сделано»? (ii) Валидация и сертификация прозрачности АС важны, поскольку они раскрывает процессы, происходящие в системе, для проведения проверки. (iii) Если происходят несчастные случаи, АС должна быть прозрачной при расследовании аварии; внутренний процесс, который привел к аварии, должен легко прослеживаться. После несчастного случая (iv) адвокатам или другим свидетелям-экспертам, которым может потребоваться предоставление доказательств, прозрачность необходима для демонстрации своих доказательств. Наконец, (v) для революционных технологий, таких как беспилотные автомобили, необходим определенный уровень прозрачности для более широких кругов общества, чтобы повысить доверие общественности к технологии. Для дизайнеров стандарт предоставит руководство для собственной оценки прозрачности в процессе разработки и даст механизмы для повышения прозрачности (например, необходимость защищенного хранения данных датчиков и данных о внутреннем состоянии аналогично регистратору данных полета или «черному ящику»).

P7002. Цель этого стандарта – сформировать единый общий методологический подход, который определяет методы управления вопросами конфиденциальности в процессах жизненного цикла систем / программ.

P7003. Этот стандарт предназначен для того, чтобы обеспечить отдельным лицам или организациям, создающим алгоритмы, в основном предназначенные для автономных или интеллектуальных систем, ориентированных на сертификацию методологий, четко сформулированную отчетность и ясность в отношении того, как и с какой целью работают алгоритмы, как происходит оценка и как

это влияет на пользователей и заинтересованные стороны, использующие указанный алгоритм. Сертификация в соответствии с этим стандартом позволит разработчикам алгоритмов сообщать пользователям и регулирующим органам, что в разработке, тестировании и оценке алгоритма использовались современные передовые методы, что, в свою очередь, позволит избежать необоснованных дифференцированных последствий для пользователей.

P7004. Этот стандарт предназначен для того, чтобы предоставить организациям, работающим с данными о детях и учащихся, процессы, ориентированные на управление, и сертификацию, гарантирующие прозрачность и подотчетность их действий в том, что касается безопасности и благополучия детей, их родителей, учебных заведений, в которых дети зарегистрированы, и сообщества и общества, где они проводят время, как в автономном режиме, так и в режиме онлайн. Стандарт также разработан для помощи родителям и преподавателям в понимании того факта, что большинство людей могут быть недостаточно технически грамотны, чтобы знать основные проблемы использования данных, но при этом должны быть соответствующим образом проинформированы о защищенности данных о своих детях, и иметь в своем распоряжении инструменты и услуги, которые обеспечивают надлежащие возможности для выбора данных о семье на основе контента и полученной ранее информации.

P7005. Этот стандарт предназначен для предоставления организациям набора четких рекомендаций и сертификатов, гарантирующих, что они хранят, защищают и используют данные сотрудников этическим и прозрачным способом. Он также предназначен для того, чтобы помочь работодателям понять, что большинство людей могут не обладать достаточными техническими навыками, чтобы понимать основные проблемы использования данных, но, тем не менее, должны быть соответствующим образом проинформированы о безопасности данных своих сотрудников, и иметь в своем распоряжении инструменты и услуги, которые обеспечивают надлежащие возможности для выбора информации, которой они обмениваются на рабочем месте, на основе контента и ранее полученной информации. Этот стандарт, разработанный по модели законодательства ЕС относительно правил защиты персональных данных GDPR (General Data Protection Regulation), будет сформулирован как разновидность «GDPR для работников». Он будет гарантировать, что работники, сталкивающиеся с широко распространенными проблемами автоматизации, связанными с потенциальной потерей рабочих мест, сохра-

нят контроль и управление своей персональной информацией, которая напрямую является базовым активом их идентичности и жизни, вне зависимости от того, получена ли она в результате мониторинга производственного процесса или из хранилища персональных данных.

Р7006. С появлением ИИ и его развитием существует риск того, что решения, принимаемые в результаты общения машин друг с другом, будут делаться на основе входных данных, имеющих характер «черного ящика» без прозрачности этих входных данных для людей. Для создания ИИ на основе этических принципов индивидуумам потребуются средства для управления и определения ценностей, правил и входной информации, которые направляли бы развитие персонализированных алгоритмов и ИИ. Им понадобится агент, который сможет представлять их индивидуальные права и возможности в системе общих социальных норм, этики и прав человека и который к тому же будет предвидеть этические последствия обработки данных и помогать их смягчить. Такой подход позволит людям безопасно организовывать и распространять свою личную информацию на машиночитаемом уровне и даст возможность персонализированному ИИ действовать в качестве их представителя при принятии решений машинами. Важнейшей целью создания этого стандарта является информирование государственных и коммерческих субъектов о том, что в их интересах предпочтительнее создание механизмов, предназначенных для индивидуальных пользователей, для обучения персональных агентов ИИ для преодоления этой асимметрии и гармонизации использования персональных данных в будущем.

Р7007. Стандарт устанавливает набор определений и их взаимосвязей, которые обеспечат развитие робототехники и систем автоматизации в соответствии с мировыми теориями в области этики и морали с упором на приведение инженерных сообществ к пониманию того, как рационально разрабатывать такие системы и гармонично их внедрять. Эти определения обеспечат точное взаимопонимание экспертов в различных областях, включая робототехнику, автоматизацию и этику. Использование онтологий для представления знаний в любой области имеет несколько преимуществ, которые включают: а) формальное определение понятий конкретного домена в представлении, независимом от языка, то есть вне зависимости от конкретного языка программирования, однако с формальным описанием для реализации на целевом языке; б) инструменты для анализа понятий

и их взаимоотношений в поисках несогласованности, неполноты и избыточности; в) язык для использования в процессе коммуникации между роботами от разных производителей и т. п.

P7008. Стандарт для этически направленных роботизированных, интеллектуальных и автономных систем устанавливает набор определений и функций и взаимосвязей между ними с преимуществами, зависящими от культурных аспектов пользователей (благосостояние, здоровье и т. п.), что позволит развивать робототехнику, интеллектуальные и автономные системы в соответствии с мировыми теориями этики и морали с упором на приведение инженерных сообществ к пониманию того, как рационально разрабатывать такие системы и гармонично их внедрять. Этот стандарт вместе с определениями обеспечит точное взаимопонимание мировых экспертов в различных областях, включая робототехнику, автоматизацию и этику.

P7009. Этот стандарт устанавливает практическую техническую основу конкретных методологий и инструментов для разработки, внедрения и использования эффективных надежных механизмов в автономных и полуавтономных системах. Стандарт включает (но не ограничивается этим): четкие процедуры измерения, тестирования и сертификации способности системы надежно работать в интервале от слабых до сильных нагрузок, а также инструкции по ее усовершенствованию в случае неудовлетворительной работы. Стандарт служит основой для разработчиков, а также для пользователей и регуляторных органов в части создания устойчивых к отказам механизмов надежным, прозрачным и ответственным образом.

P7010. Стандарт метрики благополучия для этического искусственного интеллекта и автономных систем позволяет программистам, инженерам и технологам эффективнее анализировать, как именно продукты и услуги, которые они создают, могут повысить благосостояние людей, на основе более широкого набора показателей, нежели только рост и производительность. Сегодня системы с датчиками распознавания эмоций количественно оцениваются в первую очередь по своей экономической ценности на рынке за пределами сферы их применимости в определенных областях (психология и т. д.). Хотя часто понимается, что этические соображения для интеллектуальных и автономных систем могут препятствовать инновациям из-за введения нежелательного регулирования, тем не менее, без метрик, оценивающих психическое и эмоциональное здоровье, как на уровне индивидуума, так и общества, инновации невозможно количественно оценить. Внедрение и использование этих

показателей для программистов и технологов означает, что благосостояние человека – помимо экономического роста – может быть измерено и улучшено.

P7011. Целью стандарта является устранение негативных последствий неконтролируемого распространения ложных новостей путем предоставления открытой системы легких для понимания рейтингов. При этом стандарт должен помогать в восстановлении доверия к определенным поставщикам новостей, и надлежащим образом дискредитировать других, подавлять стремление к публикации ложных новостей и пропагандировать пути улучшения для провайдеров, стремящихся к этому. Целью стандарта является репрезентативный образец набора новостных сообщений для формирования осмысленной и точной шкалы рейтингов.

P7012. Цель стандарта – предоставить людям средства для выставления своих собственных условий, касающихся личной конфиденциальности, способами, которые могут быть прочитаны, признаны и подтверждены машинами, которые управляются из сетевого мира. В более формальном смысле цель стандарта заключается в том, чтобы дать индивидуумам возможность быть главными в соглашениях с другими сторонами – в основном компаниями – выступающими в качестве вторых сторон. Следует обратить внимание, что целью настоящего стандарта является не рассмотрение политик конфиденциальности, поскольку они предполагают наличие только одной стороны и не требуют наличия соглашения. Для соблюдения условий необходимо наличие соглашения, а для соблюдения политики конфиденциальности оно не требуется.

P7013. Исследования продолжают демонстрировать, что ИИ, который используется для автоматизированного анализа лиц, подвержен тенденциозности, которая может усугублять человеческие предрассудки и систематически дискриминировать людей по признаку пола, этнической принадлежности, возраста и других факторов. Цель стандарта – предоставить руководящие принципы для разработки и сравнительной оценки автоматизированной технологии анализа лица для смягчения демографической и фенотипической предвзятости и дискриминации. Разделы и протоколы отчета, прописанные в настоящем стандарте, служат для повышения прозрачности этой автоматизированной технологии с тем, чтобы разработчики и лица, принимающие решения, могли сравнивать доступные варианты для выбора наиболее подходящей технологии на основе целевых групп населения и предполагаемых вариантов использо-

вания. Учитывая значимость биометрических данных, получаемых из анализа человеческого лица, стандарт также определяет в общих чертах надлежащее и ненадлежащее использование автоматизированного лицевого анализа на основе точности ценностей, установленных мировым сообществом.

2. Математические методы формализации понятия этики в ИИ

Проблема формализации этических норм включает в себя две основные задачи. Первая – это создание форм представлений норм, вторая – выбор соответствующего математического аппарата для работы с этими формами: сопоставления, измерения, анализа и т. д. Нечеткая, многозначная или вероятностная логика – это достаточно глубоко проработанные области, доведенные, вообще говоря, до уровня практически применимых технологий. Здесь гораздо важнее определиться с качественным уровнем представления параметров систем ИИ и этических норм. Важно отметить, что проблема формализации этических норм тесно связана с более общей задачей, а именно: с формализацией гуманитарного знания [32, 33].

Кроме того, в рамках этики ИИ требуется разработка новых норм, таких, например, как гуманность (как машины влияют на наше поведение и взаимодействие), сингулярность (как мы сможем контролировать сложную «умную» систему), безопасность и т. п. Таким образом, не всегда соответствие тем или иным нормам можно свести к классическим «да» и «нет». Поэтому здесь актуально рассмотрение и использование различных неклассических логик (например, многозначных [8, 12, 31, 32], темпоральных [10]), механизма многокритериальной классификации [17, 20], нечеткой логики и вероятностных подходов [31, 39, 41] и т. п. Рассмотрим указанный математический аппарат подробнее.

2.1. Булева алгебра

Концепция формализации различных этических понятий активно развивается на протяжении последних десятилетий. В качестве пионерской работы по исследуемому вопросу важно упомянуть книгу В. А. Лефевра «Алгебра совести» [19]. В этой книге есть целая глава, которая посвящена вопросам этики и возможным аспектам, связанным с формализацией этого понятия. Для решения рассматриваемой задачи в основном используется математический аппарат

булевой алгебры. Это имеет как положительные, так и определенные отрицательные стороны. К положительным можно отнести то, что булева алгебра к настоящему моменту очень хорошо развита, есть множество приложений, программных библиотек для самых разных инструментальных средств и т. п. К отрицательным можно отнести то, что не всегда различные этические проблемы (в том числе и относящиеся к ИИ) можно строго разделить на «белые» и «черные» [23], а механизм булевой алгебры зачастую предполагает именно такой подход. В работе Д. А. Поспелова [23] для преодоления данной проблемы вводится понятие «кольцевых» шкал, что для решения задачи формализации этики ИИ является весьма оригинальным и перспективным подходом.

2.2. Многозначные логики

В рамках развития различных неклассических парадигм в ИИ, в частности, подхода многозначных логик важно упомянуть работы отечественных специалистов, в частности А. С. Карпенко [12], В. К. Финна [32], О. П. Кузнецова [15], В. Б. Тарасова [31], В. Н. Вагина [8] и др. Использование многозначных логик для формализации понятия этики ИИ также сопряжено с определенными сложностями. В частности, переход от трехзначной логики к четырехзначной может потребовать кардинальной «переделки» соответствующих математических конструкций, что фактически означает необходимость решения указанной задачи заново.

2.3. Теория вероятностей и нечеткая логика

В этом смысле использование вероятностного аппарата и нечеткой логики [41] для формализации понятий этики ИИ [39] является весьма интересным подходом, так как нечеткую логику можно считать неким обобщением многозначной логики [31]. К известным особенностям применения нечеткой логики можно отнести определенные проблемы при построении функций принадлежности: разные способы построения таких функций приводят к разным результатам (неустойчивость методов нечеткой логики относительно исходных данных).

2.4. Вербальный анализ решений

Еще одним возможным подходом для формализации понятия этики ИИ является использование методов вербального анализа решений (ВАР) [17].

Группа методов ВАР опирается на использование достижений различных научных дисциплин: когнитивной психологии (измерения, операции получения информации при построении решающего правила, поэтапное построение решающего правила); прикладной математики (обоснование вида решающего правила, методов получения и проверки информации на непротиворечивость); теории организаций (получение объяснений); компьютерных наук (диалог человек-компьютер). Разработанные в рамках этого подхода методы принятия решений позволяют при анализе вариантов сложных решений органично сочетать качественную и количественную информацию об альтернативах, суждения экспертов и предпочтения лиц, принимающих решения, объективные и субъективные факторы, характерные для проблемной ситуации.

Например, применительно к обозначенной проблеме формализации этики в ИИ, с использованием методов ВАР, возможна следующая постановка задачи. Если в рамках этики ИИ разработать некоторый перечень норм, то степень соответствия той или иной норме можно рассматривать как задачу многокритериальной порядковой классификации [17, 20]. Соответственно, на основе анализа таких норм этики ИИ мы должны будем принять решение о том, что либо нормы соблюдены, либо есть некоторое несущественное их нарушение, либо наблюдается какой-то заметный отход от принятых норм и т. п. То есть нам будет нужно отнести определенную совокупность оценок по каждой из норм к некоторому классу решений (категории).

К положительным сторонам использования методов ВАР прежде всего можно отнести то, что к исходным данным не применяются никакие операции по их переводу в количественную форму. Известно, что перевод вербальных измерений в «цифру» зачастую весьма субъективен и не имеет строгого математического обоснования. Кроме того, методы ВАР позволяют получить объяснения принятых решений (интерпретация результата) в терминах предметной области, здесь – в терминах описания норм этики ИИ. В качестве недостатков методов ВАР можно отметить большие трудозатраты эксперта или лица, принимающего решения (ЛПР), при работе в признаковом пространстве большой размерности. В этом случае необходимо применять различные методы снижения его размерности [24, 25].

Таким образом, можно констатировать, что к настоящему моменту разработан широкий спектр инструментальных средств, опи-

рающихся на самые различные математические конструкции, позволяющие успешно решить задачу формализации этических норм в ИИ.

3. Примеры

Рассмотрим более подробно применение метода ВАР ЦИКЛ, предназначенного для построения многокритериальной порядковой классификации с привлечением ЛПР или эксперта, применительно к проблеме формализации понятия этики ИИ. При этом будем использовать следующие проекты стандартов этики ИИ IEEE: P7013 (технологии автоматизированного анализа состояния лица) и P7011 (оценка надежности новостных источников).

3.1. Постановка задачи многокритериальной порядковой классификации

Рассмотрим теперь формальную постановку задачи многокритериальной порядковой классификации. Дано [17, 20]:

- G – свойство, отвечающее целевому критерию задачи (критичность неисправности, наличие и степень тяжести заболевания, ценность кредитного проекта, степень интеллектуальности робототехнической системы и т. д.).
- $K = \{K_1, K_2, \ldots, K_N\}$ – множество критериев, по которым оценивается каждый объект.
- $S_q = \{k_1^q, k_2^q, \ldots, k_{\omega_q}^q\}$ для $q = 1 \ldots N$ – множество оценок по критерию K_q; ω_q – число градаций на шкале критерия K_q; оценки в S_q упорядочены по возрастанию характерности для свойства G.
- $Y = S_1 \times S_2 \times \ldots \times S_N$ – пространство состояний объектов, подлежащих классификации. Каждый объект описывается набором оценок по критериям $K_1, \ldots, K_N$ и представляется в виде векторной оценки $\boldsymbol{y} \in Y$, где $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)$, y_q равно номеру оценки из множества S_q.
- $C = \{C_1, C_2, \ldots, C_M\}$ – множество классов решений, упорядоченных по возрастанию выраженности свойства G.

Введем бинарное отношение строгого доминирования:

$$P = \left\{ (\mathbf{x}, \mathbf{y}) \in Y \times Y \,\middle|\, \forall q = 1 \ldots N \quad x_q \geq y_q \quad и \quad \exists q_0 : x_{q_0} > y_{q_0} \right\} \qquad (1)$$

Как нетрудно заметить, оно является антирефлексивным, асимметричным и транзитивным.

Удобно также рассматривать рефлексивное антисимметричное транзитивное бинарное отношение слабого доминирования Q:

$$Q = \left\{ (\boldsymbol{x}, \boldsymbol{y}) \in Y \times Y \;\middle|\; \forall q = 1 \ldots N \quad x_q \geq y_q \right\}$$

Требуется: с помощью ЛПР построить отображение $F : Y \to \{Y_i\}$, $i = 1 \ldots M$, (где Y_i – множество векторных оценок, принадлежащих классу C_i), удовлетворяющее свойству непротиворечивости:

$$\forall \boldsymbol{x}, \boldsymbol{y} \in Y: \boldsymbol{x} \in Y_i, \boldsymbol{y} \in Y_j, (\boldsymbol{x}, \boldsymbol{y}) \in P \Rightarrow i \geq j, \qquad (2)$$

Другими словами, объект с более характерным для свойства G набором оценок по критериям не может принадлежать к классу, соответствующему меньшей степени выраженности свойства G.

3.2. Оценка кредитного риска и проект стандарта Р7013

В современной банковской практике клиенты все чаще сталкиваются с ситуацией, когда вместо аналитиков банка, по существу решение о выдачи кредита принимают различные компьютерные системы, использующие технологии ИИ. В ряде случаев это приводит к различного рода неясностям и даже противоречиям. В работе [20] приведен пример решения такой задачи с использованием метода ВАР ЦИКЛ. Управление кредитным риском является повседневной практикой любого банка. Получение достаточно надежных оценок качества кредитов является сложной задачей, так как отсутствует единый индикатор вероятности невозврата средств. Существует множество индикаторов (факторов, критериев), которые необходимо принимать во внимание. Каждый такой фактор вносит определенный вклад в общую оценку. Для классификации банковских кредитов по группам риска была использована процедура агрегирования оценок отдельных параметров кредита, полученных от профильных специалистов банка и/или экспертов. Критерии оценки банковских кредитов разделены на несколько групп. Группа «Обеспеченность кредита» включала критерии: оценка предполагаемого обеспечения; ликвидность обеспечения; прогноз стоимости обеспечения; достаточность обеспечения. В группу «Оценка кредитного проекта» вошли критерий рентабельность проекта и предварительные условия его рассмотрения, характеризующие качество проработки проекта. Ценность заемщика для банка являлась самостоятельным критерием. Группа «Надежность заемщика» включала критерии:

статус заемщика; оценка позиции представителя заемщика на переговорах; наличие кредитной истории. В группу «Оценка финансового положения заемщика» вошли критерии: обороты по расчетным и текущим счетам в банке; тип финансовой устойчивости; наличие задолженности по кредитам другим банкам; доля задолженности 1–4 групп очередности платежей в кредиторской задолженности. В последнюю группу «Стабильность и перспективность фирмы заемщика» вошли критерии: управленческая культура организации-заемщика; наличие долговременных целей и планов их реализации; устойчивость организации-заемщика в зависимости от внешних условий на время кредитования; а также предварительные условия, характеризующие культуру управления в организации-заемщике. Перечисленные группы критериев образуют второй уровень иерархии. Третий уровень иерархии образуют следующие составные критерии: обоснованность кредита; оценка заемщика как организации; финансовое состояние и перспективы заемщика. На основании анализа этих критериев выработаны 4 категории качества кредитов: C_1 – высшая и высокая; C_2 – средняя и низкая; C_3 – сомнительная; C_4 – убытки, которые выступают в качестве классов решений. Для построения набора критериев и формирования шкал оценок составных критериев применялся метод ВАР ЦИКЛ [20] со снижением размерности признакового пространства [24, 25].

Особый интерес представляют такие системы оценки кредитного риска в связке, с представленным выше, проектом стандарта Р7013 (технологии автоматизированного анализа состояния лица) применительно к практике выдачи крупных займов корпоративным заемщикам. В работе [20] один из критериев, описывающих такого заемщика, формулируется следующим образом: «Позиция представителя заемщика на переговорах». Для получения оценок по такому критерию двадцать лет назад достаточно часто привлекали профессионального психолога, который скрытно анализировал мимику лица и поведение человека (например, представителя заемщика) на переговорах (фактически психолог осуществлял анализ характерных повадок мошенников и т. п.). В современных условиях такую задачу решает аппаратно-программный комплекс, включающий 20–30 скрытых видеокамер в переговорной комнате, а также программное обеспечение, выполняющее анализ видеоряда в реальном времени (как правило, используются нейронные сети). Именно для сертификации подобных компьютерных систем, осуществляющих анализ состояния лица, в которых используются различные тех-

нологии ИИ, и требуется повсеместное внедрение стандарта IEEE P7013 этики ИИ.

3.3. Оценка надежности новостных источников и проект стандарта P7011

Последнее десятилетие характеризуется беспрецедентными масштабами ведения различных информационных войн. Проект стандарта IEEE этики ИИ P7011 (оценка надежности новостных источников) ориентирован на устранение негативных последствий неконтролируемого распространения поддельных («фейковых») новостей путем предоставления открытой системы оценок. Очевидно, что «фейковые» новости являются одним из важнейших инструментов информационных войн, поэтому любые инициативы, направленные на «оздоровление» информационного пространства представляют определенный интерес. Если в самое ближайшее время не предпринять соответствующих усилий по предотвращению распространения «фейковых» новостей, это может привести к практически полной потере доверия конечных потребителей новостного контента к масс-медиа (телевидение, радио, интернет-ресурсы и т. п.). Потеря аудитории, в свою очередь, приведет к оттоку рекламодателей от производителей медиаконтента. А уже результатом этого будет фактическое разорение и банкротство большинства медиахолдингов. Поэтому во внедрении стандарта этики ИИ P7011 в первую очередь заинтересованы сами производители медиаконтента. Для решения указанной задачи, очевидно, могут быть использованы методы лингвистической семантики и семантического анализа текстов (семантические технологии Web) [21, 22, 30, 35]. Тем не менее, в современных условиях при распространении «фейковых» новостей необходимо проводить анализ и других компонентов (например, видеоряд, фотографии, звук и т. п.), т. к. они также могут быть фальсифицированы. Поэтому разработка специальной системы критериев для оценки надежности новостных источников может, в определенных условиях, также рассматриваться и как система этических норм в рамках стандарта этики ИИ P7011. Можно предложить следующую модельную систему критериев: 0. «Кричащий» заголовок (оценки: 0. Заголовок новости является «Кричащим»; 1. Нормальный заголовок); 1. Характеристика источника (оценки: 0. Подозрительный источник (много рекламы, странный дизайн, URL и т.п.; 1. Нормальный источник); 2. Неверная дата публикации (оценки: 0. Дата публи-

кации очевидно не соответствует появлению новостного сообщения; 1. Правильная дата); 3. Подложные фотографии (оценки: 0. Очевидно поддельные фотографии; 1. Достоверность фотографий вызывает определенные сомнения; 2. Нормальные фотографии); 4. Обилие ошибок (оценки: 0. Много орфографических и синтаксических ошибок; 1. Есть некоторое количество орфографических и синтаксических ошибок; 2. Ошибок нет); 5. Давление на жалость (оценки: 0. Присутствует давление на жалость; 1. Давления на жалость нет). Классы решений: A. Определенно фейковая новость. B. Есть подозрение, что новость является фейковой. C. Новость не является фейковой.

Построенная классификация выглядит следующим образом: Класс A (верхняя граница): (000000); Класс A (нижняя граница): (100010) (001120) (000220) (001201) (001111) (000211) (001021) (000121); Класс B (верхняя граница): (010000) (101000) (100100) (001210) (100020) (100001) (001121) (000221); Класс B (нижняя граница): (111200) (101210) (111120) (100220) (111101) (111021) (110121) (101121) (011221); Класс C (верхняя граница): (110210) (101220) (100201) (111111); Класс C (нижняя граница): (111221).

Нужно отметить, что если бы в исходной системе критериев использовались только бинарные оценки на шкалах критериев, и новости нужно было разделить только на два класса (категории), то можно использовать, например, математический аппарат булевой алгебры, т.е. применить подход, предложенный В. А. Лефевром [19] или методы расшифровки булевых функций [29], а также более эффективный с вычислительной точки зрения, по сравнению с методом ВАР ЦИКЛ, метод ВАР ДИФКЛАСС [18] для решения задачи формализации понятия этики ИИ.

Заключение

В представленной статье рассмотрена инициатива IEEE по разработке проектов стандартов этики ИИ. Предложены возможные определения этики и ИИ. Рассмотрены основные типы рисков (инженерный, модельный, экспертный и социологический), связанные с внедрением в повседневную жизнь современных технологий, использующих ИИ. Представлены примеры социально-экономических рисков, возникающих при опережающем внедрении технологий ИИ. Проанализирована особенность применения некоторых норм этики применительно к разработке и использованию современных

технологий ИИ. Представлен перечень проектов стандартов этики ИИ IEEE (P7000 – P7013), что говорит о достаточно широком спектре проблем, с которыми в самое ближайшее время столкнутся разработчики различных систем ИИ. Фактически это означает, что разработчики систем ИИ уже сейчас должны начать подготовку к прохождению различных процедур сертификации (т. е. соответствия их продукции, использующей технологии ИИ, указанным стандартам ИИ). В этой связи критический анализ различного математического инструментария, позволяющего формализовать понятие этики ИИ, будет способствовать разработке понятных и прозрачных «правил игры». К формальным методам, использование которых весьма востребовано для поставленной в статье задачи, можно отнести следующие: булева алгебра, многозначные логики, нечетная логика и теория вероятностей, а также методы ВАР. Благодаря тому, что к настоящему моменту все перечисленные концепции весьма развиты, можно достаточно оптимистично смотреть на решение задачи формализации понятия этики ИИ. В качестве положительных примеров рассмотрены варианты формализации понятия этики для проектов стандартов P7011 и P7013 с использованием метода ВАР ЦИКЛ. Из нерешенных задач необходимо отметить, что для каждого из проектов стандартов IEEE необходима разработка индивидуальных систем критериев (фактически это некий перечень норм) и соответствующих шкал оценок, по которым можно будет принять решение о степени соответствия той или иной технологии ИИ предложенным этическим стандартам.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проекты № 16-29-12901, 16-29-1278, 16-07-00865).

ЛИТЕРАТУРА

1. Анализ больших данных в интеллектуальной робототехнике / Г. В. Ройзензон, В. Э. Карпов, В. Е. Павловский, В. Б. Бритков // 10-я Всероссийская Мультиконференция по проблемам управления (МКПУ-2017). Материалы докладов / Под ред. И. А. Каляева. – Т. 2. Робототехника и мехатроника (РиМ-2017). – Ростов-на-Дону, Таганрог: Издательство Южного федерального университета, 2017. – С. 107–112.

2. Апресян, Р. Г. Этика: учебник / Р. Г. Апресян. – М.: КноРус, 2017.

3. Бек, У. Общество риска. На пути к другому модерну / У. Бек. – М.: Прогресс-Традиция, 2000.

4. Бритков, В. Б. Современные подходы анализа риска / В. Б. Бритков, Г. В. Ройзензон // Проблемы прогнозирования чрезвычайных ситуаций. XVI Всероссийская научно-практическая конференция. Сборник материалов. – М.: ФКУ Центр «Антистихия» МЧС России, 2017. – С. 22–24.

5. Бритков, В. Б. Многокритериальный подход к оценке ситуационных центров / В. Б. Бритков, Г. В. Ройзензон, А. Я. Фридман // Проблемы прогнозирования чрезвычайных ситуаций. XV Всероссийская конференция. Сборник материалов. – М.: ФКУ Центр «Антистихия» МЧС России, 2016. – С. 26–28.

6. Гусейнов, А. А. Античная этика / А. А. Гусейнов. – М.: URSS, 2017.

7. Дзикики, А. Творчество в науке / А. Дзикики; Под ред. Е. П. Велихова. – М.: URSS, 2001.

8. Достоверный и правдоподобный вывод в интеллектуальных системах / В. Н. Вагин, Е. Ю. Головина, А. А. Загорянская, М. В. Фомина; Под ред. В. Н. Вагина, Д. А. Поспелова. – М.: Физматлит, 2008.

9. Дробницкий, О. Г. Моральная философия: Избранные труды / О. Г. Дробницкий. – М.: Гардарики, 2002.

10. Еремеев, А. П. Темпоральные модели на основе логики ветвящегося времени в интеллектуальных системах / А. П. Еремеев, И. Е. Куриленко // Искусственный интеллект и принятие решений. – 2011. – № 1. – С. 14–26.

11. Интеллектуальные системы поддержки принятия решений в нештатных ситуациях с использованием информации о состоянии природной среды / В. А. Геловани, А. А. Башлыков, В. Б. Бритков, Е. Д. Вязилов. – М.: Эдиториал УРСС, 2001.

12. Карпенко, А. С. Развитие многозначной логики / А. С. Карпенко. – 3-е изд. – М.: URSS, 2016.

13. Карпов, В. Э. К вопросу об этике и системах искусственного интеллекта / В. Э. Карпов, П. М. Готовцев, Г. В. Ройзензон // Философия и общество. – 2018. – № 2(87). – С. 84–105. – DOI: 10.30884/jfio/2018.02.07.

14. Кропоткин, П. А. Этика: Избранные труды / П. А. Кропоткин. – М.: Политиздат, 1991.

15. Кузнецов, О. П. Неклассические парадигмы в искусственном интеллекте / О. П. Кузнецов // Известия РАН. Теория и системы управления. – 1995. – № 5. – С. 3–23.

16. Ларичев, О. И. Проблемы принятия решений с учетом факторов риска и безопасности / О. И. Ларичев // Вестник АН СССР. – 1987. – Т. 57, № 11. – С. 38–45.

17. Ларичев, О. И. Вербальный анализ решений / О. И. Ларичев. – М.: Наука, 2006.

18. Ларичев, О. И. Система ДИФКЛАСС: построение полных и непротиворечивых баз экспертных знаний в задачах дифференциальной классификации / О. И. Ларичев, А. А. Болотов // Научно-техническая информация. Серия 2. Информационные процессы и системы. – 1996. – № 9.– С. 9–15.

19. Лефевр, В. А. Алгебра совести / В. А. Лефевр. – М.: «Когито-Центр», 2003.

20. Метод многокритериальной классификации ЦИКЛ и его применение для анализа кредитного риска / А. А. Асанов, О. И. Ларичев, Г. В. Ройзензон и др. // Экономика и математические методы. – 2001. – Т. 37, № 2. – С. 14–21.

21. Методы и средства семантического структурирования электронных математических документов / А. М. Елизаров, Е. К. Липачев, О. А. Невзорова, В. Д. Соловьев // Доклады Академии наук. – 2016. – Т. 457, № 6. – С. 642–645.

22. Осипов, Г. С. Семантический анализ научных текстов и их больших массивов / Г. С. Осипов, И. В. Смирнов // Системы высокой доступности. – 2016. – Т. 12, № 1. – С. 41–44.

23. Поспелов, Д. А. «Серые» и/или «черно-белые» / Д. А. Поспелов // Прикладная эргономика. Специальный выпуск «Рефлексивные процессы». – 1994. – № 1. – С. 29–33.

24. Ройзензон, Г. В. Способы снижения размерности признакового пространства для описания сложных систем в задачах принятия решений / Г. В. Ройзензон // Новости искусственного интеллекта. – 2005. –№ 1. – С. 18–28.

25. Ройзензон, Г. В. Синергетический эффект в принятии решений / Г. В. Ройзензон // Системные исследования. Методологические проблемы. Ежегодник / Под ред. Ю. С. Попкова, В. Н. Садовского, В. И. Тищенко. –№36. 2011-2012.М.: УРСС, 2012. – С. 248–272.

26. Ройзензон, Г. В. Проблемы формализации понятия этики в искусственном интеллекте / Г. В. Ройзензон // Шестнадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2018). Труды конференции. В 2-х томах. – Т. 2. – М.: РКП, 2018. – С. 245–252.

27. Российская рабочая группа IEEE по вопросам этики ИИ. – 2017. – Режим доступа: http://ecai.raai.org/.

28. Согомонов, А. Ю. Этика инженера – гибкий свод моральных практик / А. Ю. Согомонов // Ведомости прикладной этики. – 2013. – № 42. – С. 14–26.

29. Соколов, Н. А. Оптимальная расшифровка монотонных булевых функций / Н. А. Соколов // Журнал вычислительной математики и математической физики. – 1987. – Т. 27, № 12.– С. 1878–1887.

30. Сулейманов, Д. Ш. Система семантического анализа ответных текстов обучаемого на естественном языке / Д. Ш. Сулейманов // Онтология проектирования. – 2014. – № 1(11). – С. 65–77.

31. Тарасов, В. Б. От многоагентных систем к интеллектуальным организациям: философия, психология, информатика / В. Б. Тарасов. – М.: Эдиториал УРСС, 2002.

32. Финн, В. К. Интеллектуальные системы и общество: Сборник статей / В. К. Финн. – М.: КомКнига, 2006.

33. Фоминых, И. Б. О формализации гуманитарного знания / И. Б. Фоминых // Одиннадцатая национальная конференция по искусственному интеллекту с международным участием (КИИ-2008). Труды конференции. – Т. 1. – М.: Ленанд, 2008. – С. 133–141.

34. Фролов, И. Т. Этика науки: Проблемы и дискуссии / И. Т. Фролов, Б. Г. Юдин. – М.: URSS, 2016.

35. Хорошевский, В. Ф. Семантические технологии: ожидания и тренды / В. Ф. Хорошевский // Открытые семантические технологии проектирования интеллектуальных систем. – 2012. – № 2. – С. 143–158.

36. Цыгичко, В. Н. Безопасность критических инфраструктур / В. Н. Цыгичко, Д. С. Черешкин, Г. Л. Смолян. – М.: URSS, 2019.

37. Швейцер, А. Культура и этика / А. Швейцер. – М.: Прогресс, 1973.

38. Шпеман, Р. Основные понятия морали / Р. Шпеман. – М.: Московский философский фонд, 1993.

39. Шрейдер, Ю. А. Проблема неполного добра в модели ценностной рефлексии по В. А. Лефевру / Ю. А. Шрейдер, Н. Л. Мусхелишвили // Системные исследования. Методологические проблемы. Ежегодник / Под ред. Д. М. Гвишиани, В. Н. Садовского. – № 25. 1997. М.: УРСС, 1997. – С. 213–224.

40. IEEE. Ethically Aligned Design. – 2016. – Режим доступа: https://ethicsinaction.ieee.org.

41. Zadeh, L. A. Fuzzy sets / L. A. Zadeh // Information and Control. – 1965. – Vol. 8, no. 3. – P. 338–353.

УДК 81.322.2; 81.367.7

ON ONE OF THE APPROACHES TO DETECTION OF UNITS OF THE PROSODIC SYSTEM OF THE RUSSIAN LANGUAGE BASED ON BRIEF INTERROGATIVE REMARKS OF SPONTANEOUS RUSSIAN SPEECH

G. M. Sagmanova
St. Petersburg State University, St. Petersburg
guzel.sagmanova@mail.ru

This article is devoted to the problem of distinguishing and classifying prosody units in the Russian language. The prosodic level of language organization is one of the most complicated to describe, because of its specific functioning in speech. There is no consensus among experts in concern of prosody units selection and the intonation inventory. Considering the quality of the research content, linguists often use only prepared speech in their studies. This article proposes one of the possible ways to elaborate a classification of prosody models based on spontaneous Russian speech within the framework of the functional-semantic type of interrogative remarks. The research attempts to use a model approach to the prosody level of speech organization based on tone movements in the intonation center, which is used for identifying typical melodic curves.

Keywords: intonation; prosody; interrogative remark; intonation center; classification of intonation units; tone; melody; live speech; Russian.

ОБ ОДНОМ ИЗ ПОДХОДОВ К ВЫЯВЛЕНИЮ ЕДИНИЦ ИНТОНАЦИОННОЙ СИСТЕМЫ РУССКОГО ЯЗЫКА НА МАТЕРИАЛЕ КРАТКИХ ВОПРОСИТЕЛЬНЫХ РЕПЛИК СПОНТАННОЙ РУССКОЙ РЕЧИ

Г. М. Сагманова
Санкт-Петербургский государственный университет, Санкт-Петербург
guzel.sagmanova@mail.ru

Данная статья посвящена проблеме выделения и классификации интонационных единиц русского языка. Просодический уровень организации языка является одним из самых сложных для описания в силу специфики его функционирования в речи. Мнения специалистов в области просодии расходятся уже на этапе выделения интонационной единицы и инвентаря интонации и находят свое отражение на классификации интонационных единиц. При этом в качестве материала исследования часто используется

лишь речь подготовленная, дикторская. В настоящей работе на материале спонтанной русской речи в рамках функционально-семантического типа вопросительных реплик предлагается один из возможных путей формирования интонационной классификации. В исследовании проведена попытка применить модельный подход к интонационному уровню организации речи, с выделением типовых мелодических кривых, опираясь на движение тона в интонационном центре.

Ключевые слова: интонация; просодия; вопросительная реплика; интонационный центр; классификация интонационных единиц; тон; мелодика; живая речь; русский язык.

The prosodic level of system of language organization is a certain degree of difficulty to linguistic study and description. Despite the fact that using of prosodic means in speech is universally for the vast majority of languages, ways and methods of its researching differ both among national linguistics and within the scientific environment studied a particular language, which is associated with some peculiarities of the functioning of intonation.

Intonation represents abstract means, which are “difficult not only to define but also to detect because of its non-objective nature” [11], heterogeneous categories related means: it can reflect semantic of the statement, structure position of the sentence part to the whole sentence, speaker’s emotional state [10]. “There is no language layer as close to universality in its categories as intonation, and at the same time there is no language form as difficult to learn and as specific as the phrasal intonation” [9]

This leads to the fact that the problem of elaborate a classification of prosody models and its inventory does not find a definite solution in the world linguistics. Researchers, depending on the chosen method, operate with different ideas about intonation units in the description of the prosodic system of language. In the Russian tradition, the classical theory is theory of E.A. Bryzgunova, which distinguishes seven types of intonation structures based on the ratio of timbre, intensity, duration of speech and, mainly, the movement of tone, manifested in a particular syntagma. Typical combinations of these parameters are correlated with the three-part division of the intonation structure into the center, post-center and pre-center parts, as well as with the value transmitted by each construction [2].

After Bryzgunova this approach develop N.B. Volskaya [3] and T.E. Janko [11]. O.T. Yokoyama combines it with an autosegmental

approach and proposes to divide the intonation into two levels: tonal and tonemic, (by analogy with phonetic and phonemic) [7].

S.V. Kodzasov, based on the functional conditionality of intonation, and, therefore, of the multiplicity of its values, does not take as a basis the concept of the model or intonational construction, and moreover does not aim to formulate the final list of "prosodems". In a combinatorial description of phrasal prosody proposed by him mora is offered as the minimal prosodic unit (in which case, for example, all complex tones are considered as bimoraic) [8].

In this paper we have made an attempt to apply a model approach to prosodic and to make an allocation of typical prosody structures based primarily on the movement of tone in the intonation center. Also it was interesting to consider the implementation of the prosodic system of the Russian language on the content of spontaneous oral speech, because "the appeal to natural, spontaneous speech, shows a greater variety of basic intonation types, which cannot be ignored if the researcher does not only describe a laboratory prepared content but sets a task to appropriately represent the process of communication" [4].

The source of the content was the sound corpus of the Russian language of everyday communication «One Speaker's Day», developed at the philological faculty of St. Petersburg state University [1], [6]. At the moment, this is the only corpus that records, first of all, spontaneous, the most natural oral speech[1]. Recording for it was made on a portable voice recorder in a living household environment without the control of collectors. Thus, the content of the ORD corpus is able to reflect the specifics of the prosody of unprepared oral speech, in contrast to the sources of the content of previous studies in this area.

The selection of the content was made manually by listening to the records of the informant's speech. At the initial stage of the study, in order to eliminate the problem of syntagmatic borders of sentences, the content was limited to short remarks, no longer than ten syllables. In addition, as a mandatory criterion for the selection of the content was the isolation of remarks, that is, the limitation of their "right" and "left" by remarks of the interlocutor or the long pause between the selected remark and the adjacent phrases of the speech flow of the speaker. This makes

[1] Stipulation is necessary about some cases of speech situation presented in the corpus content, as, for example, reading lectures, recording the vocals or monologue in a studio, acting at rehearsal, etc.

it possible to bypass the problem of influence on the chosen remark of prosodic characteristics of neighboring phrases of the informant.

Studying the prosody of a language linguists usually start from the forms of its expression or from the meaning represented by it. In the first case, groups united by common functions in speech are distinguished from the set of naturally repeating combinations of prosodic characteristics. In the second case a number of already known semantically or structurally conditioned types of sentences or parts of sentences is considered in terms of the presence in their implementation of typical repetitive combinations of prosodic characteristics. In the research functional and semantic indicators of remarks was chosen as a starting point with the disclaimer that their selection itself is largely based on the acquired stereotypes of perception of prosodic. In addition, the necessity of further correction of the data on prosodic characteristics considering synonymy and homonymy of the prosodic of the Russian language is taken into account.

At the moment, the research subcorpus contains 406 short remarks of the speech of sixteen informants: eight women and eight men – representatives of five age groups. Communicative situations meet the requirement of heterogeneity: there is communication with family and friends (including through telephone and video), colleagues, unfamiliar people. Functional and semantic markup of the content revealed a predominance of interrogative remarks– 133 units. Interrogative sentences play a key role in building of everyday communication [5] and they are reflected in most classifications of the prosodic system of the Russian language. In this regard we first turned to the research of this type of sentences. It should be noted that functional and semantic markup of statements was made by the method of auditory analysis and with the inevitable use of everyday "language intuition", but without reliance on the pragmatic aspect of the communicative situation. Thus, the content turned out to be both the actual requests for information and reflexive questions, questions of surprise, mimicking the interrogative remark of the interlocutor, etc.

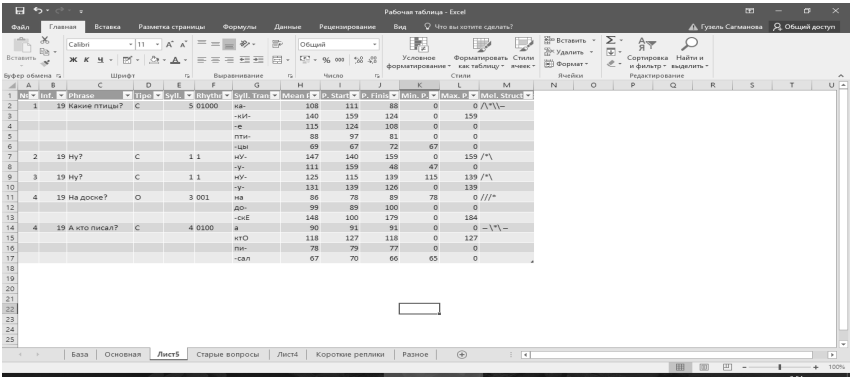
At first, the remarks selected and transcribed by the spelling principle, following the classical intonation classifications, were allocated into three functional-semantic types of questions:

– common questions – 57 units, 43%: *Чай будешь? У тебя нет синей рубашки?;*

– special questions – 70 units, 52% (*Ты во сколько подъедешь? Почему именно синяя?);*

– questions with the opposition – 6 units, 4,5% (*А у тебя как? А где закончил?*).

The prosodic stage of the research included, firstly, the use of the method of auditory analysis: the number of syllables was determined for each remark divided into syllables, then a simple rhythmic markup was made (where 1 is a stressed syllable, 0 is a unstressed syllable), the phrasal stress was determined. Secondly, the method of tool-analysis was used: prosodic markup was made with the help of the Praat software. The common pitch of each syllable, the final and initial pitch of each syllable, as well as the maximum and minimum pitch of the entire remark were recorded. The findings were entered into the table considering the average pitch of each informant.



Picture 1. The fragment of the worksheet with the tool-analysis data

On the base of the obtained melodic parameters, each remark was analyzed using the developed notation system

C – intonation center;

I – increase of tone;

D –decrease of tone;

E – even tone;

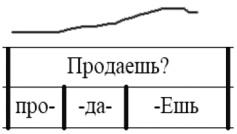
H – high start;

L –low start;

{} – wave of melodic curve.

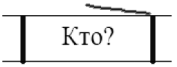
By the fluctuation of the melodic curve in the intonation center, four types of intonation structures were distinguished:

with increasing tone in intonation center (CI):



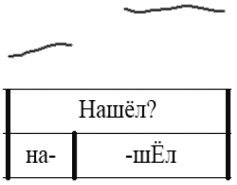
Picture 2.

– with decreasing tone in intonation center (CD):



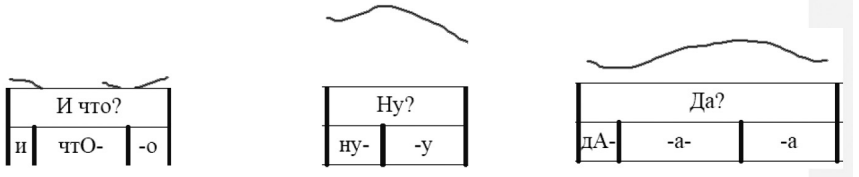
Picture 3.

– with even tone in intonation center (CE):



Picture 4.

– combined types (C(D-I), C(I-D), C(D-I-D) etc.):



Picture 5. Examples of types of structures C(D-I), C(D-I-D), C(I-D)

Table 1. **The general types of intonational structures**

General types	Percentage
CI	50%
CD	32%
Combined	10%
CE	8%

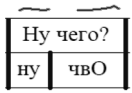
The fluctuation of the melodic curve was recorded not only for the intonation center, but also for the center and post-center parts, if they were in the remark. The formulas reflect the particular models of the mentioned above common types of intonational structures. For example, E-CI-D (13 units), E-CI (12 units), CI (8 units) were the most frequent models for structures of type CI; CD (5 units), D-CD (5 units), CD-E (4 units) were the most frequent for structures of type CD; single models were found for all eleven structures with smooth tone motion in intonation center; among the combined structures, D-(H)C(D-I) and C (I-D) particular models were found twice.

As can be seen, studied content showed a rather impressive number of the particular models of intonation types, herewith many of them are found only once. It will be possible to neutralize the bulkiness of the turned out initial typology via subsequent adjustment, which is necessary for some particular models. It is obvious, for example, that the differences between E-CI and CI models lie only in the presence or absence of pre-stressed syllables, which can be considered as an optional feature of the more general structure realizations.

Correlation of the data obtained during the prosodic analysis with the functional-semantic types of interrogative remarks revealed a predominance of structures with a rising tone in the intonation center in the common questions (48 remarks from 57) and the predominance of structures with decreasing tone in the intonation center special questions (36 remarks from 70). Such values were expected, and they can be called standard, because the correlation of prosodic and semantic features of utterances they displayed are described by linguists and listed in classic prosodic classification. For instance, the classification of Bryzgunova considered common questions with rising tone in the intonation center as one of the realizations of IK-3, and special questions with lowering the tone in the intonation center are described there as the IK-2.

The rest of the content gave the following values:

a) increasing tone in intonation center in the special questions (8remarks, 6%):

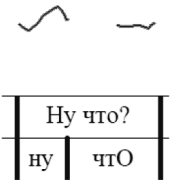


Picture 6.

b) decreasing tone in intonation center in the common questions (4remarks, 3%):

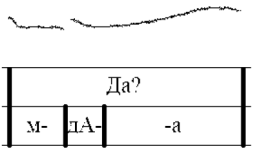
Picture 7.

c) even tone in intonation center in different functional-semantic types of questions (11 remarks, 8%):



Picture 8.

d) combined types of structures (13 remarks, 10%):

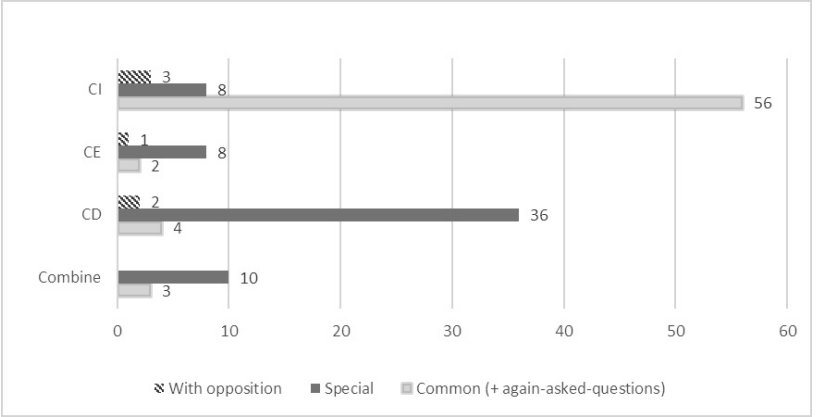


Picture 9.

e) models of questions with the opposition (5 remarks, 4%[1]):

Picture 10

[1] The quantitative value for the models of questions with oppositions is given with the deduction of one case of simultaneous occurrence of the remark also in the number of models of combined types of structures.



Picture 11. The correlation of the number of typical intonation structures and functional-semantic types of interrogative remarks

Thus, the total share of "traditional" intonation models for interrogative remarks in the considered content is 69%, the other types of structures occupy, respectively, 31%. Of course, some of the models of the second group of types require further adjustment, because in no case it is excluded that some realizations may be individual, specific only for a particular speaker. In addition, deviations from the standard ways of intonation can be caused by the speaker's emotional state and the specifics of a certain speech act. However, it seems that a group other than "traditional" types of structures worth paying attention to further, and that it is necessary to by expanding the content try to reveal sustainable intonation models that are specific to a brief remarks of the Russian oral spontaneous speech.

REFERENCES

1. Асиновский, А.С., Богданова, Н.В., Русакова, М.В., Степанова, С.Б., Рыко А.И., Шерстинова Т.Ю. The ORD Speech Corpus of Russian Everyday Communication «One Speaker's Day»: Creation Principles and Annotation. Lecture Notes in Computer Science, Text, Speech and Dialogue (5729/2009), 2009. pp. 250–257.

2. Брызгунова Е.А. Интонация, Русская грамматика. Том 1, Наука, Москва, 1982. С. 98–118.

3. Вольская Н.Б., Качковская Т.В. Принципы просодической разметки в новом корпусе русской спонтанной речи CoRuSS // Фонетика сегодня, 2016. С. 29–31.

4. Вольская Н.Б., Скрелин П.А. Система интонационных моделей для автоматической интерпретации интонационного оформления высказывания: функциональные и перцептивные характеристики // Труды третьего междисциплинарного семинара «Анализ разговорной русской речи» (АРЗ-2009). Санкт-Петербургский институт информатики и автоматизации РАН, 2009. С. 28–40.

5. Глазкова С.Н. Функционально-прагматический подход к типологии вопросительного высказывания // Проблемы истории, филологии, культуры. 2008. № 20. С. 182.

6. Звуковой корпус как материал для анализа русской речи. Коллективная монография. Часть 2. Теоретические и практические аспекты анализа. Том 1. О некоторых особенностях устной спонтанной речи разного типа. Звуковой корпус как материал для преподавания русского языка в иностранной аудитории), Отв. ред. Н. В. Богданова-Бегларян, СПб, Филологический факультет СПбГУ, 2014.

7. Йокояма О. Нейтральная и ненейтральная интонация в русском языке: автосегментная интерпретация системы интонационных конструкций // Вопросы языкознания, № 5, 2003. С. 99–122.

8. Кодзасов С.В. Исследования в области русской просодии. М.: Языки славянских культур, 2009. – 496 с.

9. Николаева Т.М. Фразовая интонация славянских языков. М.: Наука, 1977. – 279 с.

10. Светозарова Н.Д. Интонационная система русского языка. Ленинград: Издательство Ленинградского университета, 1982. – 176 с.

11. Янко Т.Е. Интонационные стратегии русской речи в сопоставительном аспекте. М.: Языки славянских культур, 2008. – 312 с.

УДК 004 934

APPLICATION OF CONTINUOUS WAVELET-TRANSFORMATION FOR FILTERING SYNTHESIZED SPEAKER SIGNAL

V. I. Semenov, A. K. Shurbin
Chuvash State University, Cheboksary
syundyukovo@yandex.ru, shurti@mail.ru

The algorithm of the inverse continuous wavelet transform in the frequency domain is presented with the use of the fast Fourier transform. The algorithm allows to increase by four orders of magnitude the speed of calculation of the inverse continuous wavelet transforming in comparison with direct numerical integration.

Keywords: wavelet transformation, scale factor, speech synthesis, Fourier transform, speech signal.

ПРИМЕНЕНИЕ НЕПРЕРЫВНОГО ВЕЙВЛЕТ-ПРЕОБРАЗОВАНИЯ ДЛЯ ФИЛЬТРАЦИИ СИНТЕЗИРОВАННОГО РЕЧЕВОГО СИГНАЛА

В. И. Семенов, А. К. Шурбин
ФГБОУ ВПО «Чувашский государственный университет им. И. Н. Ульянова», Чебоксары
syundyukovo@yandex.ru, shurti@mail.ru

В работе представлен алгоритм обратного непрерывного вейвлет-преобразования в частотной области с применением быстрого преобразования Фурье. Алгоритм позволяет на четыре порядка увеличить скорость вычисления обратного непрерывного вейвлет-переобразования по сравнению с прямым численным интегрированием.

Ключевые слова: вейвет-преобразование, масштабный коэффициент, синтез речи, Фурье преобразование, речевой сигнал.

При синтезе речевого сигнала в местах соединения фонем возникают значительные перепады частот и амплитуды сигнала, что негативно сказывается на качестве синтезируемой речи. Разрыв в точках соединения отчетливо ощутим на слух. Чтобы устранить этот недостаток, авторами используется локальная фильтрация. Вблизи точек соединения фонем высокочастотные коэффициенты приравнивают-

ся нулю, а остальные участки остаются без изменений. В результате реконструкции синтезированного сигнала в местах соединения фонем разрывы сглаживаются, что положительно сказывается на качестве синтезированной речи. Для фильтрации используется быстрое непрерывное вейвет-преобразование (ВП)**.** Непрерывное ВП одномерного сигнала $S(t)$ производится по формуле:

$$W(a,b) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} S(t)\ \psi\left(\frac{t-b}{a}\right)\ dt, \qquad (1)$$

где первый аргумент a (временной масштаб) аналогичен периоду осцилляций, а второй b – смещению сигнала по оси времени. Реконструкция выполняется с применением формулы обратного непрерывное ВП:

$$S(t) = C_{\psi}^{-1} \int_{0}^{\infty} \int_{-\infty}^{\infty} \psi\left(\frac{t-b}{a}\right)\ W(a,b) \frac{dadb}{a^{3+k}}, \qquad (2)$$

где C_{ψ} – нормализующий коэффициент:

$$C_{\psi} = \int_{-\infty}^{\infty} |F_{\psi}(\omega)|^2 \cdot \omega^{-1} d\omega < \infty,$$

$F_{\psi}(\omega)$ – Фурье–спектр базисной функции, ω – циклическая частота, k – показатель степени масштабного множителя.

Вычисление ВП прямым численным интегрированием для больших временных последовательностей по формулам (1) и (2) занимает длительное время. Для увеличения быстродействия, авторами разработан алгоритм непрерывного быстрого ВП в частотной области с использованием быстрого преобразования Фурье (БПФ). Вейвлет-преобразование используется не только для синтеза речи, но и для распознавания речи [1,2,3]. Алгоритм численного вычисления прямого быстрого непрерывного вейвлет-преобразования в частотной области приведен в работах [4,5]. Нормализующий коэффициент в формуле (2) непрерывного обратного вейвлет-преобразования $C = C_{\psi}$ в разработанном алгоритме вычисляется из аналога теоремы Парсеваля для вейвлет-коэффициентов:

$$\int S(t) S^{*}(t) dt = C^{-1} \iint W(a,b) W^{*}(a,b) \frac{dadb}{a^2}. \qquad (3)$$

После определения нормализующего коэффициента C из (1) он подставляется в формулу

$$S(t) = C^{-1} \int_0^{\infty} \int_{-\infty}^{\infty} \psi\left(\frac{t-b}{a}\right) W(a,b) \frac{dadb}{a^2}. \qquad (4)$$

Теоретической основой вычисления обратного непрерывного быстрого вейвлет-преобразования сигнала $S(t)$ в частотной области является использование формул (4) и (3). Обратным преобразованием произведения спектров вейвлет-спектра $W(a,b)$ и вейвлета $\psi(t)$ вычисляется интеграл по переменой b. Суммированием полученного интеграла по масштабному коэффициенту a рассчитывается реконструированный сигнал $S(t)$.

Алгоритм численного вычисления обратного непрерывного вейвлет-преобразования по формуле (4) в частотной области включает следующие шаги.

1. Вычисляются коэффициенты тригонометрического ряда $a_1(n)$ вейвлет-спектра $W(a,b)$ с использованием прямого БПФ по формуле:

$$a_1(n) = \frac{1}{N} \sum_{k=0}^{N-1} W(a,k) \cos\left(\frac{2\pi nk}{N}\right).$$

2. Вычисляются коэффициенты тригонометрического ряда $b_1(n)$ вейвлет-спектра $W(a,b)$ с использованием прямого БПФ по формуле:

$$b_1(n) = \frac{1}{N} \sum_{k=0}^{N-1} W(a,k) \sin\left(\frac{2\pi nk}{N}\right).$$

3. Вычисляются коэффициенты тригонометрического ряда $a_2(n)$ вейвлета $\psi(t)$ с использованием прямого БПФ по формуле:

$$a_2(n) = \frac{1}{N} \sum_{k=0}^{N-1} \psi(k) \cos\left(\frac{2\pi nk}{N}\right).$$

4. Вычисляются коэффициенты тригонометрического ряда $b_2(n)$ вейвлета $\psi(t)$ с использованием прямого БПФ по формуле:

$$b_2(n) = \frac{1}{N} \sum_{k=0}^{N-1} \psi(k) \sin\left(\frac{2\pi nk}{N}\right).$$

5. Вычисляется комплексно сопряженный спектр по формулам,

$$c_1(n) = a_1(n) \cdot a_2(n) + b_1(n) \cdot b_2(n),$$
$$c_2(n) = b_1(n) \cdot a_2(n) - a_1(n) \cdot b_2(n).$$

Большинство непрерывных вейвлетов – либо четные, либо нечетные функции. Для четных вейвлетов ряд составлен из одних косинусов, а для нечетных – из одних синусов. Для четных вейвлетов $b_2(n) = 0$ и

$$c_1(n) = a_1(n) \cdot a_2(n), \quad (5)$$

$$c_2(n) = b_1(n) \cdot a_2(n). \quad (6)$$

Для нечетных вейвлетов , $a_2(n) = 0$ и

$$c_1(n) = b_1(n) \cdot b_2(n), \quad (7)$$

$$c_2(n) = -a_1(n) \cdot b_2(n). \quad (8)$$

6. Для четного вейвлета путем $M + 1$ обратных преобразования Фурье от комплексно сопряженного спектра (5), (6) вычисляется функция $s_m'(n)$:

$$s_m'(n) = \sum_{k=0}^{N-1} c(k) \exp(i\frac{2\pi nk}{N}).$$

7. Для нечетного вейвлета путем $M + 1$ обратных преобразований Фурье от комплексно сопряженного спектра (7), (8) вычисляется функция $s_m'(n)$:

$$s_m'(n) = \sum_{k=0}^{N-1} c(k) \exp(i\frac{2\pi nk}{N}),$$

(обозначение ‘ не означает дифференцирование).

8. По формуле (3) вычисляется нормализующий коэффициент C.

9. По формуле

$$S(n) = C \sum_{m=0}^{m} s_m'(n),$$

реконструируется сигнал, где m – уровень декомпозиции. Постоянную C, можно определить проще, используя следствие формулы (3) (теоремы Парсеваля). В пространстве действительных функций плотность энергии сигнала

$$E_W(a,b) = W_l^2(a,b).$$

Локальная плотность энергии в точке t_0

$$E_\delta(a,t_0) = W_l^2(a,t_0).$$

Тогда

$$S(t_0) = C \sum_{m=0}^{m} s_m^{'}(t_0). \qquad (9)$$

Постоянная C вычисленная по формуле (9), совпадает с постоянной, найденной по формуле (9). Чтобы при вычислении по формуле (9) не было деления на ноль или умножения на отрицательное число, постоянную C лучше вычислить для функции в максимуме.

ЛИТЕРАТУРА

1. Патент на изобретение №2403628 РФ Способ распознавания ключевых слов в слитной речи. Семенов В.И. Желтов П.В., № 2008141557/09(053961). Заявл. 20.10.2008. Опубл. 10.11.2008 г.

2. Семенов В.И., Желтов П.В. Распознавание речи на основе вейвлет-преобразования./ Чуваш. ун-т. Чебоксары, 2008. 16с. Деп. в ВИНИТИ РАН 29.02.08, №174-В2008.

3. Семенов В.И., Желтов П.В. Вейлет-преобразование акустического сигнала/ КГТУ им. А.Н. Туполева. Казань, 2008. 102 с.

4. Семенов В.И., Шурбин А.К., Михеев К.Г., Михеев Г.М. Фильтрация изображений, полученных с помощью оптического микроскопа, с применением кратномасштабного анализа. Химическая физика и мезоскопия том 16 №3, Ижевск, 2014. С. 399–404.

5. В.П. Желтов, П.В. Желтов, В.И. Семенов, А.И. Трофимова, А.К. Шурбин. Распознавание слитной речи с использованием вейвлет-преобразования. Труды Казанской школы по компьютерной и когнитивной лингвистике, TEL-2014, Казань, 2014. С. 9–13.

УДК 004.912, 004.82

APPROACH TO COREFERENCE RESOLUTION BASED ON ONTOLOGICAL SIMILARITY MEASURE

E. A. Sidorova, N. O. Garanina, I. S. Kononenko, A. S. Sery
A.P. Ershov Institute of Informatics Systems SB RAS, Novosibirsk
lsidorova@iis.nsk.su, garanina@iis.nsk.su, irina_k@cn.ru,
alexey.seryj@iis.nsk.su

We suggest a logical-ontological approach to the coreference resolution in the process of text analysis and information extraction. Our approach solves the problem of comparing objects found in the text – instances of ontology classes – using the evaluation of the similarity of attributes and relations of objects.

Keywords: ontology population, text analysis, information extraction, coreference resolution, referential factors, polyadic relations.

ПОДХОД К РАЗРЕШЕНИЮ КОРЕФЕРЕНЦИИ НА ОСНОВЕ ОНТОЛОГИЧЕСКИХ МЕР СХОДСТВА

Е. А. Сидорова, Н. О. Гаранина, И. С. Кононенко, А. С. Серый
Институт систем информатики им. А. П. Ершова СО РАН, Новосибирск
lsidorova@iis.nsk.su, garanina@iis.nsk.su, irina_k@cn.ru,
alexey.seryj@iis.nsk.su

Рассматривается логико-онтологический подход к разрешению кореференции в процессе анализа текста и извлечения информации. Данный подход решает задачу сравнения найденных в тексте объектов – экземпляров классов онтологии – на основе оценки близости их атрибутов и связей.

Ключевые слова: пополнение онтологии, анализ текста, извлечение информации, разрешение кореференции, референциальные факторы, многоместные отношения.

Обнаружение референциальных отношений в дискурсе относится к числу наиболее существенных и сложных проблем, с которыми сталкиваются специалисты по автоматической обработке естественного языка. Суть отношения референции заключается в сопоставлении языковых выражений внеязыковым объектам – референтам. Разрешение анафоры и кореференции востребовано во всех NLP-

приложениях, то есть тех, что ориентированы на автоматическую обработку естественного языка. Среди них машинный перевод, реферирование текста и извлечение информации.

В рамках данной работы задача разрешения кореференции рассматривается как один из основных этапов процесса извлечения информации из текстов с целью пополнения онтологии [1]. Особое значение придается здесь тесной взаимосвязи извлечения информации и пополнения онтологии. Процесс организован таким образом, что, с одной стороны, онтология используется для представления результатов извлечения информации, с другой – знания, представленные в онтологии, помогают решать специфические задачи извлечения информации. Совокупность всех процессов образует систему извлечения информации или IE-систему. Входными данными для данной системы являются онтология предметной области, правила извлечения информации и результаты предварительной обработки текста, включающие терминологическое, тематическое и сегментное покрытия. Терминологическое покрытие является результатом лексического анализа текста – предварительного извлечения терминов из текста и формирования лексических объектов на основе семантических словарей. Сегментное покрытие – это разбиение текста на формальные и жанровые фрагменты. Тематическое покрытие формируется из фрагментов, в которых преобладает соответствующая тематика.

Онтология предметной области моделирует значимую для пользователя часть предметной области и осуществляет структуризацию информации в рамках контента информационной системы. Онтология вместе с множеством экземпляров ее понятий образует базу знаний, соответственно задача пополнения онтологии заключается в вычислении новых экземпляров из входных данных.

Система пополнения онтологии имеет модульную структуру. Модуль извлечения информации сопоставляет найденным лексическим объектам экземпляры понятий и отношений онтологии на основе правил. Правила формируются автоматически на основе моделей фактов, составляемых экспертами вручную. В модели факта заложена информация об ограничениях, накладываемых на морфологические, синтаксические, структурно-текстовые и лексико-семантические характеристики объектов с учетом рассматриваемой онтологии и языка предметной области [2]. Модуль разрешения кореференции [3] обеспечивает формирование гипотез о наличии кореферентных связей и вычисляет их вес. Модуль разрешения

неоднозначности разрешает все виды конфликтов, возникающих вследствие рассмотрения различных вариантов разбора текста, и выбирает наиболее информативный вариант [4]. Результатом работы системы является пополнение контента онтологии найденными в тексте экземплярами понятий и отношений предметной области – информационными объектами.

В классическом понимании онтологии отсутствует понятие многоместного отношения. При этом они часто востребованы при решении задач извлечения информации, поскольку естественным образом описывают пропозициональное содержание высказывания, представляющее внеязыковую ситуацию, такую как событие, действие или процесс.

За основу модели многоместных отношений нами был взят один из паттернов онтологического проектирования [5], предложенных W3C в рамках описания специфических значений, и представляющий многоместное отношение онтологическим классом, связанным со всеми аргументами обычными бинарными отношениями. На связи данного класса налагаются определенные ограничения для моделирования нужных нам многоместных отношений. В частности, такой класс должен иметь количество связей, точно совпадающее с арностью моделируемого им отношения. Совокупная область значений (range) всех связей должна в точности соответствовать аргументам моделируемого многоместного отношения.

Ранее нами уже были рассмотрены два типа факторов [3], влияющих на оценку степени референциального сходства двух информационных объектов: дискурсивные и семантические. В рамках данной работы можно выделить еще один тип факторов – логико-онтологический, который позволяет рассматривать совокупность связанных отношений между объектами. Рассмотрение таких факторов опирается на свойства отношений, заданных в онтологии. Факторы используются при оценке сходства объектов, упоминания которых были зафиксированы в тексте системой. Для каждого фактора формулируется оценка, характеризующая степень кореферентной связи между объектами в зависимости от него, и без учета других факторов.

Совокупное сходство объектов по всем известным факторам используется для их сравнения в ситуациях кореференциальных конфликтов, когда два заведомо некореферентных объекта *a* и *b* могут быть потенциально кореферентны одному и тому же третьему объекту *c*. В этом случае мы считаем, что конфликт решен

в пользу объекта *a*, если и только если *a* более схож с объектом *c*, чем объект *b*.

На данный момент наибольший вклад в оценку сходства вносят семантические факторы. Семантическое сходство $S(a,b)$ информационных объектов *a* и *b*, вычисляемое по формуле (1), определяет степень близости множеств их свойств: атрибутов и связей. Сравнивая эти множества, мы учитываем как сходство значений входящих в них элементов, так и дополнительные характеристики, основанные на онтологических свойствах этих элементов.

$$S(a,b) = S^{EQ} + (1 - S^{EQ}) \cdot S^{\Delta} \qquad (1)$$

Величина $S^{EQ} \in [0;1]$ отражает сходство значений соответствующих атрибутов и связей объектов *a* и *b* без учета дополнительных характеристик, а $S^{\Delta} \in [0;1)$ – дополнительную информацию, предоставляемую этими характеристиками. Очевидно, что полное семантическое сходство достигается только при полном совпадении значений свойств объектов. В противном случае *S* асимптотически приближается к единице, и увеличение дополнительной информации всегда приводит к росту значения *S*.

Когда мы говорим о многоместных отношениях как об одном из факторов для оценки сходства, то имеем в виду две ситуации: сравнение экземпляров многоместных отношений как обычных объектов с целью обнаружить наличие между ними кореференциальных отношений, либо использование при сравнении объектов информации о многоместных отношениях, в состав которых они входят.

Таким образом, предложенный в статье подход делает упор на предметные знания, в первую очередь на онтологию предметной области, обеспечивает расширяемость относительно правил извлечения информации и референциальных факторов, ориентирован на полностью автоматическую обработку и интегрирует вычислительные и лингвистические модели и методы анализа текста на этапе семантической обработки.

Данный подход апробируется на текстах технических заданий из предметной области автоматизированных систем управления. В рамках исследования проводится разметка корпуса и выявление случаев, когда для корректного установления кореференции может использоваться онтологическая информация о многоместных отношениях.

Благодарности. Работа выполнена при финансовой поддержке РФФИ (проект №17-07-01600).

ЛИТЕРАТУРА

1. Garanina N., Sidorova E., Kononenko I., Gorlatch S. Using Multiple Semantic Measures For Coreference Resolution In Ontology Population, // International Journal of Computing, 2017, Volume 16, Issue 3.

2. Garanina N., Sidorova E. Ontology Population as Algebraic Information System Processing Based on Multi-agent Natural Language Text Analysis Algorithms // ISSN 0361-7688, Programming and Computer Software, 2015, Vol. 41, No. 3. © Pleiades Publishing, Ltd., 2015.

3. Garanina N.O., Sidorova E.A. and Seryi A.S. Multiagent Approach to Coreference Resolution Based on the Multifactor Similarity in Ontology Population // ISSN 0361-7688, Programming and Computer Software, 2018, Vol. 44, No. 1, pp. 23–34. © Pleiades Publishing, Ltd., 2018.

4. Garanina N., Sidorova E., Anureev I. Conflict resolution in multi-agent systems with typed relations for ontology population // Programming and Computer Software, July 2016, Volume 42, Issue 4, pp. 31–45.

5. Dodds L., Davis I. Linked data patterns. Режим доступа: http://patterns.dataincubator.org/book/ (Дата обращения 06.11.2018).

УДК 81’33

APPLICATION OF KEA FOR SEMANTICALLY ASSOCIATED STRUCTURAL UNITS SEARCH IN A CORPUS AND TEXT SUMMARIZATION

E. V. Sokolova
Saint Petersburg State University, Saint Petersburg
st049868@student.spbu.ru

This paper presents results of the research on possible applications of keyphrase extraction algorithm KEA. Although this algorithm is widely used as an effective and universal tool for keyphrase extraction, our study is aimed at exploration of its further adjustments in the tasks of translation equivalents search and for semantic compression, namely, for extractive summarization. To be precise, in our first series of experiments we analyzed the output of KEA based on the text corpus developed from the United Nations documents in order to find semantically associated structural units (possible translation equivalents) among Russian and English keyphrases. The second series of experiments concerned with using keyphrases automatically extracted by KEA to compose extracts for short stories. In this case we also compiled a corpus of short stories written in (or translated into) Russian and adjusted KEA so that ranked sentences with keyphrases could be used to form previews for the stories.

Keywords: keyphrase extraction, KEA, translation equivalents, summarization.

ПРИМЕНЕНИЕ АЛГОРИТМА KEA ДЛЯ ПОИСКА СЕМАНТИЧЕСКИ СВЯЗАННЫХ СТРУКТУРНЫХ ЕДИНИЦ В КОРПУСЕ И ДЛЯ АВТОМАТИЧЕСКОГО РЕФЕРИРОВАНИЯ ТЕКСТОВ

Е. В. Соколова
Санкт-Петербургский государственный университет, Санкт-Петербург
st049868@student.spbu.ru

В статье представлены результаты исследования возможных применений алгоритма автоматического извлечения ключевых слов и словосочетаний KEA. Алгоритм уже давно широко используется в качестве эффективного и универсального инструмента для автоматического извлечения

ключевых слов и словосочетаний, наша же работа направлена на исследование его применимости в задачах поиска переводных эквивалентов, а также для семантической компрессии текста, а именно автоматического реферирования экстрактивного типа. В первом эксперименте, представленном в данной статье, мы проанализировали извлечённые KEA слова и словосочетания на русском и английском языках из корпуса документации Организации объединённых наций на предмет семантически связанных структурных единиц (возможных переводных эквивалентов). Второй эксперимент посвящён использованию автоматически извлечённых KEA ключевых слов и словосочетаний при составлении аннотаций к коротким рассказам на русском языке. Для этого эксперимента мы также собрали корпус коротких рассказов, написанных на русском языке (или переведённых на него), и использовали KEA для системы ранжирования предложений из текста, которые могут быть включены в аннотацию к нему.

Ключевые слова: извлечение ключевых слов и словосочетаний, KEA, переводные эквиваленты, автоматическое реферирование.

Ключевые слова и словосочетания чаще всего понимают как структурные единицы, содержащие наиболее важную информацию о содержании текста. На сегодняшний день они имеют довольно широкое применение в различных областях, начиная от библиотечных систем и заканчивая системами автоматического реферирования текстов, и призваны формировать у пользователя общее представление о единичном документе или их коллекции.

Данная работа посвящена исследованию одного из алгоритмов автоматического извлечения ключевых слов и словосочетаний из текста – KEA (Keyphrase Extraction Algorithm). KEA является гибридным алгоритмом, объединяющим методы статистического анализа текста, методы машинного обучения, а также лингвистические модули обработки естественного языка. Наше исследование направлено на поиск и изучение новых возможных применений KEA в области NLP (Natural Language Processing), а именно поиск семантически связанных структурных единиц среди извлечённых KEA ключевых слов и словосочетаний из текстов на русском и английском языках, а также автоматическое создание аннотаций для коротких рассказов на русском языке.

В ходе первого эксперимента был собран корпус из 60 текстов, представляющих собой разного рода документацию Организации объединённых наций (ООН), 30 из которых написаны на английском, а оставшиеся 30 являются их переводами на русский (либо наоборот, в зависимости от языка оригинала). 25 документов из каж-

дого подкорпуса были взяты для обучения, оставшиеся пять вошли в тестовую выборку. Для каждого документа из тестовой выборки с помощью KEA автоматически были получены 20 ключевых слов и словосочетаний, после чего итоговые списки вручную анализировались на предмет наличия в них переводных эквивалентов. В результате эксперимента оказалось, что процент переводных эквивалентов среди общего числа извлечённых ключевых слов и словосочетаний для нашего корпуса превысил 60%, что доказывает лингвонезависимость KEA и его пригодность для решения подобных задач.

В рамках второго эксперимента был собран корпус из 30 коротких рассказов на русском языке, 25 из которых использовались в обучающей, а оставшиеся 5 – в тестовой выборке. Из каждого рассказа тестовой выборки было извлечено по 20 ключевых слов и словосочетаний. Для непосредственного создания аннотации (в данном случае, реферата экстрактивного типа, т. е. набора наиболее содержательных отрывков из оригинального текста в их первоначальном виде) был разработан и реализован алгоритм на Python, который осуществляет поиск извлечённых ключевых слов и словосочетаний в оригинальном тексте и извлекает предложения, в которых они содержатся. В ходе выделения наиболее значимых предложений им назначались веса, а затем первые четыре (и первое предложение из оригинального рассказа для завязки) с весом равным 2 и более, формировали итоговую аннотацию. Для определения качества полученных результатов привлекалась экспертная оценка, которая показала, что, как правило, показатель качества итоговой аннотации выше среднего.

Таким образом, в результате данного исследования мы предложили и протестировали два новых возможных применения KEA. Оказалось, что алгоритм справляется с предложенными задачами и показывает приемлемые результаты.

ЛИТЕРАТУРА

1. Keyphrase Extraction Algorithm [сайт]. URL: http://www.nzdl.org/Kea/index.html (дата обращения: 27/04/2017).

2. Kaur J., Gupta V.: Effective Approaches For Extraction Of Keywords // IJCSI International Journal of Computer Science Issues. Vol. 7. Issue 6. November 2010. URL: http://www.ijcsi.org/papers/7-6-144-148.pdf (дата обращения: 27.04.2017).

3. Beliga S.: Keyword extraction a review of methods and approaches.

URL: http://langnet.uniri.hr/papers/beliga/Beliga_KeywordExtraction_a_review_of_methods_and_approaches.pdf (дата обращения: 27.04.2017).

4. Соколова, Е. В., Митрофанова, О. А.: Автоматическое извлечение ключевых слов и словосочетаний из русскоязычных текстов с помощью алгоритма KEA. // Труды XX Международной объединенной научной конференции «Интернет и современное общество», IMS-2017, СПб., 2017.

5. Nenkova, A., McKeown, K.: Automatic summarization. // Foundations and Trends in Information Retrieval, vol. 5, № 2–3, pp. 103–233 (2011).

6. Witten, I.H., Paynter G.W., Frank E., Gutwin C., Nevill-Manning C.G.: KEA: Practical Automated Keyphrase Extraction. // Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, IGI Global, pp. 129–152 (2005).

7. The United Nations [сайт], URL: http://www.un.org/ru/index.html (дата обращения: 2018/05/27)

8. RNC [сайт], URL: http://www.ruscorpora.ru/ (дата обращения: 2018/05/27).

9. Kazantseva, A., Szpakowicz, S.: Summarizing short stories. // Computational Linguistics, vol. 36, № 1, pp. 71–109 (2010).

10. Luhn, H.P.: The automatic creation of literature abstracts. // IBM Journal of research and development, vol. 2, № 2, pp. 159–165 (1958).

11. Hovy, E., Lin, C.Y.: Automated text summarization and the SUMMARIST system. // Proceedings of a workshop on held at Baltimore, Maryland: October 13–15, 1998, Association for Computational Linguistics, pp. 197–214 (1998).

12. FantLab [сайт], URL: https://fantlab.ru/ (дата обращения: 2018/05/27).

13. Korobov, M.: Morphological Analyzer and Generator for Russian and Ukrainian Languages. // Analysis of Images, Social Networks and Texts, pp 320–332 (2015).

УДК 81’322.2

STUDYING TEXT COMPLEXITY IN RUSSIAN ACADEMIC CORPUS WITH MULTI-LEVEL ANNOTATION

V. D. Solovyev, M. I. Solnyshkina, V. V. Ivanov, A. V. Danilov
Kazan Federal Univaersity, Kazan; Innopolis University, Kazan
maki.solovyev@mail.ru

The problem of compiling a large multi-level annotated corpus of Russian academic texts was sparked by the demand to measure complexity (difficulty) of texts assigned to certain grade levels in terms of meeting their cognitive and linguistic needs Measuring text complexity called for linguistic annotations at various language levels including POS-tags, dependencies, word frequencies.

Keywords: annotated corpus, Russian language, academic texts, text complexity.

ГЛУБОКО АННОТИРОВАННЫЙ КОРПУС ДЛЯ ИЗУЧЕНИЯ СЛОЖНОСТИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ

В. Д. Соловьев, М. И. Солнышкина, В. В. Иванов, А. В. Данилов
Казанский федеральный университет, Казань;
Университет Иннополиса, Казань
maki.solovyev@mail.ru

Статья посвящена описанию глубоко аннотированного корпуса учебных текстов на русском языке. Корпус создан для решения задачи автоматического измерения сложности текста. Для этого требуется лингвистическая аннотация на разных уровнях языка, включая морфологический, синтаксический, лексический.

Ключевые слова: аннотированный корпус, русский язык, учебные тексты, сложность текста.

В этой статье мы представляем текущий проект, выполненный в Казанском федеральном университете, направленный на сбор и аннотирование корпуса российских учебных текстов. Насколько нам известно, спорадические исследования читаемости текста на русском языке не затрагивали учебные материалы. Цель нашего проекта – представить многоуровневый аннотированный корпус российских учебных текстов как для исследования собственно сложности текстов, так и для дискурсивных и когнитивных исследований.

Самые ранние исследования по удобочитаемости текстов, начиная с конца 19-го века, в основном были направлены на разработку формул сложности и использовали ограниченное количество количественных признаков: средняя длина предложения, средняя длина слова и частота слов. Данный подход подвергался критике ввиду использования слишком малого числа параметров. Современные методы автоматической обработки естественного языка позволяют получать многие лексические и синтаксические признаки текста.

При этом параметры могут быть как универсальными, так и языковоспецифическими. Это означает необходимость для русского языка выделения и проверки релевантных признаков в корпусе значительного размера. В настоящее время исследователи рассматривают многочисленные параметры, как когнитивно значимые: средняя длина предложения, количество абстрактных слов, число омонимов, количество технических терминов и т. д. К сожалению, в настоящее время отсутствуют корпусы ученых текстов на русском языке.

Мы использовали два набора учебников по общественным наукам, написанных для средней школы под редакцией Боголюбова и Никитина. Общий размер корпуса из 14 учебников – более 60 тыс. слов. Воспроизводимость результатов обеспечивается представлением корпус на веб-сайте проекта. Обычная предварительная обработка включала токенизацию и разделение текста на предложения.

На этапе предварительной обработки мы исключили все чрезвычайно длинные предложения (длиной более 120 слов), а также слишком короткие предложения (короче 5 слов). Чрезвычайно короткие предложения чаще всего появляются в виде имен глав и разделов книг или в результате неправильного расщепления предложения. Мы опускаем эти предложения, потому что средняя длина предложения является очень важной особенностью в оценке сложности текста и, следовательно, не должна быть предвзятой из-за ошибок предобработки.

Все аннотации в корпусе выполняются на трех уровнях: текстовом уровне, уровне предложений и уровне слова. Мета-аннотации текстового уровня включают число предложений и слов (точнее, токенов), автора и уровень сложности текста (номер класса). На уровне слов присутствует разметка частей речи и морфологические характеристики слова, полученные с использованием TreeTagger.

Мы также аннотируем каждую лемму в корпусе ее относительной частотой согласно частотному словарю, построенному по НКРЯ. На уровне предложения корпус содержит синтаксическую

разметку в формате XML. Разметка выполняется с помощью анализатора ЭТАП-3 в рамках модели синтаксических зависимостей. Набор тегов доступен на веб-сайте проекта, статистические данные собраны в таблицы.

Планируется добавление семантических аннотаций, путем связывания с синсетами большого электронного русского тезауруса RuWordNet. В ряде статей авторов корпус используется для разработки и корректировки формул удобочитаемости текстов на русском языке. Но даже очень простые статистические данные могут быть полезны при изучении сложности текста. Например, можно увидеть, что среднее количество уникальных прилагательных возрастает с номером класса. В то же время среднее число глаголов (а также наречий) уменьшается.

Исследование читабельности текстов на русском языке представляется актуальным по следующим причинам. Исследование международной организации PISA показало средний уровень понимания российскими старшеклассниками прочитанных текстов – Россия заняла 27 место среди 70 обследованных стран. Также ряд исследований свидетельствуют о том, что русские школьники не проявляют интереса к чтению, что вызвано неудачным подбором учебных материалов. Корпус является ценным инструментом для дискурсивных исследований, поскольку его данные обеспечивают прочную основу для сравнительного квантитативного исследования современных русских текстов.

Благодарности. Работа выполнена при поддержке гранта РНФ «Сложность текстов на русском языке», № 18-18-00436.

ЛИТЕРАТУРА

1. Kompetentnostnyy podkhod v vysshem professionalnom obrazovanii (2012) (pod redaktsiyey A.A. Orlova, V.V. Gracheva, Tula, 2012. 261 s.

2. Berezhkovskaja E. Problema psihologicheskoj negotovnosti k poluchenіju vysshego obrazovanija u studentov mladshih kursov (2017). M.: prospect. 62 p.

3. Britton, B.K., & Gulgoz, S. (1991). Using Kintsch's computational model to improve instructional text: Effects of repairing inference calls on recall and cognitive structures. Journal of Educational Psychology, 83, 329–404.

4. Chall J., Dale E. (1995) Readability revisited: The new Dale-Chall readability formula. Brookline Books.

5. Coleman M., Liau T. L. (1975) A computer readability formula designed for machine scoring. Journal of Applied Psychology, 60:283284.

6. Cornoldi, C., & Oakhill, J. (1996) (Eds.). Reading comprehension difficulties:Processes and intervention. Hillsdale, NJ: Erlbaum.

7. Crossley, S. A., Allen, L. K., Kyle, K., & McNamara, D. S. (2014). Analyzing discourse processing using a simple natural language processing tool (SiNLP). Dis- course Processes, 51, 511–534.

8. Dzmitryieva A. (2017) Iskusstvo yuridicheskogo pis'ma: kolichestvennyy analiz resheniy Konstitutsionnogo Suda Rossiyskoy Federatsii [The art of legal writing: a quantitative analysis of the Russian Constitutional Court rulings]. Sravnitel'noe konstitutsionnoe obozrenie, no.3, pp.125133. (In Russian).

9. Heilman M., Thompson K. C., Callan J., and Eskenazi M. (2007) Combining lexical and grammatical features to improve readability measures for first and second language texts. In Human Language Technologies 2007: The Conference of the North American Chapter of the Associa298 tion for Computational Linguistics (HLT-NAACL-07), pages 460467, Rochester, New York

10. Jackson, G. T., Guess, R. H., & McNamara, D. S. (2009). Assessing cognitively complex strategy use in an untrained domain. In N. A. Taatgen, H. van Rijn, L. Schomaker, & J. Nerbonne (Eds.), Proceedings of the 31st Annual Meeting of the Cognitive Science Society (pp. 2164–2169). Amsterdam, The Netherlands: Cognitive Science Society.

11. Ivanov V.. K voprosu o vozmonosti ispolzovanija lingvisticeskix xarakteristik slonosti teksta pri issledovanii okulomotornoj aktivnosti pri ctenii u podrostkov [toward using linguistic profiles of text complexity for research of oculomotor activity during reading by teenagers]. Novye issledovanija [New studies], 34(1):4250, 2013.

12. Karpov N., Baranova J., and Vitugin F. (2014). Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST).

13. Kincaid J. P., Fishburne R. P. Jr., Rogers R. L., and Chissom B. S. (1975). Deriva- tion of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Research Branch Report 8-75, Naval Technical Training Command, Millington, TN.

14. Krioni N., Nikin A., and Fillipova A.. Avtomatizirovannaja sistema analiza slozhnosti uchebnyh tekstov [the automated system of the analysis of educational texts complexity]. Vestnik UGATU (Ufa), 11(1):28, 2008.

15. McNamara, D.S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. Canadian Journal of Experimental Psychology, 55, 51–62.

16. McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. Cognition & Instruction, 14, 1–43.

17. Obobroneva I.. Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov. M.: RAS Institut soderzhaniya i metodov obucheniya, 2006.

18. Okladnikova S.. Modelkompleksnoj ocenki citabelnosti testovyx materialov na etape razrabotki [a model of multidimensional evaluation of the readability of test materials at the development stage]. Prikaspijskij urnal: upravlenie i vysokie texnologii, 3:6371, 2010.

19. Popova Ja.I., Shishkevich E.V. Standartizacija uchebnoj literatury srednej shkoly po kriteriju udobochitaemosti In Sevastopol'skij nacional'nyj universitet jadernoj jenergii i promyshlennosti // Nauchnye vedomosti BelGU. Ser. Gumanitarnye nauki. 2010. 12. Vyp. 6. S. 142147.

20. Shpakovskiy Y et al. Otsenka trudnosti vospriyatiya i optimizatsiya slozhnosti uchebnogo teksta. PhD thesis, 2007.

21. Solnyshkina M.,. Harkova E, and Kiselnikov A. Comparative coh-metrix analysis of reading comprehension texts: Unified (Russian) state exam in english vs cambridge first certificate in english. English Language Teaching, 7(12):65, 2014.

22. Ozuru, Y., Rowe, M., O'Reilly, T., & McNamara, D. S. (2008). Wheres the difficulty in standardized reading tests: The passage or the question? Behavior Research Methods, 40, 1001-1015.

23. Reynolds R. (2016) Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories, San Diego, CA: 16 June 2016. In: Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications, pp.289300.

24. Sinclair, J. (1997) Corpus Evidence in Language Description, in A. Wichmann, S. Fligelstone, T. McEnery and G. Knowles (eds.) Teaching and Language Corpora (London/New York: Longman), 27–39.

25. Solovyev V., Solnyshkina M, Ivanov V., Timoshenko S/ (2018) Complexity of Russian Academic Texts as the Function of Syntactic Parameters (in press).

26. Karpov N., Baranova J., and Vitugin F.. Single-sentence readability prediction in Russian. In International Conference on Analysis of Images, Social Networks and Texts, pages 91100. Springer, 2014.

27. e statisticheskih parametrov [Tekst] / L. V. Ustinova, L. S. Fazylova // araandy untini habarshysy. Matermatika ser. = Vestnik Karagand.un-ta.Ser. Matematika. – 2014. – 1. – S. 96–103.

28. Loukachevitch, N. V., Lashevich, G., Gerasimova, A. A., Ivanov, V. V., & Dobrov, B. V. (2016). Creating Russian Wordnet by conversion. Kompjuternaja Lingvistika i Intellektualnye Tehnologii, 15, 405-415.

УДК 81.33; 811.512.145

LEXICAL AND GRAMMATICAL POTENTIAL OF TURKIC LANGUAGES FOR THE DEVELOPMENT OF NEW INFORMATION PROCESSING TECHNOLOGIES

Dzhavdet Suleymanov, Dilyara Yakubova
Kazan Federal University, Kazan, Russia
Institute of Applied Semiotics of Tatarstan Academy of Sciences, Kazan, Russia
dvdt.slt@gmail.com, suleymanovad@gmail.com

This article presents the results of studying the lexical and grammatical characteristics of the Turkic languages on the example of the Tatar language. These features are of particular methodological and practical interest for the development of software for the efficient processing of information in natural language. they reduce the time and memory for processing and storing information, and also provide for the coding and processing of fuzzy information. The authors argue that consideration of these properties when developing new technologies for processing knowledge will increase their efficiency and intelligence.

Keywords: technological aspect of natural languages · affixal morphemes · morphotactics · knowledge activeness · morphological ellipsis · recursion

ЛЕКСИКО-ГРАММАТИЧЕСКИЙ ПОТЕНЦИАЛ ТЮРКСКИХ ЯЗЫКОВ ДЛЯ РАЗВИТИЯ НОВЫХ ТЕХНОЛОГИЙ ОБРАБОТКИ ИНФОРМАЦИИ

Джавдет Сулейманов, Диляра Якубова
Казанский федеральный университет, Казань, Россия
Институт прикладной семиотики Академии наук Республики Татарстан, Казань, Россия
dvdt.slt@gmail.com, suleymanovad@gmail.com

В данной статье представлены результаты изучения лексико-грамматических характеристик тюркских языков на примере татарского языка. Эти особенности представляют определенный методологический и практический интерес для разработки программного обеспечения для эффективной обработки информации на естественном языке. Показано, что учет этих характеристик позволяет уменьшить время и память для обработки и хранения информации, в том числе нечеткой информации.

Ключевые слова: technological aspect of natural languages · affixal morphemes · morphotaksika · activity of knowledge · morphological ellipsis · recursion.

1. Introduction

Currently, it has become apparent not only for professionals in the field of Computer Science, but also for ordinary Internet users that the modern means of knowledge accumulation and processing in natural language can barely cope with such tasks as fast search and selection of relevant information in distributed databases, knowledge extraction and semantic analysis of textual information. One reason for this is that modern information and communication technologies are inherently unintelligent. They are based on primitive artificial programming languages that in practice present a subset of synthetic and analytic languages or artificial structures based on them.

Therefore, research of natural language potential for the creation of new programming languages and systems is promising and has great potential for the development of intelligent information processing technologies.

It is most productive to study natural languages in the following three aspects: cognitive, communicative and technological. The cognitive aspect of the language characterizes it in terms of conceptualizing reality and refers to its capability for describing a picture or model of the world and for explicit knowledge representation. The communicative aspect of the language reflects the ability of a natural language to encode, receive, transmit and organize a dialogue. The technological aspect of the language determines its formal and conceptual potential to implement means for efficient processing, adequate description and compact storage of information on this language. It refers to the capability of the natural language to create ergonomic technical equipment and technologies that take into account the language specifics (for example, the frequency of the alphabet letters in keyboard design) and to develop intelligent software tools and software, including operating systems.

Obviously, at the basis of artificial languages and programming systems lie deep structures that represent the mentality of the natural language, and therefore these systems implement the descriptive and computing capacity of the corresponding natural language.

In systems of knowledge processing, the following characteristics are important and determine their efficiency and intelligence: 1) information processing time; 2) memory for information storage; 3) availability of means for information "compression" and its compact storage; 4) possibility for coding and processing of fuzzy information; 5) knowledge activeness. Besides, the first three parameters determine the effectiveness

whereas the parameters 4 and 5 refer to the intelligence of systems and technologies.

This article dwells on the technological aspect of the Tatar language and reveals a number of features that determine the effectiveness of the lexical and grammatical model of the Tatar language in terms of the development of intelligent information processing systems. Tatar language belongs to the Turkic group of languages, being one of two official languages in the Republic of Tatarstan and the second largest language in Russia after the Russian language according to the number of its native-speakers.

2. Regularity and Natural Complexity of Tatar Morphology

Studies show that the Tatar morphology is a regular and almost automaton [1, 2] formation, which has natural complexity and provides efficient coding of complex information in the word structure. It is known that automaton grammars have minimum characteristics of time and space functions, which implies that generation and processing of information takes much less time and less intermediate storage than with context-free and context-sensitive grammars. Let us consider in more detail a number of features that characterize the regularity and natural complexity of the Tatar language morphology on the example of nominal and verbal forms.

The regularity of morphology means that the same scheme of morphemic combination (morphotactics) is inherent to all or nearly all nominal and verbal forms. This feature allows almost automate creating of word-forms with the same deep affixal meanings using the same scheme.

An important property of Tatar morphology, along with its regularity, is the fixed position of affixes in the sequence of affixal morphemes, as well as a clear connection between the position of an affix and its type. In nominal and verbal paradigms, adding an affix on the right is directly determined by the previous affix. Thus, both properties of automaton or regular grammar are present: the right-linear property and the property of "short memory" [3].

For example,

1. *At, atlar, atlarım, atlarıma* ('horse, horses, my horses, to my horses')

2. *Kul, kullar, kullarım, kullarıma* ('hand, hands, my hands, to my hands')

In these examples, nominal root morphemes *at* ('horse') and *kul* ('hand') have the same sequences of affixal morphemes with identi-

cal meanings. They can be described by the same scheme: X(Noun), X(Noun)+lar(Plural affix), X(Noun)+lar(Plural affix)+ım(Possessive affix, 1 person, Singular), X(Noun)+lar(Plural affix)+ım(Possessive affix, 1 person, Singular)+a(Dative case affix).

Positions of affixal morphemes that compose a word-form have relation to certain types of morphemes and are mutually invariable. Affixal morphemes of a certain type can appear only in a certain position or they fall out together with the position. Such possibility allows determining the presence or absence of some feature or property of the described meaning (plurality, modality, recurrence and others) basing on a corresponding position. It is another technological feature to consider in order to increase the quality of information processing.

The following morphotactics corresponds to nominal word-forms: <Nominal wordform>::=<STEM> [<plurality>][<possessiveness>] [<case>][<modality>]. In this scheme, the <STEM> is either a root morpheme, or a root morpheme plus derivational affix, or a word-form created by adding a certain type of affixes that transform a word-form into a nominal form (e.g. case-like possessive affix *-nıkı*: *apanıkı* 'aunt's'). [<Plurality>] corresponds to the type of affixes that encode the plurality of meanings: *{-LAr}* (hereinafter affixes in capital letters mean variation, for example, *-LAr*: *apalar* 'aunts', *enelər* 'brothers', *uramnar* 'streets'). [<Possessiveness>] is the type of affixes that determine the belonging of the word-form's meaning: {Im} (e.g. *apam* 'my aunt', *ulım* 'my son'). [<Case>] is the type of case and case-like affixes (e.g. *urman* [Nominative case] – 'forest', *urmanga* [Directive case] 'to the forest', *urmanda* [Locative case] 'in the forest', *urmannı* [Accusative case] 'the forest', etc.). [<Modality>] is the type of affixes that reflects the emotional and subjective attitude: question, statement, doubt, admiration, exclamation, etc. (e.g. *urmanmını* 'really a forest?', *urmandır* 'probably a forest, undoubtedly a forest', *urmanmı* 'a forest?').

Examples of nominal forms: *kitap* ('book' – Noun), *kitaplar* ('books') – *kitap* ('book' – Noun) + *lar* (Plural affix), *kitabım* ('my book') – *kitap* ('book' – Noun) + *ım* (Possessive affix, 1 person, Singular), *kitapmı?* ('a book?') – *kitap* ('book' – Noun) + *mı* (Question affix), *kitaplarımnanmını? (kitap-lar-ım-nan-mını?)* ('really from my books?') – *kitap* ('book' – Noun) + *lar* (Plural affix) + *ım* (Possessive affix, 1 person, Singular) + *nan* (Ablative affix) + *mını* (Question affix, surprise), *kitaplarımdırmı* ('whether these are my books') – *kitap* ('book' – Noun) + *lar* (Plural affix) + *ım* (Possessive affix, 1 person, Singular) + *dır* (Modality affix, doubt)+ *mı* (Question affix).

Another characteristic feature of Tatar grammar is that an affixal morpheme (or sequence of morphemes) that joins on the right a nominal word-form, which is the most right constituent in the sequence of word-forms, corresponds to the entire nominal group. This follows from the description of morphotactics and can be used as a technological feature when creating a model of text processing.

For example,

1. *Balaçaqtan yaratqan kitaplarım* 'the books I love from the childhood' -(Balaçaqtan yaratqan kitap)+lar+ım

2. *Balaçaqtan yaratqan kitabıbız* 'the book we love from the childhood' -(Balaçaqtan yaratqan kitap)+ ı+bız

In the Tatar language, the properties of morphotactic regularity and fixed positions of corresponding types of affixal morphemes are also inherent to verbal groups. The following morphotactics corresponds to verbal word-forms: <Verbal form>::=<STEM>[<voice>][<negation>] [<time>] [<person>][<modality>].

In this scheme, <STEM> is defined as a verbal root morpheme in its lemma form.

For example, *kal* ('stay' – verb), *kaldir* ('do so that X stay(s)') – *kal* ('stay') + *dır* (Causative voice), *kaldırma* ('do so that X do/does not stay) – *kal* ('stay') + *dır* (Causative voice) + *ma* (negation), *kaldirdiğiz* ('you did so that X stayed') – *kal* ('stay') + *dır* (Causative voice) + *dı* (past tense) + *ğız* (3 person, plural), *kaldirmadigiz* ('you did not do so that X stayed)' – *kal* ('stay') + *dır* (Causative voice) + *ma* (Negation) + *dı* (past tense.) + *ğız* (3 person, plural), *kaldırmadığızmını* ('didn´t you do so that X stayed') – *kal* ('stay') + *dır* (Causative voice) + *ma* (negation) + *dı* (past tense) + *ğız* (3 person, plural) + *mını* (question, surprise).

Just as in the case with the nominal group, the sequence of morphemes that describes the meanings of a role situation encoded by a verbal group is determined for a verbal word that occupies the most right position in the sequence of word-forms in a verbal group.

For example, in the verbal group: *çabıp barıp qarap alıp kayttığızmı?* (literally, 'having run + having gone + having looked + having taken did you return?') – the sequence of affixes *-tı+ğız+mı* is added to the last verbal form *kayt* (imperative, 2 person, singular) (*'return'*). It acts as a certain out-of-bracket chain that finishes a verbal group and refers to the entire verbal group.

Therefore, in semantic explication of this expression it is correct to represent the verbal group (e.g. in text annotation for machine applications) using the following bracket record: (*çabıp* 'having run' *barıp* 'hav-

ing gone' *qarap* 'having looked' *alıp* 'having taken' *qayt* 'return') + *tı* (Past tense) + *ğız* (3 person, Plural) + *mı* (Modal, question)?

Let us consider a number of features that determine natural complexity of Tatar morphology:

1. The possibility of joining to a word-form of certain affixal morphemes, which change the word type, e.g. convert a nominal word-form into a verbal or adjective form, and vice versa.

2. Morphological (synthetic) means of expressing modality, mood, emotional and personal attitude towards a situation, object or process described by a word-form.

3. Contextual variety of meanings of an affix.

It is known that a nominal group, as a rule, encodes a certain semantic role situation, whereas a verbal group encodes contextual relations beyond these roles. Thus, the possibility of switching from a nominal form to a verbal one and vice versa by joining the corresponding affixes allows describing simultaneously within a single word-form both complex role situation and contextual relations among semantic roles. This is how the compactness of description and information storage is provided. The synthetic or affixal method of inflexion permits the encoding of a certain meaning within a single word-form, which is described in inflected-analytical languages (for example, in English) by several word combinations and even sentences.

As an example of feature (1) realization, let us see the following word-form, which is correct for the Tatar language: *"Tatarçalaştırğalaştıruçıl ardağınıqılarğamını?"* ('Really (to those/on those) that belongs to (that) that is on those, who (that) from time to time do Tatar localization?'). It has the following structure: *Tatar* (Noun) + *ça* (Adverb) + *la* (Verb) + *ştır* (Verb, Voice) + *ğala* (Verb, Voice) + *ştır* (Verb, Voice) + *u* (Noun, Gerund) + *çı* (Noun) + *lar* (Plural) + *dağı* (Noun, Locative) + *nıqı* (Noun, Possessive) + *lar* (Plural) + (Directive) + *mını* (Question, surprise).

The ability to express modality with affix in the Tatar language is also a feature that contributes to adequate interpretation of the word meaning and minimizing of the time for its recognition. It is different in languages where this feature is expressed emotionally and prosodically or by means of additional word-forms.

The third feature of complexity of the Tatar morphology determines contextual variety of affixal meanings. Almost all affixal morphemes are polysemic. In particular, as shown in our study [2], the affixal morpheme *-GA* has about 20 meanings, so it is used to encode up to 20 different contextual meanings. Table 1 shows some examples with the affix *-GA*.

Table 1. **Contextual variety of the affix –*GA***

Coded value		Example
1. Performer of the process	+	*Ukituçığa hat ukıtu* ('to let the teacher read the letter'): the teacher reads the letter
2. Object	-	
2.1. Direct	-	
2.2. Indirect	+	*Ukuçığa kuşu* ('to give a task to the student')
3. Aim of the process	+	*Utınğa baru* ('to go for firewood')
4. Reason of performing of the process	+	*Maturlıkka soklanu* ('to admire beauty')
5. Time of performing of the process	+	*Biş atnağa kaitu* ('to come back for five weeks')
6. Place of performing of the process	-	
6.1. Final destination point	+	*Urmanğa baru* (′to go to the forest′)
6.2. Point of departure	-	
7. Way of performing of the process	+	*Barlık kəçkə çabu* (‘to run for dear life')
8. The extent of the process or action	-	
8.1. The extent of the process	+	*Ber atnağa kaitu* ('to come back for a week')
8.2. The extent of the object of the process	+	*Un sumğa alu* ('to buy for ten rubles')
9. Means of achieving the aim	+	*Aqçağa alu* ('to buy for money')

3. Morphological Ellipsis and Recursion in the Tatar Language

Other features of the Tatar morphology that increase the technological potential of the Tatar language are morphological ellipsis and recursion. Morphological ellipsis is the possibility of omitting a sequence of affixes in homogeneous nominal word-forms, maintaining it in the last word-form. Thus, there is a possibility of moving any sequence of affixes, which is common for homogeneous members, to the right, beyond the

sequence of homogeneous members, and joining them to the last on the right homogeneous member.

For example:

1. *Işek aldı tawıqlarğa, kazlarğa, sarıklarğa tulı = Işek aldı tawıq, kaz, sarıqlarğa tulı*. 'A court is full chickens, geese, and sheep'.

2. *Min kırlarıbızğa, urmannarıbızğa, yılgalarıbızğa şatlanam = Min kır, urman, yılğalarıbızğa şatlanam*. 'I am glad to our fields, forests, rivers'.

Recursion is the possibility of cyclic generation of a new meaning by successive application of the same "formula", in other words, a repeated joining of the same affix.

Such properties are possessed by affixal morphemes *–DAğI* (Locative 2, space-time case 2) and *–nIkI* (Possessive case) that can also be called ambiguity affixes as they provide the added lexemes with ambiguity.

For example, let us consider the lexeme *tau* ('mountain'). Joining of affix *–dağı* generates new objects or properties that are indefinite: *taudağı* 'something on the mountain'; *taudağıdağı* 'something on something on the mountain'; *taunıqı* 'something that belongs to the mountain'; *taunıqınıqı* 'something that belongs to something that belongs to the mountain'.

Following this formula, a word-form of practically any length can be created. Naturally, such long sequences of morphemes in normal speech are almost never used. In our opinion, it is mainly related to the problem of memory depth and convenience of communication among people. Nevertheless, such inflexion is correct from the point of view of the Tatar language grammar. Thus, a word-form created by joining a sequence of any length hypothetically always makes sense, while a concrete meaning is acquired when it is "immersed" in a certain context.

Let us consider the following word-form: *taunıkındağınıkınıkındağı*. It can be separated into the following constituents – *tau+nıqı+ +nda+ğı+nıqı+nıqı+ndağı* (Noun + Possessive + Locative2 + Possessive + Possessive + Possessive + Locative2). This word-form means the following: 'something situated on/in something belonging to something belonging to something situated on/in something belonging to the mountain'.

It is easy to notice that by explicitly setting parameters after every morpheme it is possible to achieve a contextual definiteness of a word-form. That is, when concrete meanings substitute indefinite affixes, the word-form also receives a concrete meaning. In real cases, the context of speech or discourse fills such parameters with a concrete meaning. It can be seen on the following example.

Let every indefinite affix be followed by the parameters: *tau+dağı*

(x1)+ndağı(x2)+nıqı(x3)+nıqı(x4)+ndağı(x5)+nıqı(x6), where *xi* are contextual objects, i.e. objects that acquire a concrete meaning either in a context or it is set by the user (i=1,6).

Giving value to parameters*: x1= məğərə 'cave', x2= ayu 'bear', x3= ayak 'paw', x4= tırnak 'claw', x5= bal 'honey'* – we get the following contextual meaning: 'something (value x6 remains indefinite) that is inherent to the honey that is on the claw that belongs to the paw that belongs to the bear that is situated in the cave that is located on the mountain'.

The place of the root morpheme can also be occupied by an indefinite parameter: *X+nıqı(x1)+ndağı(x2)+nıqı(x3)+nıqı(x4)+ndağı(x5)+nıqı (6)*. Here the place of X can be occupied by any concept, either set implicitly and exposed through a context, or set explicitly by the user. In the last example, *X=tau* ('mountain'). In the speech, the pronoun *ni* is often used for X, which performs as a metaword or intensional and acquires a concrete meaning in the context: *ni/dəğe/ndəğe/neke/neke/ndəğe /neke*.

Let us consider the manifestation of the recursion feature on the example of entire sentences. *Kır kuyanı kolaklarındağı kara taplarda maturlyk bar. Urman kuyanınıkılardağılarnıkınnan bashkarak*. ('There is beauty in black spots on the ears of field hares. It is a little different from the beauty of the black spots on the ears of forest hares'). In the word form *kuyanınıqılardağılarnıqınnan = kuyanı* ('his/her hare')+*nıqı*(x0)+*lar*(plural)+*dağı*(x1)+*lar*(plural)+*nıqı*(x2)+*nnan*(initial case) the set of concepts (x0, x1, x2) is given implicitly; however it becomes explicit in the previous context (i.e. on presupposition): x0 = *qolaq* ('ear'); x1 = *qara tap* ('black spot'); x2 = *maturlıq* ('beauty'). The second sentence, if expressed explicitly, has the following aspect: "*Urman quyanı qolaqlarındağı qara taplardağı maturlıqtan başqaraq"* ('It is a little different than the beauty of the black spots on the ears of a forest hare').

Even on this short example, simple calculation shows that use of recursion affixes brings to the compression of information and substantial memory economy. If recursion is used, the amount of words in the example grows more than two times short and the number of characters in-use diminishes on 23 (in the variant without recursion: 7 words, 64 signs; in the variant with recursion: 3 words, 41 sign). Thus, the context provides quite a simple explication of ambiguities.

4. Fuzziness in description of commands and actions

It is known that, as a rule, lexical description of predicates (commands, actions, relations) is carried out by verbal word-forms. Let us consider

the following two features of technological effectiveness that reflect the natural cognitive mechanisms manifested in verbal word-forms:

1. Possibility to set fuzzy commands and to describe ambiguous actions and relations among objects by means of recursion.

2. Possibility to describe the actions that relate to the entire role situation within one word-form by means of recursion.

The property (1) is encoded by verbal affixes that occupy the voice position, i.e. immediately after the verbal stem, – *GAlA*, – *çtIr*.

For example: *yu* ('wash' – 3 person, Singular, Imperative)*, yuğala* ('wash from time to time') – *yu* ('wash') + *ğala* ('from time to time')*, yugalaştır* ('wash from time to time, from time to time – rarer') – *yu* ('wash') + *ğala* ('from time to time') + *ştır* ('from time to time'), *yuğalaçtırğala* ('wash from time to time, from time to time, from time to time – even rarer') – *yu* ('wash') + *ğala* ('from time to time') + *ştır* ('from time to time') + *ğala* ('from time to time')*, yuğalaçtırğalaçtırğala* ('wash from time to time, from time to time, from time to time – and even rarer') – *yu* ('wash') + *ğala* ('time from time – rare') + *ştır* ('time from time – rarer') + *ğala* (yet rarer) + *shtyr* (yet rarer) + *ğala* (yet rarer).

The degree of rarity of washing required is determined by contextual information or by the model of the world. For example, *futbolkanı yuıştırğala* ('wash the T-shirt from time to time') may imply the command: wash the T-shirt after putting it on several times, while *galstuknı yuıştırğala* ('wash the tie from time to time') is more likely to imply: wash the tie once a year or even rare, when needed.

The property (2) is exteriorized through a number of special verbal affixes that can occupy a voice position: *-n, -Iş, -t, -DIr*.

The following example with the verbal word-form *tashla* ('throw') describes the change in the role situation caused by joining of corresponding affixes.

Participants of the action: subject S, object-item O_k, where $k \geq 1$.

The role situation for the word-form *tashla* ("throw"): S influence O_k.

Adding of the affixes *-n, -Iş, -t, -DIr* results in changes that are described below.

1. *–n*: *taşlan – taşla+n* ("throw yourself")

Role situation: S influence S (reflection)

2. *–Iş*: *taşlaş – taşla+ş* ('help to throw/throw together')

Participants of the action: subject S, object-actor $A_{i,j}$, object-item O_k, where i is the number of the object-actor group, $i \geq 1$; j is the number of participants in the group i, $j \geq 1$.

Role situation: S influence (help) $A_{i,j}$ and (S & $A_{i,j}$) influence (throw) O_k.

3. *-t, -DIr*:

a) *taşlat – taşla+t* (‘do so that they throw’)

Role situation: S influence $A_{i,j}$ -> $A_{i,j}$ influence (throw) O_k. Here the arrow -> means implication.

b) *taşlattır – taşla+t+tır* (‘do so that they do so that they throw’)

Role situation: S influence $A_{i,j}$ -> $A_{i,j}$ influence $A_{l,m}$ -> $A_{l,m}$ influence (throw) O_k.

c) *taşlattırt – taşla+t+tır+t* (‘do so that do so that they do so that they throw’)

Role situation: S influence $A_{i,j}$ -> $A_{i,j}$ influence $A_{l,m}$ -> $A_{l,m}$ influence $A_{s,t}$ -> $A_{s,t}$ influence (throw) O_k.

This formula makes it possible to create new role situations every time and to describe processes at the lexical level by adding new concrete affixes. For example, adding the affix *-ıl* to the last word-form *taşlattırtıl* converts the subject into an object-item, object of influence, i.e. S= O_k.

The role situation in this case is as follows: S influence $A_{i,j}$ -> $A_{i,j}$ influence $A_{l,m}$ -> $A_{l,m}$ influence $A_{s,t}$ -> $A_{s,t}$ influence (throw) S.

5. Activeness of knowledge

It is known that sentences in the English language are built according to the scheme S-V-O (subject-verb-object), and in the Tatar language according to the scheme: S-O-V. That is, when English speakers tell of their intention to go to the cinema, they first say if they will or will not go, and only then give the information on where, how, why, with whom, when, etc. In the sentence: *“I’ll go to see the “Atilla” film with my friend in the afternoon,”* – the action controls the situation. After such obvious expression of subject’s intention, further information becomes passive and it does not influence the choice of action mode neither complicates it. On the contrary, in the Tatar language information and analysis come first, and only then, possibly taking into account the listener’s reaction, the action itself is determined – either positive or negative. *“Min dustım belən tөştən soñ bulası “Atilla” kinosna baram/barmıym”* – literally: ‘I with my friend in the afternoon to the film “Atilla” will go/will not go’.

In artificial intelligence systems, this feature receives the name of knowledge activeness, and it is one of the most important signs of system’s intelligence [2]. For intelligent systems, the following style of reflection is natural and fundamental: analysis-action, reflection-aims-algorithms, whereas modern technologies and programming languages and systems

that are based on the mentality of the English language the command style is implemented action-analysis, algorithm-aim. That is, the system based on the Tatar language mentality in order to describe a situation first analyses and processes information, and only then carries out the action, namely selects an adequate model of knowledge representation or determines the appropriate algorithms and implementation schemes.

6. Conclusions

Lexical and grammatical means for description, collection, storage, processing and transmission of information on the natural language are technological tools, which allow capturing, verbalizing and showing explicitly on the surface level those deep patterns, cognitive models and mechanisms that describe situations and processes in the "model the world".

On the example of the Tatar language, this article studies a number of lexical and grammatical features that determine the technological potential of Turkic languages and present methodological and practical interest for the creation of software for efficient processing of natural language information.

The compactness in the transmission of the meaning of the text on the lexical level is also due to the capability of the language to encode the meaning synthetically with a word-form, whereas in other languages, such as English and Russian, it is done analytically, often with several sentences. The agglutinative nature of the Tatar language, presence of algorithmic patterns and a powerful meta-system, minor exceptions and sufficient rigidity of the syntax allow setting the task of constructing a language of intermediate translation and developing new information processing technologies based on Turkic languages.

Acknowledgement. The reported study was funded by Russian Science Foundation according to the research project №16-18-02074.

REFERENCES

1. Suleymanov, D.Sh.: Formal elegance and natural complexity of the Tatar language morphology (in Russian). E-conference: Information technologies in humanities (Kazan, May 25–31, 1998). Kazan (1998).

2. Suleymanov, D.Sh., Gatiatullin, A.R.: Structural and functional computer model of Tatar morphemes (in Russian). Kazan, Fen, 220 p. (2003).

3. Gladkiy, A.V., Melchuk, I.A.: Elements of mathematical linguistics (in Russian). Moscow, Nauka. 192 p. (1969).

УДК 81'33

COMPARATIVE ANALYSIS OF ONTOLOGICAL CONCEPTS FOR DESCRIPTION OF GRAMMATIC CATEGORIES IN DIFFERENT TURKIC LANGUAGES

D. Sh. Suleymanov, A. R. Gatiatullin
Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan, Kazan
dvdt.slt@gmail.com, ayrat.gatiatullin@gmail.com

The article describes the work on the creation of an ontological model of the grammar of Turkic languages. At the current stage, the authors developed a thesaurus to describe the morphology of the Tatar language. In the process of preparing the thesaurus, a comparative analysis of the grammatical categories and notation systems of these categories in the Tatar, Kazakh, Kyrgyz, Uzbek and Turkish languages was conducted.

Keywords: grammatical categories, Turkic languages.

СРАВНИТЕЛЬНЫЙ АНАЛИЗ ОНТОЛОГИЧЕСКИХ КОНЦЕПТОВ ДЛЯ ОПИСАНИЯ ГРАММАТИЧЕСКИХ КАТЕГОРИЙ В РАЗНЫХ ТЮРКСКИХ ЯЗЫКАХ

Д. Ш. Сулейманов, А. Р. Гатиатуллин
Институт прикладной семиотики Академии наук Республики Татарстан, Казань
dvdt.slt@gmail.com, ayrat.gatiatullin@gmail.com

В статье описывается работа по созданию онтологической модели грамматики тюркских языков. Авторами на текущем этапе разрабатывался тезаурус для описания морфологии татарского языка. В процессе подготовки тезауруса был проведен сравнительный анализ грамматических категорий и систем обозначений этих категорий в татарском, казахском, киргизском, узбекском и турецком языках.

Ключевые слова: грамматические категории, тюркские языки.

На конференциях по компьютерной обработке тюркских языков TurkLang неоднократно обсуждались вопросы по разработке унифицированной морфологической разметки текстов на тюркских

языках для использования в корпусах и других системах автоматической обработки текста. Подобная унифицированная система разметки должна также служить в качестве универсального средства глоссирования текстовых примеров (например, в международных публикациях).

В рамках проекта AP05132249 «Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний» проводились работы по разработке метаязыка понятий морфологических правил тюркских языков. В рамках этого проекта велась работа параллельно для пяти языков: татарский, казахский, киргизский, узбекский, турецкий. Авторы статьи участвовали в разработке татарской части тезауруса.

Также нами был проведен сравнительный анализ как самих грамматических единиц и категорий, так и системы обозначений в этих языках для всех вышеперечисленных тюркских языков. В статье рассматриваются результаты сравнительного анализа.

1. Сингармонизм

Одно из основных различий в языковых единицах, в данном случае в морфемах, заключается в количестве алломорфов. Одним из факторов, влияющих на количество алломорфов, является закон сингармонизма по гласным звукам. Так, в узбекском языке нет чередования по гласным, и во всех алломорфах используется только один вид гласного звука.

Например, в таблице 1 приведены примеры алломорфов узбекских аффиксальных морфем.

Таблица 1

Язык	Морфема	Алломорфы	Категория
узбекский	-lar	lar	Plural
	-[i]m	im	POSS.1SG
	-Ga	ga	DIR

В морфемах татарского и казахского языков используются алломорфы с двумя типами гласных: твердая и мягкая (Таблица 2).

Таблица 2

Язык	Морфема	Алломорфы		Категория
татарский	-ЛАр	лар	лəр	Plural
казахский	-ЛАр	лар	лер	Plural
татарский	-[Ы]м	ым	ем	POSS.1SG
казахский	-[Ы]м	ым	ім	POSS.1SG
татарский	-[Г]А	га	гə	DIR
казахский	-[Ғ]А	ға	ге	DIR

А в морфемах турецкого и киргизского языков встречается 4 вида алломорфов с чередованиями гласных (Таблица 3).

Таблица 3

Язык	Морфема	Алломорфы				Категория
турецкий	-lar	lar	ler			Plural
киргизский	-ЛАр	лар	лер	лор	лөр	Plural
турецкий	-[I]m	ım	um	im	üm	POSS.1SG
киргизский	-[Ы]м	ым	им	ум	үм	POSS.1SG
турецкий	-[у]А	уа	уе			DIR
киргизский	-[Г]А	га	ге	го	гө	DIR

На примере в таблице 3 показано, что в турецком языке категории plural и DIR имеет только два варианта алломорфов, а в киргизском по четыре. Из этого примера видно, что 4 варианта алломорфов, различающихся по гласным звукам, также для разных языков реализованы по разному.

Разное количество алломорфов в морфемах, выражающих одни и те же грамматические категории в разных языках, определяют и разное количество правил сочетания алломорфов.

В том варианте тезауруса, который подготовлен на этом этапе проекта, пока отсутствуют концепты, позволяющие отображать морфонологические свойства морфем и законы сингармонизма. На

следующем этапе необходимо реализовать в тезаурусе возможность представлять такую информацию.

2. Грамматические категории

В разрабатываемом тезаурусе основными концептами являются концепты, представляющие морфологические единицы и грамматические категории. Все концепты в создаваемом тезаурусе классифицированы по частям речи. Сравнительный анализ показывает, что именные категории в тюркских языках, в отличие от глагольных, имеют намного больше сходства, однако среди них также есть ряд различий.

Одно из таких различий в именных категориях выражено в категории Инструментального падежа или Инструментатива. Так, в казахском и турецком языках инструментатив может выражаться, как аффиксально, так и с помощью послелогов. А в татарском, киргизском и узбекском языках он выражается только с помощью послелогов (Таблица 4).

Таблица 4

Категория	Татарский	Казахский	Киргизский	Узбекский	Турецкий
Инструментатив	белән	-[Б]ен: -бен, -мен, -пен	менен	bilan	-[y]lA: -yla,-yle

В качестве особенности киргизского языка, отличающей ее от других рассматриваемых тюркских языков, является категория вежливости, представленная в таблице 5.

Таблица 5

Категория	Морфема	Алломорфы			
POSS.2SG	-[Ы]ң	ың	иң	уң	үң
POSS.2SG.P	-[Ы]ңЫз	ыңыз	иңиз	уңуз	үңүз
POSS.2PL	-[Ы]ңАр	ңар	ңер	ңор	ңөр
POSS.2PL.P	-[Ы]ңЫзДАр	ыңыздар	иңиздер	уңуздор	үңүздөр

В татарском и узбекском языках существует категория DIR_LIM, которая выражена аффиксами -[Г]АчА в татарском языке и -Gacha в узбекском.

Таким образом, получается многоязычный тезаурус, в котором есть концепты классы, общие для всех языков и концепты экземпляры, которые присутствуют или отсутствуют в отдельном тюркском языке.

3. Системы обозначений

Кроме того, что грамматики разных тюркских языков отличаются наличием или отсутствием языковых единиц и категорий, они еще отличаются обозначением этих категорий в разных тюркских языках. Пример обозначений этих категорий представлен в таблице 6.

Таблица 6

Русский	Татарский	Казахский	Киргизский	Узбекский	Турецкий
Падеж	килеш	септік	мүчө	kelishik	hal
Залог	юнәлеш	етіс	мамиле	ovoz	çatı
Союз	теркәгеч	жалғаулық	байламтал	bog'lovchi	bağlaç

Как видно из таблицы 6, практически отсутствуют похожие названия одних и тех же грамматических категорий. Это показывает, что несмотря на большое совпадение грамматических категорий и грамматических единиц в тюркских языках, существует очень большая разница в обозначениях этих категорий. Такое различие указывает на то, что в период образования этих терминов процессы по изучению и описанию грамматик для тюркских языков развивались практически полностью автономно от других тюркских языков и не было заимствований терминов из другого тюркского языка.

Заключение

Проведенный анализ показывает, что создание единой онтологической системы разметки для электронных корпусов тюркских языков, а также представление в этой модели обозначений на разных тюркских языках, позволит сформировать «взаимопонимаемость», согласованность терминов, обозначающих идентичные грамматические категории.

Полученный многоязычный тезаурус, реализованный на языке представления онтологий OWL должен стать эффективным ресур-

сом для систем многоязычного поиска и систем машинного перевода между тюркскими языками.

Благодарности. Работа выполнена при финансовой поддержке проекта «AP05132249 «Разработка электронных тезаурусов тюркских языков для создания систем многоязычного поиска и извлечения знаний».

ЛИТЕРАТУРА

1. Ергеш Б.Ж., Шарипбай А.А., Гатиатуллин А.Р. Тезаурусы: обзор моделей, применение // Вестник КазНИТУ. – 2018. – №4 (128) – С. 208–212.

УДК 004.01, 81'367.628

ESTABLISHMENT AND DEVELOPMENT OF TERMINOVAISM ON INFORMATICS AND INFOCOMMUNICATION TECHNOLOGIES IN THE TATAR LANGUAGE

D. S. Sulejmanov, A. F. Galimjanov
Institute "Applied Semiotics" of the Academy of Sciences of the Republic of Tatarstan, Kazan
dvdt.slt@gmail.com, anis_59@mail.ru

The article describes the principles of creating terminology in computer science and information technology, provides a classification of principles and rules and prospects for the creation of new terms based on these rules.

Keywords: Tatar terminology, informatics, infocommunication technologies.

СОЗДАНИЕ И РАЗВИТИЕ ТЕРМИНОТВОРЧЕСТВА ПО ИНФОРМАТИКЕ И ИНФОКОММУНИКАЦИОННЫМ ТЕХНОЛОГИЯМ НА ТАТАРСКОМ ЯЗЫКЕ

Д. Ш. Сулейманов, А. Ф. Галимянов
Институт «Прикладная семиотика» АН РТ, Казань
dvdt.slt@gmail.com, anis_59@mail.ru

В статье описываются принципы создания терминологии по информатике и информационным технологиям, приводится классификация принципов и правил и перспективы создания новых терминов, опирающихся на указанные правила.

Ключевые слова: татарская терминология, информатика, инфокоммуникационные технологии.

Как было сказано в наших прежних работах ([7], [8]), принципы образования понятий, терминов можно разделить на два типа: принципы отвечающие на вопрос – "Каким образом порождаются татарские понятия и термины?" и принципы второго типа, отвечающие на вопрос – "Какими должны быть татарские понятия и термины?".

1.1. Первый тип принципов порождения терминов и понятий в татарском языке

1.1.1. Поиск готовых понятий и терминов в самом языке, в том числе в диалектах. Например, санак (посох пастуха для "фиксирования зазубринами" расчетов с хозяевами) – компьютер (рус.), тәрәз (диалектн.) – window (англ.) – окно (рус.) (для компьютерных систем). Данный принцип имеет большой потенциал для порождения новых понятий. Тому же принципу отдают предпочтение и другие ученые из тюркских стран. Например, казахский ученый Байтырсунов А.В. считает, что для терминов предпочтительно перед другими языками брать казахские слова с вполне соответствующим данному понятию значением. По мнению ученого, общеупотребительные международные термины должны приниматься, но с изменениями, соответствующими природе казахского языка. При наличии казахских слов, могущих заменить их, должны помещаться оба, чтобы право выбора предоставить обществу. Все неказахские слова, не согласующиеся с природой казахского языка, точно должны подвергаться изменениям соответственно казахскому говору. Тот же принцип вполне приемлем и применяется в терминотворчестве на татарском языке.

1.1.2. Порождение новых понятий, используя правила татарского языка.

1. <Корень>+<аффиксы>, т. е., порождение нового понятия путем добавленя к корню аффиксов. Например, бегунок – шудырма: шудыр [глагол в пов. накл., 1 л., ед.ч.] + ма [аффикс субстантивизации], меню – сайлак: сайла [повел. накл., 2 л.] + у [афф. им. действ.] + к [словообр. афф. (инструмент)].

Татарский язык в силу агглютинативности обладает богатыми возможностями словообразования и словоизменения с помощью аффиксов. И с учетом того, что некоторые аффиксы являются одновременно и словообразующими и словоизменяющими, а некторые из них слова одной части речи превращают в другие (имя сущ. в глагол, глагол в имя сущ., имя сущ. в прилаг., глагол в наречие и т. д.), то на примере татарского языка можно видеть, каким мощным морфологическим потенциалом обладают тюркские языки для порождения новых понятий. Сегодня наиболее активны аффиксы -чы/-че, -лык/-лек, -лы/-ле, -ла/-лә и еще несколько таких аффиксов. В то же время, аффиксы –мә/-ма (ярма: яр+ма – сечка), -[ә]к/-[а]к (тарак: тара+к – расческа), –чә/-ча (кулча: кул+ча – кольцо, аркача:

арка+ча – рукавицы), -чек/-чык (уенчык: уен+чык – игрушка), образующие из существительного глагол, практически уже вышли из активного использованя и изучаются только в плане диахронии. Их можно было бы назвать “спящим” потенциалом.

2. <Корень>+<Корень>, т. е., порождение из двух корневых морфем понятия с одним значением. Например, видеопамять – видеохәтер: видео [корень] + хәтер [корень], полуслово – ярымсүз: ярым [корень] + сүз [корень].

3. Определение одного значения с помощью словосочетания. Например, архитектура ЭВМ – ЭХМ архитектурасы, ворона – кара карга.

4. Порождение нового значения с помощью парных слов. К первому слову через дефис добавляется второе вспомогательное слово. В итоге с использованием второго слова образуется обощенная форма первого слова. Например, посуда – савыт-саба, дети – бала-чага (бала – ребенок).

Естественно, потенциал татарского языка по образованию новых понятий исходя из соих внутренних возможностей не ограничивается представленными вариантами. Использование парных корневых слов, синонимия, использование слоосочетаний и т. д. – это мощный потенциал для определения новых поятий, который еще ждет своего серьезного изучения.

1.1.3. Заимствование терминов из тюркских языков. Сегодня немало тех, кто среди тюркских народов на своем языке пишет научные труды, использует свой язык в компьютерных технологиях, порождает свои термины и понятия в соответствии с требованием времени. Здесь наиболее активны турки, татары, казахи, азебайджанцы, якуты, киргизы, чуваши. В эпоху глобализации актуальность интеграции родственных языков возрастает. Межъязыковая связь обусловлена моноистоком языков, входящих в одну языковую семью. Сходство и родство тюркских языков исследовались и исследуются различными учеными (Баскаков Н., Богородицкий В., Дмитриев Н., Щербак А., Томанов М., Курышжанов Г., Сагиндыкулы Б., Сабыр М. и др). Однако на сегодняшний день взаимосвязь и взаимопроникновение тюркских языков, обмен научными знаниями и, соответственно, обмен терминами слабо развит. Такого же мнения придерживается и казахский ученый Ф.Р. Зейналов: «Несмотря на общность словарного состава языков тюркской семьи и ряд других объединяющих их особенностей и черт, восприятие издающейся на том или ином тюркском языке литературы, особенно научной, затруднено для представителей других тюркоязычных народов, что

в определенной мере осложняет научные контакты между представителями одной профессии и не способствует успешному развитию отдельных отраслей науки, на что указывается в специальной лингвистической литературе» [3].

В вопросе терминологии тюркских языков существует мнение о наличии устоявшейся терминологической системы в отдельных языках, а также и о ее отсутствии. Например, если ученый К.М. Мусаев считает, что в тюркских языках еще не вполне устоялись терминологические сферы [4] то, по мнению Х.Ф. Исхаковой: «Современный этап развития тюркских языков характеризуется высокой степенью разработанности терминологии, связанной прежде всего с гуманитарными науками, и только по ним на родном языке печатаются в настоящее время научные труды и ведется преподавание в высшей школе» [4].

В качестве аргумента К.М. Мусаев приводит следующее: терминологическая система в значительной степени определяется существованием национально-русского билингвизма и русского одноязычия в ряде сфер. Кроме того, развитие терминологии в разных тюркских языках наряду с общими моментами отличается спецификой в каждой из республик, поскольку у литературных языков и соответственно у национальных терминологических систем – различные исторические судьбы [6].

Поддерживая мысль об изолированном развитии терминологии тюркских языков друг от друга, имеющемся различии в исторических судьбах, особенностях процесса формирования национальной культуры, количественном составе носителей, этническом окружении народа, исследователь Л.Х. Махиева утверждает, что «несмотря на общность исконного словарного состава языков тюркской системы, в современных тюркских языках наблюдается большое расхождение в терминотворчестве. Здесь сказывается специфика каждого литературного языка» [5].

Из всего сказанного следует, что родство тюркских языков под влиянием экстралингвистических факторов утрачивается и ослабевает на современном этапе. Безусловно, в результате этого процесса в языке наглядно разделяется лексика того или иного языка, она становится менее понятной носителям разных тюркских стран. Появляются новые законы в языке, вытесняя старые посредством другого доминирующего языка. Например, ослабевание закона сингармонизма в татарском языке и т. д. Однако типология языка и мотивация словотворчества у тюркских языков на сегодняшний день в до-

статочной мере близки. В них еще наблюдается метаязык. Об этом говорил и ученый Н.З. Гаджиев: «Основные лексико-семантические группы слов – термины родства, названия диких и домашний животных, названия предметов (имена существительные), названия качеств (имена прилагательные), названия действий (глаголы) и др. – составляют общетюркский словарный фонд, представленный во всех ареальных группах тюркских языков. В отличие от языков другой типологической структуры, тюркские корни, да и производные основы, фонетически относительно мало трансформировались.

Тем не менее, можно сказать, что практически отсутствует обмен терминами между тюркскими народами, заимствование из других тюркских языков понятий или терминов, не сохранившихся на своем языке. Особенно показательно то, что мы сегодня находимся еще только в начале процесса, который ведет к постепенному отдалению тюркских языков друг от друга, но не продвигаемся дальше разговоров, резолюций конференций, каждый тюркский народ продолжает порождать свои “особые” термины и понятия. Это особенно наглядно видно на примере инфокоммуникационных технологий.

На наш взляд, сегодня так же актуален вопрос создания единого терминологического фонда, где могут применяться общие принципы терминообразования. Для этих целей и для дальнейшей интенсивной интеграции необходимы целенаправленные исследовательские работы по тюркским языкам, сравнительный анализ каждого языка с другими тюркскими языками. Таким образом, мы можем усилить терминотворчество каждого тюркского языка. Однако это не должно привращаться в массовое копирование. Терминотворчество каждого языка должно развиваться непосредственно по модели единого терминологического фонда тюркских языков с сохранением национальных особенностей языков.

1.1.4. Заимствование понятий из других нетюркских языков. Это один из самых активных принципов, работающих сегодня. Например, утюг (рус.) – үтүк, file (англ.) – файл (рус.) – файл, computer (англ.) – компьютер (рус.) – компьютер. Заимствование терминов из другого языка без перевода – это достаточно распространенный и естественный процесс. Во многих случаях это оптимальный путь, оправдывающий себя при преподавании предметов на татарском языке. Скажем, можно признать, что перевод таких понятий как интеграл, функция, дифференциал, анализ, синтез на татарский язык, или обозначение их каким–либо другим татарским понятием является, как минимум, бесполезным, а то и во вред пониманию.

Однако так же бесспорно, что эти слова максимально должны быть ассимилированы, т. е. должны быть подчинены правилам фонетики татарского языка.

1.1.5. Замствование понятий из других языков путем их перевода.

1. Прямой перевод. Перевод корневых слов: mouse – мышь – тычкан, window – окно – тәрәз; перевод словосочетаний: булево значение – буль кыйммәте, архитектура электронных вычислительных машин – электрон хисаплау машиналарының архитектурасы.

2. Перевод по смыслу. Порождение нового понятия на основе смысла. Например, если понятие mouse – мышь – тычкан было дано устройству из-за длинного шнура – “хвоста”, то сейчас уже широко распростараняется вариант мыши без хвоста и это устройство по форме можно было бы назвать “бака – жаба”, хәтта “кабартма – пончик”. Функциональный смысл этого устройства в управлении стрелкой на экране монтора. Исходя из этого его можно было бы назвать также “укйөрткеч” (водитель стрелки), я “укидарә” (управляющий стрелкой).

Пять принципов, приведенных выше, это практически основные принципы, которые известны давно и используются многими поколениями для образования понятий и терминов на татарском языке.

К этим известным принципам можно добавить еще три новых принципа. Ниже раскроем эти принципы, которые практически еще не изучены или недостаточно отражены в татарской лексикологии.

1) Принцип “Формального гнезда” (образование новых слов из каркаса татарского слова, в котором пропускаются все гласные буквы и слово-схема заполняется другими гласными буквами).

2) Принцип “возвращенных слов”, принцип “восстановления слова” (возвращение слова, которое этимологичсеки является тюркским, сохранилось в других языках, или же не применяется в языке и обзначно как архаизм).

3) Принцип “Блендинг” – синтетический принцип (образование новой метафоры из нескольких метафор с обозначением ее новым именем). Этот принцип активно прменяется для языков индоевропейской группы.

1.1.6. Принцип “формального гнезда”. Структура татарских корневых слов представляет собой матрицу из согласных букв, заполненных гласными буквами. Покажем несколько примеров. Как известно, в татарском языке 9 гласных букв (а-ә, о-ө, у-ү, ы-е, и). Соответсвенно, заполняя схему “т-з”, можно образовать сле-

дующие слова: таз-тәз-тоз–төз-туз-түз-тыз-тез-тиз. Здесь 7 слов из 9-и представлены в татарском словаре: таз (тазик), тоз (соль), төз (стройный), туз (береста), түз (терпи), тез (строй), тиз (быстро). В схеме “к-н” из 9 возможностей в современном татарском языке используется 6 слов: кан (кровь), кон (кон), көн (день), кун (взлетев садись), күн (кожа), кын (ножны). Хотя слово кон (при игре в карты) и не ассимилировано на татарский язык, оно включено в словари, соответственно, признается как татарское. В то же время по схеме “ж-л” образуется всего одно слово, это слово җил (ветер), и нет ни одного татарского литературного слова, которое можно породить по по схеме “б-н”. Такие схемы мы называем “формальными гнездами”, которые практически являются готовыми структурами для порождения новых слов.

Какой вывод можно сделать из приведенных примеров?

Во-первых, по разным схемам образуется разное количество татарских слов. Соответственно, интересно было бы провести исследование, почему такой разброс при использовании формальных схем, почему сегодня одни заполненные формы используются в языке, а другие нет?

Во-вторых, это явление «подталкивает» к интересным исследованиям: не сохранились ли эти “спящие” слова в диалектах? Также и поиск в других тюркских языках, определение их значения может привести к интересным результатам.

Даже эти несколько схем, которые приведены выше, показывают, что принцип “формальных гнезд” открывает новый потенциал для порождения новых понятий и терминов в татарском языке. Как видно из примеров, слова, образованные из таких схем путем добавления гласных, хотя и не используются в современном татарском языке, однако и по написанию, и по произношению могут быть признаны татарскими, то есть не противоречат закономерностям языка. Соотвественно, в последующем, с возрастанием потребности в новых понятиях и терминах в технологиях и различных науках, можно будет обратиться к данному принципу “автоматного” порождения новых слов. Очевидно, технология такого словообразования достаточно простая. По образцу слов, имеющихся в толковом и орфографическом словарях, воспринимая их как “формальные гнезда”, путем замены гласных букв в структуре слова на согласные, создается база всевозможных новых слов. Далее из этой базы исключаются слова, уже имеющиеся в толковом и орфографическом словаре и те слова, которые являются словоформами, получаемыми из словар-

ных слов путем присоединения аффиксальных морфем, чтобы исключить появления омонимов. Исключаются также слова, которые не соответсвуют определенным закономерностям татарского языка, например, слова, в которых последняя “клетка гнезда” заполняется буквой “ө” (есть слово төтен (дым), но нет слова “төтөн”).

1.1.7. Блендинг (Blending): гармоничное слияние метафор в одном понятии (термине). Принцип создания слова, означающего метафору, образованной путем слияния нескольких метафор. Например, слиянием слов пүчтәк (рус.: пустяк, пустячок) и күчтәнәч (презент) можно образовать слово пүчтәнәч. В итоге данное новое слово передает новое значение: маленький, недорогой подарок, образованное из значений двух метафор.

1.1.8. Возвращение, восстановление. Известно, что татарский язык, как один из тюркских языков, имеет многовековые корни, и это является признанным фактом, обоснованным многочисленными публикациями. Тюркско-татарские слова можно обнаружить не только в русских словарях, что вполне естественно для соседствующих, взаимопроникающих языков, но также и в лексике таких народов, как греки, арабы, англичане, немцы, китайцы. Как это можно объяснить? Перешли ли эти слова, понятия, термины с татарского языка в тот период, когда, как известно из истории, тюркско-татарский язык был глобальным языком – имел сильные позиции в развитии цивилизаций, активно участвовал в описании мировых явлений и процессов, или неоторые из этих языков сами являются продолжением древнетюркского языка, изменившимся до неузнаваемости? На этот счет нет единого мнения даже среди маститых ученых. Для этого нужны серьезные исследования с привлечением всего научно-инструментального потенциала лингвоархеологии, этимологии и других смежных наук. В качестве примера покажем ряд таких слов, которые претендуют на роль возвращенных и восстановленных.

В электронной почте для обозначения адреса используется символ ‘@’. Американцы первыми ввели этот символ в оборот в качестве коммерческого символа. Этот символ произносится как “эт”. Очевидно, это имеет свое объяснение. Этот глиф (знак), заимствованный у народа майа, сильно напоминает голову собаки с перекинутым через голову хвостом, и вполне можно предположить, что символ ‘@’ обозначал собаку, что по-татарски звучит как эт. Видимо, не случайно, на русском языке этот символ так и произносится: “собачка”. Таким образом, для возвращения названия данного симола в татарский язык необходимо вместо слова “собачка” вернуть

начальное, тюркское – “эт”. К восстановленным словам и понятиям мы относим те слова, которые когда-то были активны в языке, а в настоящее время не употребляются или стали архаизмами. В качестве примера можно привести следующие слова: хисап (счет), әсбап (предмет), мөгаллим (учитель).

2.2. Вторая группа принципов образования понятий и терминов в татарском языке

Вторая группа принципов определяет то, какими должны быть новые термины и понятия.

Первый принцип: чем короче слово, определяющее понятие или термин – тем лучше. Наиболее пердпочтительной формой слова является корневая.

Второй принцип: наиболее выигрышным является обозначение понятия или термина одним словом (особенно, с точки зрения технологий).

Третий принцип: “избегание” омонимии. Понятия, термины должны иметь только одно значение. Например, слово печать – бастыру предпочтительнее слова язу, потому что писать (язу) можно и на экране, однако печатать (бастыру) можно только на принтере.

Четвертый принцип: “понятность”, “ясность” понятий, их распространенность, привычность; активное и широкое использование терминов, иногда даже в ущерб некоторым принципам, например, фонетическому – компьютер, дифференциал, функция).

Пятый принцип: благозвучие. Фонетичекая ассимиляция с татарским “произношением” (cash: руское произношение – кэш, татарское: кәш, stack – стек – стәк, tag – тег – тәг).

Шестой принцип: использование редких вариантов синонимов (как правило, диалектные варианты) (окно – тәрәз/тәрәзә: тәрәз; активный рабочий стол – актив өстлек/актив өслек: актив өстлек).

Седьмой принцип: прямой перевод иностранных понятий и терминов на татарский язык (не через третий язык), исправление некорректных переводов (кальки), полученных через русский язык (арифметическое действие – арифметик гамәл – арифметика гамәле).

Восьмой принцип: уход от неологизмов. Не порождать новых слов, которые создают трудности в понимании и использовании термина. Любые заимствованные слова должны быть удобны для использования в татароязычной среде, не должны нарушать закономерностей, правил языка, а наоборот, должны им подчиняться.

Девятый принцип: иностранные слова должны заимствоваться строго в качестве корневых слов. Любые грамматические, просодические и другие изменения и проявления данного слова в языке должны подчиняться правилам татарского языка.

ЛИТЕРАТУРА

1. Байтурсынов А. Қазақ тіл білімінің мәселерлері. – Алматы: Абзал-ай, 2013. – 640 стр.

2. Гаджиев Н.З. Языки мира. Тюркские языки. – М., 1997. – С. 17–34.

3. Зейналов Ф.Р. О необходимости создания сравнительного словаря лингвистических терминов тюркских языков // Советская тюркология. – 1973. – № 4. – С. 68–71.

4. Исхакова Х.Ф. Структуры терминологических систем. Тюркские языки. М.: Наука, 1987. – 125 с.

5. Махиева Л.Х. Формирование и развитие лингвистической терминологии карачаево-балкарского языка. дис. канд. филол. наук.- Нальчик. – 2003 166 с.

6. Мусаев К.М. Современные проблемы терминологии на тюркских языках СССР // Советская тюркология. 1989. – № 4. – С. 18–30.

7. Сулейманов Д.Ш., Галимянов А.Ф. Система татарских терминов в компьютерных технологиях и информатике // В сб. Трудов Казанской школы по компьютерной и когнитивной лингвистике ТЕЛ-2012. – Казань: Изд-во «Фэн» Академии наук РТ, 2012. – С. 61–69.

8. Сулейманов Д.Ш., Галимянов А.Ф., Валиев М.Х., Желтов П.В., Желтов М.П., Желтов В.П. Англо-русско-татарско-чувашский словарь терминов по информатике и информационным технологиям (с толкованиями на татарском языке). Приложение к Материалам Третьей Международной конференции по компьютерной обработке тюркских языков (TurkILang 2015, Казань, 17–19 сентября 2015 г.) // Казань, изд-во АН РТ, 2015. – 400 с.

УДК 81.33; 811.512.145; 81.367.5

REALIZATION FEATURES OF SEMANTIC-SYNTACTIC ANALYZER OF TATAR SENTENCE

D. Sh. Suleymanov, A. R. Gatiatullin, M. M. Ayupov, A. M. Bashirov, R.R. Gataullin
Institute of Applied Semiotics of the Academy of Sciences of the Republic of Tatarstan, Kazan
dvdt.slt@gmail.com, agat1972@mail.ru, madehur@mail.ru, a.basheerov@gmail.com, ramil.gata@gmail.com

The article presents the features of a semantic-syntactic analyzer developed for the Tatar simple sentence. The analyzer is a computer program which inputs sentences in the Tatar language, and outputs their syntactic structures in the form of immediate constituent trees. The nodes of such trees represent verb and noun groups of the input sentences. The peculiarity of these syntactic trees is that their nodes also contain semantic information about the sentence fragments mapped to those nodes, together with their semantic and syntactic roles in the sentence. The technical implementation of the semantic-syntactic analyzer software is carried out using the NLTK (Natural Language ToolKit) open libraries. The results and resources obtained while creating the project should contribute to solving a wide range of tasks in computer processing of the Tatar language.

Keywords: semantic-syntactic analyzer, Tatar language, trees of direct components.

ОСОБЕННОСТИ РЕАЛИЗАЦИИ СЕМАНТИКО-СИНТАКСИЧЕСКОГО АНАЛИЗАТОРА ТАТАРСКОГО ПРЕДЛОЖЕНИЯ

Д. Ш. Сулейманов, А. Р. Гатиатуллин, М. М. Аюпов, А. М. Баширов, Р. Р. Гатауллин
Институт прикладной семиотики Академии наук Республики Татарстан, Казань
dvdt.slt@gmail.com, agat1972@mail.ru, madehur@mail.ru, a.basheerov@gmail.com, ramil.gata@gmail.com

В данной статье представлены особенности реализации семантико-синтаксического анализатора татарского простого предложения. Данный анализатор представляет собой компьютерную программу, на вход которой поступает предложение на татарском языке, а на выходе получаются синтаксические структуры в виде деревьев непосредственных состав-

ляющих. Особенности этих синтаксических деревьев в том, что узлы этого дерева, представляющие собой именные группы, содержат информацию о семантических ролях, которые выполняют фрагменты текста, представленные в этих узлах предложения. Техническая реализация программного обеспечения семантико-синтаксического анализатора осуществляется с помощью открытых библиотек NLTK (Natural Language ToolKit). Результаты и ресурсы, полученные в ходе реализации данного проекта, должны способствовать решению целого круга задач по компьютерной обработке татарского языка.

Ключевые слова: семантико-синтаксический анализатор, татарский язык, деревья непосредственных составляющих.

В настоящее время существует несколько подходов по разработке синтаксических анализаторов для анализа предложений естественного языка: классический (на основе правил), статистический и нейросетевой. Последние два требуют наличия синтаксически размеченных корпусов большого объема. Для русского языка примером такого размеченного корпуса является СинТагРус [Богуславский, 2008], который использован для обучения многих самообучающихся программных разработок. Однако, для татарского языка такого размеченного корпуса в настоящее время не существует, а его создание требует достаточно больших временных затрат. Создание компьютерных моделей, баз данных и программного обеспечения, основанных на правилах, позволяет создавать лингвистические ресурсы с описанием языка, которые могут быть использованы также и в других проектах. Например, в информационно-справочных и обучающих системах. Эти факторы способствовали принятию решения о создании семантико-синтаксического анализатора, основанного на правилах.

Результатом работы этого семантико-синтаксического анализатора должны стать деревья непосредственных составляющих (НС-деревья), получаемых из предложений татарского языка. Выбор деревьев непосредственных составляющих обусловлен тем, что по синтаксической классификации тюркские языки (в том числе и татарский язык) являются языками проективного типа. В узлах деревьев непосредственных составляющих можно представлять не только отдельные слова, но и многословные выражения, между элементами которого невозможно поставить отношения зависимости.

Одним из недостатков НС деревьев является, то что НС-деревья не определяют никаких отношений среди составляющих одного уровня. Этот недостаток отсутствует в ориентированных деревьях

непосредственных составляющих (ОНС-деревья). Однако, ОНС-деревья, в свою очередь, наследуют недостаток деревьев зависимостей, а именно, неспособность адекватно описывать неподчинительные связи. Для таких случаев предусмотрен вариант частично-ориентированных деревьев непосредственных составляющих (ЧОНС-деревья). В ЧОНС-деревьях главные непосредственные составляющие выделяются не для всех элементов, а только для некоторого подмножества.

Для реализации семантико-синтаксического анализатора простого выделения главного элемента среди непосредственных составляющих одного уровня недостаточно, для этого необходимо присутствие семантической информации. Такой семантической информацией в нашем проекте являются ролевые структуры предикатов. В работе [Харламов, 2013] указано, что "Использование словаря валентности глаголов позволяет уменьшить количество вариантов разбора, поскольку по нему выбирают предпочтительные связи между глаголом и его актантами, запоминают семантический класс глагола и типы предикатной связи присутствующих в предложении актантов". В своем проекте мы решили проверить эту гипотезу и создать программу, которая строит комбинированные НС-деревья. Комбинированность деревьев проявляется в том, что для предикатов определяются их семантические актанты и информация об этих ролях прописывается в узлах НС-деревьев. На остальных уровнях деревьев непосредственных составляющих предполагается выделение главных НС-узлов этого уровня.

Что касается технической реализации программного обеспечения, то здесь работа велась с учетом того, что в настоящее время уже существует множество разработок для синтаксического анализа предложений на других языках, поэтому нужно попытаться максимально использовать готовые разработки. Следует отметить, что локализация существующих программных продуктов предполагает большой объем экспериментальной работы, а значит для реализации нужно выбрать такое программное обеспечение, которое позволяет проводить эти эксперименты. В качестве такого программного продукта нами выбрана библиотека NLTK (Natural Language Toolkit) [Bird, 2009]. Где NLTK (www.nltk.org) – это пакет библиотек и программ для символьной и статистической обработки естественного языка, написанных на языке программирования Python. Синтаксические правила для работы анализатора представляются в виде правил контекстно-свободных грамматик.

Алгоритм работы семантико-синтаксического анализа включает следующие основные этапы:

1. Морфологический анализ.
2. Определение многословных выражений (аналитических форм).
3. Определение валентностей глаголов (из словаря валентностей).
4. Построение структуры НС-деревьев с семантическими ролями.

Для проведения экспериментов с правилами и отработки технологий семантико-синтаксического анализа создан сайт семантико-синтаксического анализа татарского предложения. Информация о семантических валентностях глаголов закладываются непосредственно в сами правила контекстно-свободных грамматик. Эта информация берется из словаря валентностей татарских глаголов.

Для представления результатов семантико-синтаксического анализа также необходимо разработать систему семантико-синтаксических тегов. Эти теги должны содержать информацию, как о синтаксическом элементе, так и о семантической роли.

Данная статья выполнена при поддержке Российского фонда фундаментальных исследований РФФИ 18-47-160014 «Разработка интегральной компьютерной модели и программного инструментария для семантико-синтаксического анализа татарских текстов».

СПИСОК ЛИТЕРАТУРЫ

1. Bird, Steven. Natural Language Processing with Python. – O'Reilly Media Inc, 2009.

2. Богуславский И.М., Валеев Д.Р., Иомдин Л.Л., Сизов В.Г., Синтаксический анализатор системы ЭТАП и его оценка с помощью глубоко размеченного корпуса русских текстов // Труды Международной конференции «Корпусная лингвистика – 2008». СПб.: Санкт-Петербургский государственный университет, 2008. С. 56–74. ISBN 978-5-288-04769-5.

3. Харламов А.А., Ермоленко Т.В., Дорохина Г.В. Сравнительный анализ организации систем синтаксических парсеров // Инженерный вестник Дона, № 4, 2013, http://www.ivdon.ru/ru/magazine/archive/n4y2013/2015.

УДК 007.51:372.851

PROBLEMS AND SOLUTIONS IN THE DESIGN OF ONTOLOGY SCHOOL COURSE OF PLANIMETRY

L. R. Shakirova, M. V. Falileeva, A. V. Kirillovich, E. K. Lipachev, Sh. M. Khaidarov
Kazan Federal University, Kazan
liliana008@mail.ru, mmwwff@yandex.ru, alik.kirillovich@gmail.com, elipachev@gmail.com, 15jkeee@gmail.com

Digitalization of education requires the creation of high-quality subject databases, on the basis of which it is possible to create training intelligent systems. The task of our research is to design educational mathematical ontology. The section “Planimetry” of school mathematics was chosen as a pilot project. To translate the school system of knowledge into a formal language, a system analysis was required, which showed that there are gaps indicating incompleteness, lack of system and non-timeliness of presenting information on planimetry. Not only the problems of translating planimetry into a formal language, but also the imperfection of the methodological system of teaching this section of mathematics are revealed. Among the students of the Institute of Mathematics and Mechanics. N.I. Lobachevsky Kazan Federal University using a test built on a fragment of ontology, conducted a study of the quality of understanding of generic concepts, graphic representations of geometric figures, the results of which showed the interrelation of problems in the presentation of school planimetry with the quality of students’ knowledge.

Keywords: conceptualization of mathematical knowledge, planimetry, ontology, OntoMathEdu.

ПРОЕКТИРОВАНИЕ ОБРАЗОВАТЕЛЬНОЙ МАТЕМАТИЧЕСКОЙ ОНТОЛОГИИ: ПРОБЛЕМЫ И МЕТОДЫ РЕШЕНИЯ НА ПРИМЕРЕ КУРСА ПЛАНИМЕТРИИ

Л. Р. Шакирова, М. В. Фалилеева, А. В. Кириллович, Е. К. Липачев, Ш. М. Хайдаров
Казанский федеральный университет, Казань
liliana008@mail.ru, mmwwff@yandex.ru, alik.kirillovich@gmail.com, elipachev@gmail.com, 15jkeee@gmail.com

Цифровизация образования требует создания качественных предметных баз данных, на основе которых можно создавать обучающие интеллектуальные системы. Задача проводимого нами исследования состоит в

проектировании образовательной математической онтологии. В качестве пилотного проекта выбран раздел «Планиметрия» школьной математики. Для перевода школьной системы знаний на формальный язык потребовался системный анализ, который показал, что существуют пробелы, указывающие на неполноту, бессистемность и несовременность подачи информации по планиметрии. Выявлены не только проблемы перевода планиметрии на формальный язык, но и несовершенство методической системы обучения этому разделу математики. Среди студентов Института математики и механики им. Н.И. Лобачевского Казанского федерального университета с помощью теста, построенного по фрагменту онтологии, проведено исследование качества понимания родовидовых понятий, графических представлений геометрических фигур, результаты которого показали взаимосвязь проблем в подаче содержания школьного курса планиметрии с качеством знаний студентов.

Ключевые слова: концептуализация математического знания, планиметрия, онтология, OntoMathEdu.

Введение

Данная работа посвящена разработке OntoMathEdu – образовательной математической онтологии.

Онтология планируется как центральный компонент цифровой образовательной платформы.

Проектируемая онтология позволит, в частности, решить следующие задачи: (1) семантическая разметка математических учебников; (2) автоматическая рекомендация учебных материалов в соответствии с индивидуальным профилем обучающегося; (3) автоматическая генерация тестовых заданий для проверки знаний. Кроме того, данная онтология должна выступать основой для справочной базы знаний, ориентированной непосредственно на конечных пользователей.

В настоящее время существует ряд математических онтологий, одной из которых является онтология профессиональной математики OntoMathPro [1, 2]. Данная онтология (см. рис. 1) лежит в основе платформы семантической публикации [3], предназначенной для извлечения структурированных данных из текстовой коллекции математических публикаций и интеграции их в облако Открытых связанных данных (LOD). Платформа семантической публикации, в свою очередь, является центральным компонентом экосистемы OntoMath [4], объединяющей множество сервисов по управлению математическим знанием, такие как семантический поиск по мате-

матическим формулам [5] и рекомендательная система математических публикаций [6].

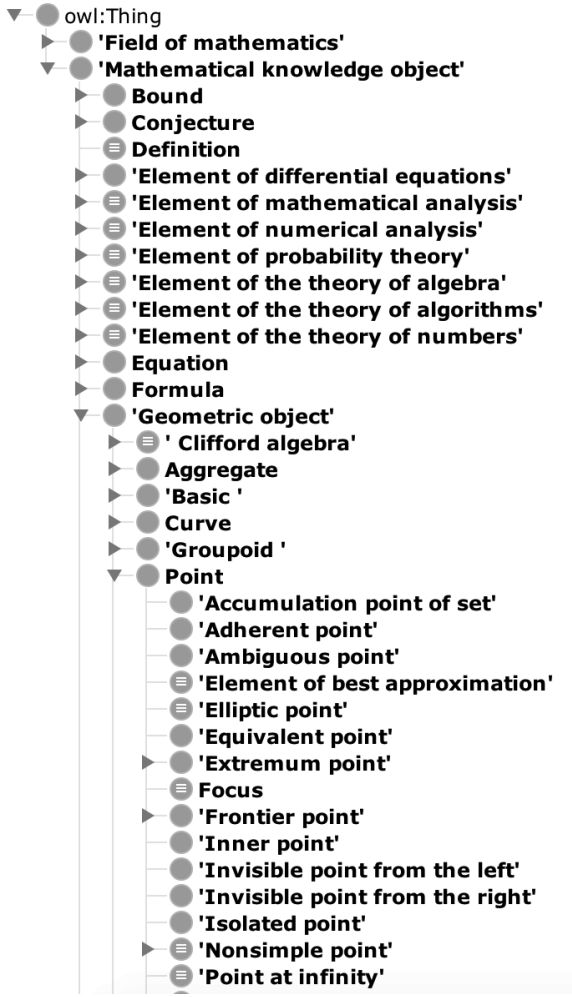


Рис. 1. Фрагмент иерархии онтологии OntoMathPro

Однако онтология OntoMathPro ориентирована на профессиональную математику и не отражает образовательный аспект, что создает проблемы при ее использовании в качестве основы для циф-

ровой образовательной платформы OntoMathEdu. Их можно сгруппировать следующим образом.

• *Терминология*. В онтологии OntoMathPro для обозначения концептов используется профессиональная математическая терминология, в то время как в онтологии цифровой образовательной среды для обозначения концептов должна использоваться терминологии школьной математики (например, вместо термина «коммутативность» нужно использовать термин «переместительный закон»).

• *Отбор понятий*. Онтология цифровой образовательной среды должна содержать понятия из разделов школьной математики.

• *Дидактические связи между понятиями*. Онтология цифровой образовательной среды должна содержать отношения, отражающие дидактическую зависимость между понятиями (в частности, эти отношения отражают последовательность и преемственность изучения понятий).

• *Концептуализация*. Онтология OntoMathPro отражает концептуализацию профессиональной математики, в то время как онтология цифровой образовательной среды должна отражать концептуализацию школьной математики (концептуализации школьной и профессиональной математики имеют существенные отличия, например, в школьной математике понятие «число» является неопределяемым, базовым, а в профессиональной математике оно определено как подвид множества).

• *Точки зрения*. Образовательная онтология должна представлять не только математику как таковую, но и математику с некоторых точек зрения (например, с точек зрения различных уровней подготовки или разных учебников).

В связи с вышеизложенным возникает необходимость разработки новой образовательной онтологии OntoMathEdu. В данной статье опишем опыт создания фрагмента этой онтологии на примере курса планиметрии.

Состав онтологии

Онтология OntoMathEdu состоит из следующих модулей: 1) иерархия типов; 2) иерархия материализованных отношений; 3) иерархия ролей и 4) сеть точек зрения.

Текущая версия онтологии содержит 585 концептов, относящихся к курсу планиметрии 7–9 классов школы (рис. 2).

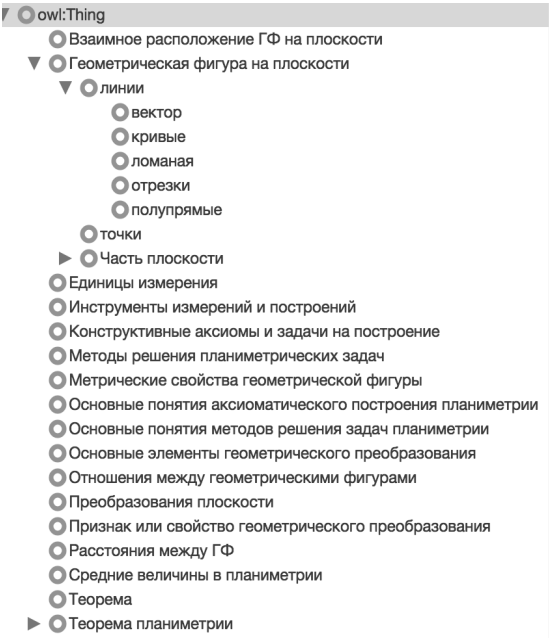


Рис. 2. Фрагмент иерархии онтологии OntoMathEdu

Иерархия типов

Базовой иерархией онтологии является иерархия типов. Тип – это концепт, который является семантически жестким и онтологически независимым [7]. Так, например, концепт *треугольник* является типом, т. к. любой треугольник всегда является треугольником вне зависимости от его отношения с другими фигурами.

Верхний уровень иерархии типов состоит из следующих концептов:

1. *Геометрическая фигура на плоскости*. Главный класс, включающий в себя все геометрические фигуры планиметрии. Особенности проектирования этого класса будут рассмотрены подробнее далее.

2. *Аксиома планиметрии*. В данный класс входят подклассы аксиом аксиоматик Гильберта, Вейля и аксиомы теории геометрических построений. Например, одним из концептов этого класса является аксиома «принадлежности» о единственности прямой, проходящей через две точки.

3. *Теорема планиметрии*. В данный класс входят подклассы «Свойства геометрических фигур», «Признаки геометрических фигур», «Теоремы геометрических преобразований» и др. Например, теорема Пифагора входит в подкласс «Свойства прямоугольного треугольника», который является подклассом «Свойства треугольника», который в свою очередь является подклассом класса «Свойства геометрической фигуры».

4. *Планиметрическая задача.* Класс включает в себя класс «Задачи на построение» (задачи о построении биссектрисы угла циркулем и линейкой, о спрямлении окружности и др.), известные планиметрические задачи (задача Герона о нахождении наименьшей суммы расстояний от точек до прямой и др.).

5. *Единица измерения*. В класс входят подклассы «Единица измерения площади», «Единица измерения длины» и «Единица измерения угла», в которые соответственно, например, включены квадратный сантиметр, сантиметр, градус и др.

6. *Инструмент построения и измерения*. Приведем примеры подклассов – «Инструмент измерения углов» (транспортир, астролябия), «Инструмент построения параллельных прямых» (рейсшина, малка, рейсмус) и т. д.

Иерархия материализованных отношений

Отношения между понятиями представлены в онтологии в материализованном виде, т. е. не в виде онтологических концептов, а в виде онтологических отношений. Благодаря такого рода представлению, отношения между концептами являются сущностями первого порядка, и выступают в качестве объекта утверждений.

Верхний уровень иерархии материализованных отношений состоит из следующих концептов:

1. *Преобразование плоскости*. Приведем примеры концептов: параллельный перенос, гомотетия, инверсия, симметрия и др.

2. *Метрическое свойство геометрической фигуры*. В классе выделены подклассы: «длина линии», «мера угла», «площадь ограниченной части плоскости», «тригонометрические функции острого угла прямоугольного треугольника», «степень точки относительно окружности», «эксцентриситет эллипса», «эксцентриситет параболы», «длина вектора». Общим выражением каждого геометрического понятия является число.

3. *Отношение сравнения геометрических фигур*. Класс включает подклассы: «сравнение геометрических фигур», «равные гео-

метрические фигуры», «равносоставленные геометрические фигуры», «равновеликие геометрические фигуры», «подобные фигуры», «пропорциональные отрезки».

4. *Взаимное расположение геометрических фигур на плоскости*. В этот класс, например, входят концепты: описанный многоугольник, треугольник с вершинами в точках Эйлера, вписанный угол и др.

Иерархия ролей

Роль – это концепт, который является семантически нежестким и онтологически зависимым [7]. Объект, относящийся к ролевому концепту, относится к этому концепту только в силу его отношения с другим объектом. Так, например, концепт *вершина треугольника*, является ролью, т.к. точка является вершиной треугольника не сама по себе, а только в отношении к некоторому треугольнику.

Одним из наиболее важных ролевых концептов является концепт *основной элемент геометрического преобразования*. Этот класс отличается разнородностью концептов, включенных в него. Например, в подкласс неподвижные элементы плоскости при преобразовании входят и «ось симметрии», и «неподвижная окружность при инверсии», так же в этот класс входят и коэффициент гомотетии.

Точки зрения

Помимо универсальных утверждений о математических понятиях, онтология содержит утверждения, привязанные к отдельным точкам зрения. Точки зрения моделируются с использованием шаблона проектирования «Descriptions and Situations» и базируются на онтологии верхнего уровня DOLCE+DnS Ultralite [8-10].

В настоящее время существуют следующие виды точек зрения:

1. Определения: с точки зрения одного определения некоторое понятие определяется через одно понятие, а с точки зрения другого определения, оно определяется через другое понятие.

2. Образовательные уровни.

Разработка иерархии точек зрения находится на начальном этапе.

Отношения

Онтология содержит следующие отношения:

1. *Отношение «целое-часть»* – данное отношение указывает, из каких элементов состоит геометрическая фигура. Например, «многоугольник» (целое) имеет части «сторона многоугольника», «угол многоугольника», «вершина многоугольника».

2. *Отношение «определяется»* связано с элементами, которые определяют, но не являются частью самой фигуры. Например, окружность определяется центром окружности и ее радиусом.

3. *Отношение онтологической зависимости* связывает ролевое понятие с понятием, от которого данная роль зависит. Например, «фокус эллипса» онтологически зависит от понятия «эллипс», а «хорда окружности» зависит от понятия «окружность».

4. *Отношение «теорема-свойство»* показывает, какими свойствами, представленными в школьном курсе в виде теорем, обладает геометрическое понятие. Например, концепт «треугольник» связан отношением «теорема-свойство» с концептами «соотношения между углами и сторонами треугольника», «неравенство треугольника», «теорема о сумме острых углов прямоугольного треугольника»; концепт «биссектриса треугольника» – с концептом «свойство биссектрисы угла треугольника» и т. п.

5. *Отношение «теорема-признак»* связывает геометрическое понятие с его признаками, представленными теоремами. Например, концепт «ромб» связан отношением «теорема-признак» с концептом «признак ромба» и др.

6. *Отношение «находится по формуле»* связывает геометрические фигуры с их метрическими свойствами. «Площадь треугольника» связывается с концептами «нахождение площади треугольника по формуле Герона», «нахождение площади треугольника через высоту и основание», «нахождение площади треугольника по двум сторонам и углу между ними» и т. д.

Аксиомы

Онтология содержит аксиомы классификации понятий. Например, она содержит аксиомы о том, что концепты «выпуклый многоугольник», «невыпуклый многоугольник» являются непересекающимися и их объединение равняется понятию «многоугольник».

Источники понятий

В качестве источников понятий при построении онтологии были использованы учебники по геометрии для общеобразовательных учреждений следующих авторов: 1) Атанасяна Л.С, Бутузова В.Ф. и др. (2018) [11], 2) Погорелова А.В. (2018) [12]; 3) Смирновой И.М и Смирнова В.А. (2015) [13]. Были также проанализированы учебник Шарыгина И.Ф. (2018) [14], реализующий наглядно-эмпирическую

концепцию построения школьного курса геометрии, и пособие для углубленного изучения математики в школе «Геометрия. Дополнительные главы к учебнику» авторов Атанасяна Л.С, Бутузова В.Ф. и др. (2005) [15]. Данная учебная литература позволила выделить список геометрических понятий, изучаемых на базовом и профильном уровнях обучения математике в школе (7–9 классы).

Принципы именования понятий

Именование понятий онтологии основывается на образовательном и на эмпирическом базисах. Образовательный базис: выбор имени концепта определяется тем, под каким именем он присутствует в учебной программе. Эмпирический базис: если имя концепта отсутствует в школьной программе, то выбор этого имени определяется на основе тестирования студентов.

Рассмотрим подробнее класс «геометрическая фигура на плоскости». Непростым стал вопрос деления этого класса на формальные подклассы, поскольку в планиметрии образование геометрических фигур идет эмпирически с 5 класса и/или аксиоматически – с 7 класса. Для данного класса были сначала отобраны понятия, удовлетворяющие утверждению: *геометрическое понятие X* – это геометрическая фигура. Таким образом, через родовое понятие «геометрическая фигура» определили около 60 понятий курса планиметрии школьного базового и профильного уровней (часть плоскости, линия, точка, кривая, ломаная, вектор, отрезок, луч, окружность, угол, многоугольник и др.), которые в свою очередь необходимо выстроить в систему родовидовых отношений (формальную таксономию).

Затруднением при построении формальной таксономии для класса «геометрическая фигура на плоскости» стало отсутствие отдельных планиметрических понятий, позволяющих разбить видовые геометрические понятия по родам. Например, «угол», «многоугольник», «полуплоскость», «круг», «сегмент», «сектор» не имеют в школьном курсе родового понятия, кроме универсального понятия «геометрическая фигура». Между тем фигуры отличаются между собой принципиально. Так, площадь угла и полуплоскости неизмеримы, поэтому они не будут связаны с понятием «площадь геометрической фигуры». Круг и многоугольник определяются как части плоскости, ограниченные замкнутой линией, поэтому в построенной онтологии мы дали классу условное короткое название – «ограниченная часть плоскости», в него включены подклассы «круг»,

«многоугольник», «части круга». С термином, обозначающим родовое геометрическое понятие для понятий «угол» и «полуплоскость», существует ряд проблем методического характера. Поскольку «геометрическая фигура – это часть поверхности, ограниченная линией» [14, с. 22], то угол и полуплоскость так же «ограниченная часть плоскости», но в представлениях учащихся и студентов она не воспринимается так. Этот вывод следует из проведенного исследования методологических представлений студентов о геометрический фигурах и отношениях между ними. В нем приняли участие 72 студента Института математики и механики им. Н.И. Лобачевского Казанского федерального университета (бакалавры 3, 4 и 5 курсов педагогического отделения, магистры 1 и 2 года обучения математических направлений). Данное исследование было направлено на изучение качества сформированности геометрических понятий у обучающихся, их графических представлений, родовидовых связей между ними и классификацию геометрических понятий. Студентам было предложено три задания:

1) обозначить родовидовую связь между 12 плоскими геометрическими фигурами (например, если студенты определили, что множество треугольников принадлежит множеству выпуклых многоугольников, то должны нарисовать стрелку от понятия «треугольник» к понятию «выпуклый многоугольник»;

2) определить истинность четырех высказываний по вопросу классификации геометрических понятий (например, множество четырехугольников можно представить в виде суммы двух множеств: выпуклых и невыпуклых многоугольников);

3) написать какими геометрическими фигурами изображаются четыре геометрических понятия (например, биссектриса угла изображается на плоскости лучом).

Анализ результатов тестирования студентов показал, что только 3 студента построили родовидовую связь между понятиями «угол» и «ограниченная часть плоскости». Кроме этого студент А показал 4 из 18 возможных и три неверных родовидовых взаимосвязей; студент Н – 6 из 18 и одну ошибочную, студент И – 4 из 18 и одну ошибочную взаимосвязь. Между тем 21 респондент связал угол с понятием «неограниченная часть плоскости», которое мы ввели как пробный интуитивно понятный термин. Эту взаимосвязь показала и студентка с лучшими результатами тестирования, указавшая 15 верных родовидовых отношений (из 18). Поэтому в рамках онтологии класс «неограниченная часть плоскости» подразумевает гео-

метрическую фигуру, ограниченную незамкнутой линией и на данный момент включает в себя подклассы «полуплоскость», «угол», «внешняя область многоугольника». Вопрос о том, как математически правильно назвать данный класс так, чтобы он был удобен для обучения планиметрии, пока остается открытым.

Построение родовидового дерева геометрических фигур на плоскости создает возможности для построения и использования для осознанного запоминания учащимися и студентами этих понятий.

Вторым пробелом стал вопрос различной трактовки определений отдельных геометрических фигур. В учебниках предлагаются следующие определения:

1) *многоугольник* – простая замкнутая ломаная [12, с. 169];

2) *плоский многоугольник* или *многоугольная область* – конечная часть плоскости, ограниченная многоугольником [12, с. 170];

3) *многоугольник* – фигура составленная из отрезков *AB*, *BC*, …, *FA* так, что смежные отрезки не лежат на одной прямой, несмежные отрезки не имеют общих точек [11, с. 98];

4) фигуру, состоящую из сторон многоугольника и его внутренней области, так же называют *многоугольником* [11, с. 99];

5) замкнутая ломаная, не имеющая самопересечений, ограничивает *многоугольник* [14, с. 59];

6) фигура, образованная простой замкнутой ломаной и ограниченная ею внутренней областью, называется *многоугольником* [13, с. 34].

Определение позволяет найти положение геометрического понятия в таксономии и в дальнейшем построить многочисленные отношения между ним и другими понятиями, поэтому такое разнообразие требует приведения определения к единому стандарту. Мы остановились на определениях, связывающих многоугольник с частью плоскости, поскольку нельзя измерить площадь ломаной.

Заключение

На данном этапе развития онтологии OntoMathEdu выделяются и другие аксиомы и отношения для формализации семантических связей между геометрическими понятиями. Происходит качественный анализ различных определений концептов как на уровне представлений для младших школьников, так и для профильной математической подготовки учащихся.

Условиями построения таксономии по элементарной математике для учащихся средней школы и будущих учителей математики являются:

– выделение и дифференциация концептов, выявление всевозможных семантических связей между ними для использования в генерировании обучающих систем с возможностью автоматического создания тестов;

– проектирование таксономии геометрических понятий, которую можно использовать в обучении в соответствии с ФГОС, ее апробация, коррекция и выявление ее обучающего потенциала;

– существование постоянной возможности расширения таксономии с удобным алгоритмом фильтрации понятий.

Современные требования к образованию требуют создания новых подходов и методов, которые будут эффективно решать проблемы как традиционно существующие, так и возникающие в настоящее время. Ведущие специалисты в области образования пришли к единому знаменателю: в центр обучения необходимо ставить обучаемого и создавать ситуации, способствующие формированию определенных умений или компетенций.

Благодарности. Работа выполнена при финансовой поддержке РФФИ и Правительства Республики Татарстан в рамках научного проекта № 18-47-160007.

ЛИТЕРАТУРА

1. Nevzorova O., Zhiltsov N., Kirillovich A., Lipachev E. OntoMathPRO Ontology: A Linked Data Hub for Mathematics // Pavel Klinov, Dmitry Mouromstev (eds.) Proceedings of the 5th International Conference on Knowledge Engineering and Semantic Web (KESW 2014). Communications in Computer and Information Science, vol. 468. Springer, Cham, 2014. Pp. 105–119.

2. Elizarov A.M., Kirillovich A.V., Lipachev E.K., Nevzorova O.A., Solovyev V.D., Zhiltsov N.G. Mathematical knowledge representation: semantic models and formalisms // Lobachevskii Journal of Mathematics, October 2014, Vol. 35, No. 4. Pleiades Publishing, 2014. Pp. 348–354.

3. Nevzorova O., Zhiltsov N., Zaikin D., Zhibrik O., Kirillovich A., Nevzorov V., Birialtsev E.. Bringing Math to LOD: A Semantic Publishing Platform Prototype for Scientific Collections in Mathematics // Harith Alani, et al. (eds.) Proceedings of the 12th International Semantic Web Conference (ISWC 2013). Lecture Notes in Computer Science, vol. 8218. Springer Berlin Heidelberg, 2013. Pp. 379–394.

4. Elizarov A., Kirillovich A., Lipachev E., Nevzorova O. Digital Ecosystem OntoMath: Mathematical Knowledge Analytics and Management // Kalinichenko L., Kuznetsov S., Manolopoulos Y. (eds). XVIII

International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2016). Communications in Computer and Information Science, vol 706. Springer, Cham, 2017. Pp. 33–46.

5. Elizarov A., Kirillovich A., Lipachev E., Nevzorova O. Semantic Formula Search in Digital Mathematical Libraries // Proceedings of the 2nd Russia and Pacific Conference on Computer Technology and Applications (RPC 2017). IEEE, 2017. Pp. 39–43.

6. Елизаров А.М., Жижченко А.Б., Жильцов Н.Г., Кириллович А.В., Липачев Е.К. Онтологии математического знания и рекомендательная система для коллекций физико-математических документов // Доклады РАН. – 2016. – Т. 467. – № 4. – С. 392–395.

7. Guarino N., Welty C.A. A Formal ontology of properties. Dieng, R., Corby, O. (eds.) Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management (EKAW ‘00). Lecture Notes in Computer Science, vol. 1937. Springer, Heidelberg, 2000. Pp. 97–112.

8. Borgo S., Masolo C. Ontological Foundations of DOLCE // Roberto Poli, Michael Healy, Achilles Kameas (eds). Theory and Applications of Ontology: Computer Applications. Springer, 2010 +.

9. Borgo S., Masolo C. Foundational Choices in DOLCE // Nicola Guarino et al (eds). Handbook on Ontologies. Springer, 2009.

10. Gangemi A., Mika P. Understanding the Semantic Web through Descriptions and Situations // OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003.

11. Геометрия. 7–9 классы: учеб. для общеобразоват. организаций / Л.С. Атанасян, В.Ф. Бутузов, С.Б. Кадомцев и др. – М.: Просвещение, 2018. – 383 с.

12. Погорелов А.В. Геометрия. 7–9 классы: учеб. для общеобразоват. организаций / А.В. Погорелов – М.: Просвещение, 2018. – 240 с.

13. Смирнова И.М. Геометрия. 7–9 классы: учеб. для общеобразоват. организаций / И.М. Смирнова, В.А. Смирнов. – М.: Мнемозина, 2015. – 376 с.

14. Шарыгин И.Ф. Геометрия. 7–9 классы: учеб. для общеобразоват. организаций / И.Ф. Шарыгин – М.: Дрофа, 2018. – 364 с.

15. Геометрия. Дополнительные главы к учебнику 9 кл.: Учеб. пособие для учащихся школ и классов с углубл. изуч. математики / Л.С. Атанасян, В.Ф. Бутузов, С.Б. Кадомцев и др. – 5-е изд. – М.: Вита-Пресс, 2005. – 176 с.

УДК 81.33; 81.322.2

TEXT SUMMARIZATION USING FUZZY LOGIC

A. Sharipbay[2], A. Zulkhazhav[1], G. Bekmanova[2], T. Aidynov[2]

[1]L.N.Gumilev Eurasian national university
altinbekpin@gmail.com, sharalt@mail.ru, gulmira-r@yandex.kz, tolegen.ch@gmail.com

In this paper describes the methods of Kazakh text summarization based on the processing of sentences, determining their characteristics and applying fuzzy logic. For the experiments we used different Internet resources on the Kazakh language.

Keywords: text summarization, natural language processing, Kazakh language, fuzzy logic.

РЕФЕРИРОВАНИЕ ТЕКСТА С ИСПОЛЬЗОВАНИЕМ НЕЧЕТКОЙ ЛОГИКИ

А. Шәріпбай[2], А. Зулхажав[1], Г. Бекманова[2], Т. Айдынов[2]
[1]Евразийский национальный университет им. Л. Н. Гумилева, Астана
altinbekpin@gmail.com, sharalt@mail.ru, gulmira-r@yandex.kz, tolegen.ch@gmail.com

В настоящей работе описывается методы реферирование текстов на казахском языке на основе обработки предложений, определения их характеристик и применения нечеткой логики. Для проведения экспериментов мы использовали разные интернет ресурсы на казахском языке.

Ключевые слова: реферирование текста, обработка естественного языка, казахский язык, нечеткая логика.

1. Введение

Реферирование представляет собой создание реферата на основе свертывания и сжатия смысловых структур первичного текста. Реферирование вручную требует колоссальных человеческих ресурсов, поэтому и возникла задача создания методов автоматического реферирования [1].

В условиях быстрого роста объема информации в сети Интернет, создание методов автоматического реферирования является

очень актуальной и важной проблемой компьютерной обработки естественно-языковых текстов. В этом качестве автоматическое реферирование относится к фундаментальным технологиям искусственного интеллекта.

Автоматическое реферирование представляет собой реферат текста, извлеченный машиной для представления наиболее важной части в более короткую версию исходного текста. Это позволяет пользователю быстро понять и отсортировать большой объем информации.

Автоматическое реферирование может быть выполнено двумя способами: извлечение и абстракция. В способе извлечения осуществляется выборка предложений или фраз, имеющих высокие оценки важности, и объединение их в новый короткий текст, без изменения исходного текста. В способе абстракции применяются лингвистические методы для смыслового анализа и интерпретации текста. В этом случае из текста извлекается основное содержание и перефразируется. Большинство исследований реферирования ориентированы на методы извлечения, то есть на критерий оценки важности [2].

В этой работе мы рассматриваем способ извлечения на основе нечетких вычислений. Метод состоит из трех стадий:

1) предварительная обработка текста;

2) вычисление функции;

3) нечёткие вычисления контента.

2. Предварительная обработка и вычисление функций предложения

В качестве исходных данных для реферирования мы использовали тексты казахских новостных статей, а также созданные нами базы данных стоп слов и морфологический словарь.

Предварительная обработка включает в себе:

1) удаления лишних отступов, пробелов, знаков препинаний и других специфических символов;

2) сегментация, которая предполагает произвести разбивку текста на предложения и токенизацию каждого предложения;

3) замена личных местоимений на имена тех лиц, которые они указывают.

4) удаление стоп-слов, которые часто встречаются в тексте, но не представляют особое значение для определения важности контента.

Например, әр – каждый, осы – это, деп – говоря, және – и, немесе – или.

5) удаление окончаний путем проведения стемминга;

Обычно в процессе удаления стоп-слов личные местоимения такие, как “я”, “он”, “она”, “они” автоматически относятся к стоп-словам и удаляются из текста. Но часто, они несут определенную значимость при определении важности контента, так как могут указывать на разные лица в зависимости расположения в тексте. Поэтому путем проведения морфологического и синтаксического анализа личные местоимения заменяются на имена тех лиц, на кого они указывают.

После предварительной обработки текста необходимо произвести вычисления функций предложения, результатами которых являются вектора из семи элементов для каждого предложения. Элементы каждого вектора принимают значения в интервале [0, 1]. Мы будем рассматривать следующие функции предложения:

1) Функция заголовка (Ф1): определяется как отношение числа соответствий заголовочных слов в текущем предложении к количеству слов заголовочного предложения:

$$Ф1 = \frac{\text{Количество заголовочных слов в текущем предложении}}{\text{Общее количество слов в заголовке}} \quad (1)$$

2) Функция длины предложения (Ф2): определяется как отношение количество слов текущего предложения к количеству слов самого длинного предложения в тексте:

$$Ф2 = \frac{\text{Количество слов в текущем предложении}}{\text{Количество слов самого длинного предложения в тексте}} \quad (2)$$

Эта функция необходима для фильтрации из выборки коротких и неполных предложений, такие как автор статьи, дата статьи и т.п.

3) Функция позиции предложения (Ф3): определяется как максимум следующих двух отношений:

$$Ф3 = \text{Максимум}\left(\frac{1}{\text{позиция пред.}}; \frac{1}{\text{количество всех предл.} - \text{позиция предл.} + 1}\right) \quad (3)$$

В формуле (3) если предложение находится в начале текста, то первое выражение является максимумом, если предложение находится в конце текста, то максимум значение примет второе выражение. Эта функция имеет важное значение при выборке, так как более

информативные предложения обычно располагаются в начале или в конце текста.

4) Функция тематического слова (Ф4): определяется как отношение количества тематических слов в текущем предложении к максимуму количества тематических слов, вычисляемом. по всем предложениям текста:

$$Ф4 = \frac{\text{Количество тематических слов в предложении}}{\text{Максимум (Количество тематических слов в каждом предложении)}} \quad (4)$$

Под тематическими словами понимаются наиболее часто употребляемые слова в тексте. Они имеют прямое отношение к основной тематике текста. В качестве тематических мы выбрали пять самых часто встречающихся слов в тексте

5) Функция веса предложения (Ф5): определяется как отношение суммы частот появления терминов в предложении (текущих терминов) к сумме частоты появления терминов в тексте:

$$Ф5 = \frac{\text{Сумма (Частота появления термина в предложении)}}{\text{Сумма (Частота появление текущих терминов в тексте)}} \quad (5)$$

Чтобы вычислить вес предложения, мы находим частоту появления термина в предложении и частоту появления этого же (текущего) термина в тексте:

6) Функция имен собственных (Ф6): определяется как отношение количества собственных имен в предложении к длине предложения:

$$Ф6 = \frac{\text{количество собственных имен в предложении}}{\text{длина предложения}} \quad (6)$$

Собственные имена, встречающиеся в предложении, несут в себе много информации о персональных фактах. Поэтому предложения с наиболее количеством собственных имен являются важной частью контента.

7) Функция числовых данных (Ф7): определяется как отношение количества числовых данных в предложении к длине предложения:

$$Ф7 = \frac{\text{количество числовых данных в предложении}}{\text{длина предложения}} \quad (7)$$

Обычно числовые данные имеют конкретные важные значения для реферирования. Поэтому числовых данных в тексте нельзя пропускать.

3. Нечеткие вычисления контента

Нечеткие вычисления состоит в извлечении наиболее важного контента с применением нечеткой логики, в которой используются следующие понятия: нечеткое множество, функция принадлежности, нечеткие логические операции, лингвистические переменные, лингвистические термы, нечеткие логические значения, нечеткий логический вывод [3].

Нечеткие вычисления состоит из следующих этапов:

1) фазификация;

2) логический вывод по база нечетких знаний;

3) дефазификация.

При фазификации определяется соответствия между четким численным значением входной переменной и значением функции принадлежности соответствующего ей терма лингвистической переменной. В нашем случае лингвистическими переменными выступают имена семи функций, определенных нами выше. Они принимают значения из множества слов такие, как “незначительный”, “низкий”, “средний”, “высокий”, “очень высокий”. Эти слова называются терм-множествами и принимают значения в интервале [0,1] (рисунок 2). Одним словом, фазификация – это процесс перехода от четкого представления к нечеткому [4]

Процесс фазификации зависит от функции принадлежности для соответствующих лингвистических термов. Одной из основных проблем применения нечеткой логики является выбор функций принадлежности лингвистических переменных. Основными видами функций принадлежности являются треугольные, трапециевидные, кусочно-линейные, гауссовы, сигмоидные и другие функции. Выбор функции принадлежности конкретной переменной представляет собой плохо формализованную задачу, решение которой основано на интуиции и опыте [6]. Для нашей задачи мы предпошли более подходящей треугольную функцию принадлежности, использующейся для задания неопределенностей типа: «приблизительно равно», «среднее значение», «расположен в интервале», «подобен объекту», «похож на предмет» и т.п.

Качество нечеткого вывода зависит от правильного построения “если-то” правил. Мы получили правила для нечеткой базы знаний на основе анализа ручных рефератов.

Пример из правил нечеткой базы знаний:

Правило1 = ЕСЛИ((заголовок['средний'] ИЛИ заголовок['хо-

роший']) И (тематическое['средний'] ИЛИ тематическое['хороший'])) ТО важность['высокая'])

Поскольку все функции принадлежности лингвистических переменных нам известны, и определены нужные нам правила, переходим к процессу агрегации. Агрегирование представляет собой процедуру определения степени истинности условий по каждому из правил системы нечеткого вывода. При этом используются полученные на этапе фазификации значения функций принадлежности термов лингвистических переменных. Если условие нечеткого продукционного правила является простым нечетким высказыванием, то степень его истинности соответствует значению функции принадлежности соответствующего терма лингвистической переменной. Если условие представляет составное высказывание, то степень истинности сложного высказывания определяется на основе известных значений истинности составляющих его элементарных высказываний при помощи, введенных ранее нечетких логических операций [4].

После логического вывода, путем обращения к нечеткой базе знаний получаем нечеткие значения и с помощью дефазификации нечетких значений лингвистических переменных получаем четкие значения для вывода (рисунок 1).

Дефазификацией (defuzzification) называется процедура преобразования нечеткого множества в четкое число. В теории нечетких множеств процедура дефазификации аналогична нахождения характеристик положения (математического ожидания, моды, медианы) случайных величин в теории вероятности. Простейшим способом выполнения процедуры дефазификации является выбор четкого числа, соответствующего максимуму функции принадлежности [4]

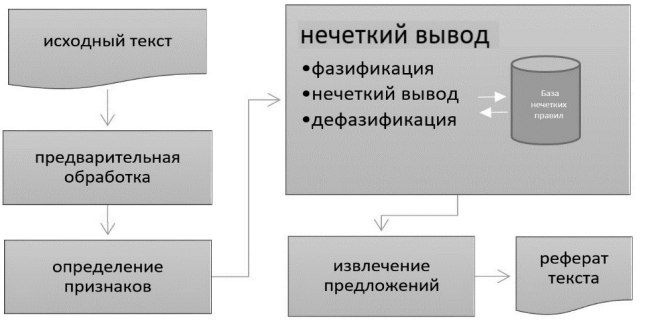


Рис. 1. Реферирование текста с применением нечёткой логики

Для программной реализации реферирования текста на основе нечеткой логики мы использовали язык python и пакет skfuzzy [7]. Мы построили функцию принадлежности для каждой значений функции из пяти нечетких множеств: незначительный, низкий, средний, высокий, очень высокий. Пример функции принадлежности для функции заголовка (рисунок 2)

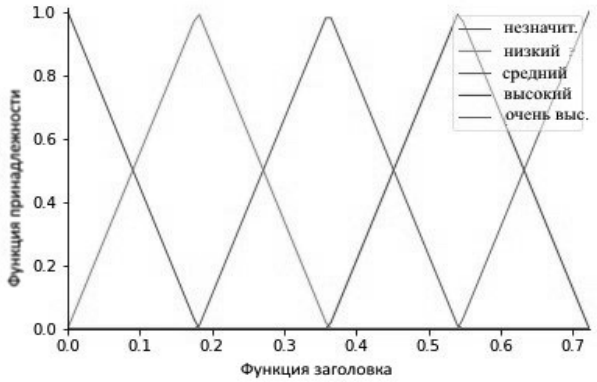


Рис. 2. Лингвистическая переменная «функция заголовка»

Последним шагом нечеткого вывода является дефазификация, то есть выходная функция принадлежности, которую мы разбили на три: незначительный, средний, важный (рисунок 3).

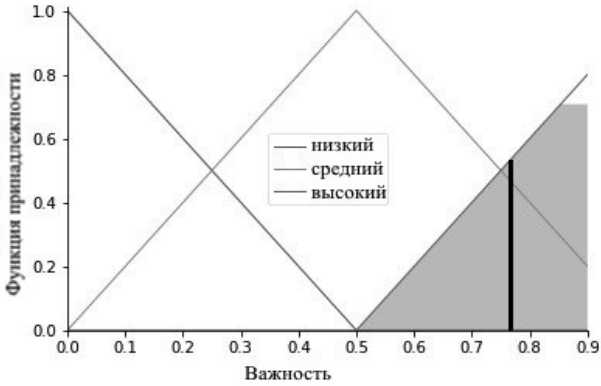


Рис. 3. Лингвистическая переменная «функция заголовка»

Пример из кода:

```
def get_importance(functions):
    functions_list = functions [0]
#определяем входные функции принадлежности для лингвистических переменных
    topic = ctrl.Antecedent(np.arange(0, 1, 0.01), 'topic')
    topic.automf(5)
    # выходная функция принадлежности для дефаззификации
    importance['low'] = fuzz.trimf(importance.universe, [0, 0, 0.5])
    importance['medium'] = fuzz.trimf(importance.universe, [0, 0.5, 1])
    importance['high'] = fuzz.trimf(importance.universe, [0.5, 1, 1])
    #определяем «если-то» правила для нечеткой базы знаний
    #rule1 = ctrl.Rule(termfreq['poor'] & topic['poor'] & thematic['poor'] , importance['low'])
    #rule2 = ctrl.Rule(termfreq['good'] & topic['good'] & thematic['good'] , importance['high'])
#модуль логического вывода
importance_ctrl = ctrl.ControlSystem([rule3, rule2, rule1])
importance_val = ctrl.ControlSystemSimulation(importance_ctrl)
#получаем нечеткий вывод
importance_val.compute()
#путем дефазификации получаем четкое значение
val = importance_val.output['importance']
```

4. Результаты

Для предварительной оценки качества работы алгоритмы были вычислены показатели ROUGE-L (наибольшая общая подпоследовательность – задача поиска последовательности, которая является подпоследовательностью нескольких последовательностей), ROUGE-2, ROUGE-1 нескольких обзорных рефератов полученных из новостных статей на казахском языке (источник статей https://www.kt.kz) (рисунок 4), для которых имеется ручная аннотация (таблица 1):

Таблица 1. **Результаты работы программы автоматического реферирования**

Ручная аннотация	Автоматический реферат	Вес предложения (показатель важности)	Вектор функций
Биылғы мамыр-шілде айларында 2 миллион тоннадан астам көмір тасымалданды. Бұл өткен жылдың сәйкес мерзімімен салыстырғанда 35 % жоғары көрсеткіш. Маусымда 850 мың тонна көмір тасымалданса, ол өткен жылдың сәйкес мерзімімен салыстырғанда 74% жоғары көрсеткішті құрады.	Жыл басынан бері 144 миллион тонна жүк тиеліп, өткен жылмен салыстырғанда 7% өсті	0.76666-высокий	[0.31, 0.62, 0.5, 1.0, 1.0, 0.25, 0.25]
	Биылғы мамыр-шілде айларында 2 миллион тоннадан астам көмір тасымалданды.	0.766127-высокий	[0.0, 0.77, 0.25, 0.0, 0.62, 0.1, 0.2]
	Маусымда 850 мың тонна көмір тасымалданса, ол өткен жылдың сәйкес мерзімімен салыстырғанда 74% жоғары көрсеткішті құрады	0.766127-высокий	[0.0, 0.46, 0.16, 0.33, 0.45, 0.17, 0.17]

Показатели ROUGE-N представляет собой статистическую меру, выражающую какой процент лексических единиц (N-gram,- последовательностей из N лексем), входящих в состав ручной, построенной независимым экспертом, аннотации, попадает в обзорный реферат 5]

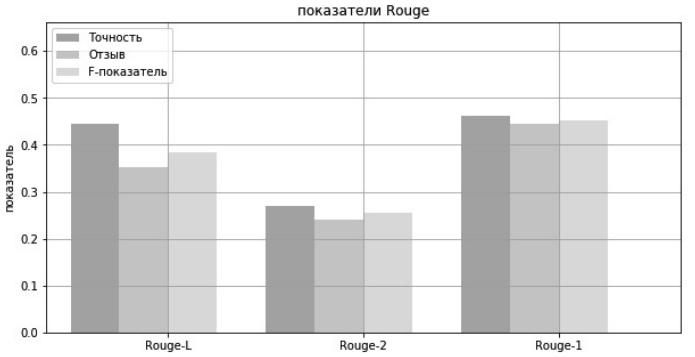


Рис. 4. График Rouge показателя примера из новостной статьи Kazakhstan today (www.kt.kz)

Заключение

В нашей стране казахский язык является государственным языком, и сфера его применения быстро расширяется. Поэтому результаты работы будут очень востребованными для быстрого восприятия обществом краткого содержания большого потока информации. Алгоритм экстрактивного метода реферирования с применением нечеткой логики показал себя эффективной для задач автоматического реферирования корпусов новостных сюжетов на казахском языке. Все же алгоритм требует усовершенствования и включения в алгоритм дополнительных методов, которые позволят не только извлечь важный контент, но и перефразировать, чтобы более соответствовать к ручному реферату. Также стоит задача сформировать базу данных статей с ручным аннотированием, так как набора данных на казахском языке ничтожно мало для большего проведения экспериментов и улучшения качества автоматического реферирования.

Благодарности. Работа выполнена при содействии коллег из Университета Техаса в Сан-Антонио: Ramin Sahba, professor Mo Jamshidi (https://acelab.wixsite.com/acelab)

REFERENCES

1. https://methodological_terms.academic.ru/1639/РЕФЕРИРОВАНИЕ_ТЕКСТА

2. Ramin Sahba, Nima Ebadi, Mo Jamshidi, Paul RadAutomatic. Text Summarization Using Customizable Fuzzy Features and Attention on the Context and Vocabulary. 2018 World Automation Congress (WAC), pp. 68-73, 2018. https://ieeexplore.ieee.org/abstract/document/8430483/?part=1;

3. Заде Л. Понятие лингвистической переменной и его применение к принятию приближенных решений. – М.: Мир, 1976. – 166 с.;

4. СД Штовба. Введение в теорию нечетких множеств и нечеткую логику. http://matlab.exponenta.ru/fuzzylogic/book1/index.php

5. С.Д. Тарасов. Алгоритм ранжирования связанных структур для задачи автоматического составления обзорных рефератов новостных сюжетов. XI Национальная конференция по искусственному интеллекту с международным участием (КИИ-2008);

6. М.Г. Семененко, И.В. Князева, С.И. Черняев. Проблемы выбора функций принадлежности нечетких множеств. Современные проблемы науки и образования. – 2013. – № 5.;

7. Набор алгоритмов нечеткой логики, предназначенных для использования в стеке SciPy, написанный на языке программирования Python. https://pythonhosted.org/scikit-fuzzy/overview.html.

УДК 81’33, 81.512.1

ABOUT THE PARAMETERS OF TEXT COMPLEXITY IN THE TATAR LANGUAGE (ON THE EXAMPLE OF EDUCATIONAL TEXTS)

I. I. Fatkhullina, B. E. Khakimov
Kazan Federal University, Kazan
fathulich@gmail.com, khakeem@yandex.ru

The article describes the existing approaches and methods for assessing the text complexity. Taking into account the typological features of the Tatar language, subjective and objective parameters for determining the complexity of the Tatar texts are proposed. The example of experimental evaluation using the questionnaire on the material of educational texts is given.

Keywords: text complexity, Tatar language.

К ВОПРОСУ О ПАРАМЕТРАХ СЛОЖНОСТИ ТЕКСТА В ТАТАРСКОМ ЯЗЫКЕ (НА ПРИМЕРЕ УЧЕБНЫХ ТЕКСТОВ)

И. И. Фатхуллина, Б.Э. Хакимов
Казанский федеральный университет, Казань
fathulich@gmail.com, khakeem@yandex.ru

В статье охарактеризованы существующие подходы к оценке сложности текста. Предлагаются возможные субъективные и объективные параметры определения сложности текста для татарского языка, в том числе, с учетом типологических особенностей языка. Представлен пример экспериментальной оценки предложенных параметров с помощью анкетирования на материале учебных текстов.

Ключевые слова: сложность текста, татарский язык.

Проблема изучения сложности текста в разных аспектах представляет актуальное направление современной лингвистики. Практическую значимость подобных исследований учебных текстов трудно переоценить, так как регулирование сложности текста для тех или иных обучающихся помогало бы им лучше реализовать свои познавательные способности. Главное требование к учебным текстам – это понятность и доступность информации, передаваемой в тексте. В то же время, некоторые авторы учебников понимают диа-

лектическую связь понятности и научности односторонне и считают, что понятность текста ограничивает его научность. К сожалению, зачастую критерии определения сложности учебного текста ограничиваются профессиональным опытом и чутьем человека. Однако для объективной оценки сложности текста необходим обоснованный выбор, конкретные параметры и критерии определения сложности, подходящие выбранному языку и стилистике текста. В свою очередь, это позволяет ставить и решать задачи автоматизации оценки уровня сложности текста с помощью специальных компьютерных программ.

Сложность является важной прагматической характеристикой текста, состоящего из множества элементов, объединенных различного рода связями. В XX веке при рассмотрении понятия сложности текста встречаются такие синонимичные понятия, как удобочитаемость, читабельность, трудность текста и благозвучие. Они отражают, насколько удобным для зрительного либо слухового восприятия является текст, а факторами выступают размер букв, цвет шрифта и фона, наличие жаргонизмов и неологизмов и т. п. [5].

С началом XXI века при описании сложности текста стал превалировать подход, оперирующий общей идеей понимания сложности как количества затрачиваемых ресурсов для описания какого-либо объекта. Так выделяют два фактора сложности текста: 1) объективный (абсолютный); 2) субъективный (относительный). К объективным факторам сложности текста относят: абстрактную, лексическую и синтаксическую сложность. К субъективным факторам относятся: индивидуальные, возрастные факторы и информативность. Впрочем, выделение субъективных и объективных факторов сложности текста довольно условно, так как, например, информативность и абстрактность текста могут быть оценены как относительно, так и безотносительно к субъекту [5].

Статистические параметры сложности текста обычно представлены на нескольких уровнях. Например, согласно [5], первый – макроуровень, здесь рассматривается уровень текста, абзаца. В него входят такие параметры, как длина текста в абзацах, длина текста в словах, длина текста в буквах, средняя длина текста (в словах или буквах). Второй уровень – синтаксический: средняя длина предложения в фразах, средняя длина предложения в словах или в слогах, средняя длина предложения в буквах. Третий уровень – лексический со следующими параметрами: средняя длина слова в буквах, процент простых слов, процент сложных слов, процент неповторяющихся

слов, средняя частота повторяющихся слов, процент определенных частей речи (существительное, прилагательное, глагол).

Для определения сложности учебного текста на татарском языке также можно выделить субъективные и объективные факторы. К субъективным факторам относятся индивидуальные (уровень владения татарским языком), возрастные (собственно возраст и класс, курс обучения), а также фактор уровня знаний (читатель должен обладать определенными базовыми знаниями по теме, чтобы понимать содержание текста). К объективным факторам на данном этапе мы отнесли следующие параметры: синтаксические – средняя длина предложения в словах, количество простых предложений, количество сложных предложений, количество т.н. "синтетических" и "аналитических" придаточных предложений; лексические – средняя длина предложения в слогах, процент простых слов, процент сложных слов. В группу простых слов в первую очередь входят слова с конкретным значением, наиболее употребительные в языке и встречающиеся в бытовом общении. К сложным словам отнесены термины и слова с абстрактным значением. Так как татарский язык является агглютинативным языком и морфемный состав слова имеет важное значение, выделяется также и морфологические параметры – количество аффиксов, присоединяемых к основе.

С целью апробации предложенных параметров сложности и выявления корреляции между объективными параметрами сложности текста и субъективной оценкой трудности со стороны учеников, нами был осуществлен сравнительный анализ учебных текстов из учебников для 6–11 класса на татарском языке. Методом анкетирования учащихся школ РТ получены статистические данные о восприятии сложности текстов обучающимися. Далее тексты были подвергнуты лингвистическому сравнительному анализу по объективным параметрам.

Эксперимент проводился при помощи метода анкетирования в средней общеобразовательной школе села Большие Ключи Зеленодольского района Республики Татарстан в апреле 2018 года. В эксперименте приняли участие 69 учеников 6, 7, 8, 9, 10 классов. В ходе эксперимента участникам предлагались случайным образом отобранные отрывки из учебников для средних общеобразовательных школ на татарском языке. Объем текста составлял 10–12 предложений. Для точности эксперимента для каждого класса были отобраны тексты по гуманитарным предметам (история или обществознание) и по естественным наукам (биология или география).

Эксперимент был организован следующим образом: 1) обучающиеся читают текст 1 раз; 2) в ходе чтения выделяют незнакомые слова; 3) оценивают сложность текста по 5 бальной шкале: 1 очень легко, 2 – легко, 3 – средне, 4 – сложно, 5 – очень сложно; 4) результаты анкетирования обрабатываются, результаты фиксируются в таблицу; 5) текст анализируется по объективным статистическим параметрам; 6) осуществляется сравнение объективных и субъективных параметров; 7) подводятся итоги.

В качестве примера приведем некоторые результаты анализа текстов на татарском языке для 6 класса. В 6 классе в анкетировании приняли участие 19 учеников, их возраст составил 12–13 лет. Участникам эксперимента было предложено два отрывка из текстов по географии [4] и истории [2].

По оценке учеников, средняя сложность первого текста составила 2,4 по пятибалльной шкале. Всего незнакомых слов 46 слов, среднее количество сложных слов – 2,5. Объективные параметры текста: текст состоит из 10 предложений, 142 слов и 371 слогов. Слов с более чем 3 аффиксами встретилось всего 8. В одном предложении в среднем 14,2 слов. В тексте 5 сложных предложений. Сложных слов – 21, простых слов – 121.

Средняя сложность второго текста составила 2,6 по пятибалльной шкале. Незнакомых слов 39, это в среднем для каждого ученика по 2 слова. Объективные параметры текста: текст состоит из 12 предложений, 155 слов, 429 слогов. Слов с более чем 3 аффиксами – 8. Сложных предложений – 3. Количество простых слов равно 120, сложных – 35.

Сравнив показатели двух текстов, можно сделать следующие выводы:

1) по субъективным оценкам (по мнению учеников), второй текст сложнее;

2) второй текст объемнее, соответственно, количество сложных слов, как и общее количество слов, в нем больше;

3) в первом (естественнонаучном) тексте неизвестных ученикам слов больше (первый текст – в среднем 2,5 слова, второй текст – 2 слова). На это мог повлиять и тот факт, что при чтении второго текста внимание учеников снижается, следовательно, тексты лучше было бы предлагать в разное время, но в аналогичных условиях;

4) во втором тексте больше предложений, но они короче: средняя длина предложения в первом тексте – 14,2 слова, во втором тексте – 12,9 слова;

5) в первом тексте больше сложных предложений: 5 против 3 предложений во втором тексте.

Итак, путем сравнения субъективных и объективных факторов определения сложности текста, мы можем сделать вывод, что второй текст сложнее. С одной стороны, он сложнее по мнению учеников, с другой стороны – больше по объему, а также содержит больше терминов, многие из которых не являются широко употребительными.

Результаты анализа других текстов, использованных в эксперименте, также обнаруживают корреляцию между субъективной оценкой трудности со стороны учеников и объективными параметрами сложности текста. Таким образом, предложенные параметры могут быть взяты за основу при изучении сложности татарских текстов, но для их корректировки необходима выборка текстов большего объема.

На современном этапе изучения сложности текста предпринимаются попытки автоматизации процесса ее расчета, появляются онлайн-сервисы, позволяющие оценить сложность вводимого пользователем текста. В дальнейшем такие программы необходимо реализовать и для татарского языка.

ЛИТЕРАТУРА

1. Miestamo, M. Grammatical complexity in a cross-linguistic perspective. In: Language complexity: Typology, Contact, Change. – pp. 23–42. John Benjamins, Amsterdam (2008).

2. Бойцов М.А., Шукуров Р.М. Гомуми тарих. Урта гасырлар тарихы : татар урта гомуми белем бирү мәктәбенең 6 нчы сыйныфы өчен дәреслек / М.А. Бойцов, Р.М. Шукуров. – Казан: Мәгариф; Москва: Русское слово, 2008. – 318 б.

3. Кисельников А.С. Экзаменационный текст: сущность, специфика, функции (на материале русского и английского языков): дис. ... канд. филол. наук / А.С.Кисельников. – Казань, 2017. – 243 с.

4. Коринская В.А. h.б. Материклар географиясе: 6 сыйныф өчен дәреслек / В.А. Коринская, Л.Д. Прозоров, В.А. Щенев. – Казан: Татар. кит. нәшр., 1986. – 280 б.

5. Мизернов И.Ю., Гращенко Л.А. Анализ методов оценки сложности текста / И.Ю. Мизернов, Л.А. Гращенко // Новые информационные технологии в автоматизированных системах. – 2015. – №18. – С. 572–581.

6. Оборнева И.В. Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис. ... канд. пед. наук / И.В. Оборнева. – М., 2006. – 165 с.

7. Солнышкина М.И., Кисельников А.С. Параметры сложности экзаменационных текстов / М.И. Солнышкина, А.С. Кисельников // Вестник Волгоградского государственного университета. Серия 2: Языкознание. – 2015. – №1 (25). – С. 99–107.

8. Солнышкина М.И., Кисельников А.С. Сложность текста: этапы изучения в отечественном прикладном языкознании / М.И. Солнышкина, А.С.Кисельников // Вестник Томского государственного университета. Филология. – 2015. – № 6(38). – С. 86–99.

9. Татар грамматикасы: II том / проект җит. М.З. Зәкиев. – Тулыландырылган 2 нче басма. – Казан: ТӘhСИ, 2016. – Т. II. – 432 б.

УДК 004.82

METHODS AND TOOLS TO REVEAL SEMANTIC CONSTRUCTS BASED ON PROJECT DOCUMENTS PARSING OUTPUT

A. A. Kulikova, E. A. Trifonova, E. E. Rybnikova
Ulyanovsk State Technical University, Ulyanovsk
a.push1206@gmail.com, catherine21tea@gmail.com,
scarletknight390@gmail.com

The article presents methods and tools, which help to reveal semantic construct including concepts (terms) and relations of certain types in a text (particularly in a project documents) based on a parser output in order to use them in the ontologi-cal support mechanisms for software system design. In the first section some pre-requisites of the research and related works are described. The second section tells about the existing text analysers including Pullenti parser, LPaRus, Stand-ford parser and others, which help to retrieve both syntactic and semantic rela-tions from Russian and English documents. In the third section new solutions to reveal necessary semantic constructs (concepts and relations) with the help of the abovementioned instruments are presented. Some examples are considered in or-der to illustrate retrieving possible semantic meaning from pairs of words or phrases which are syntactically linked to each other. The forth section presents the tools developed within the research which retrieve part-and-whole as well as casual relations from the texts in English. The following relation types were cho-sen because retrieving them could be useful when building up an architectural model of a software system or its logical structure correspondently. The ad-vantages of these tools for software design are also described. In the end, some conclusions are conducted.

Keywords: project ontology, semantics, semantic relation, syntactic relation, automatic text analysis, software system design, computer-aided design.

МЕТОДЫ И СРЕДСТВА ОБНАРУЖЕНИЯ СЕМАНТИЧЕСКИХ КОНСТРУКТОВ В РЕЗУЛЬТАТАХ ПАРСИНГА ПРОЕКТНЫХ ДОКУМЕНТОВ

А. А. Куликова, Е. А. Трифонова, Е. Е. Рыбникова
Ульяновский государственный технический университет, Ульяновск
a.push1206@gmail.com, catherine21tea@gmail.com,
scarletknight390@gmail.com

В статье рассматриваются методы и средства, позволяющие обнаруживать семантические конструкты – концепты и следы семантических

связей некоторых типов в результатах автоматического синтаксического анализа (парсинга) текстов рассуждений и проектных документов с целью дальнейшего их использования в механизмах онтологического сопровождения процесса проектирования автоматизированных систем. Приводятся существующие инструменты для автоматизированного анализа русскоязычных и англоязычных текстов на предмет наличия в них синтаксических и семантических связей, а также описываются собственные решения для извлечения необходимых связей с использованием данных инструментов.

Ключевые слова: проектная онтология, семантика, семантическая связь, синтаксическая связь, автоматический анализ текстов, автоматизированная система, системы автоматизации проектирования.

Введение

В сфере разработки сложных автоматизированных систем (АС) с программным обеспечением в настоящее время наблюдается ряд серьезных проблем. Одна из них – низкий процент успешности. Согласно исследованиям Standish Group [1], в середине 90-х годов всего 16% всех разрабатываемых в мире систем были успешно завершены. Сейчас этот показатель достиг 40%, но, тем не менее, такие цифры нельзя назвать удовлетворительными.

Важнейшей причиной низкого процента успешности разработок являются негативные проявления человеческого фактора, важным источником которых является феномен понимания, отвечающий за корректность употребления языка. Некорректность приводит к дорогостоящим семантическим ошибкам в рассуждениях и документах, а через них – в проектных решениях и их материализациях в проектах АС.

Соответственно, успех разработки АС существенным образом зависит от адекватности и качества понимания ситуаций, событий и связывающих их системных отношений в процессах решения проектных задач – в первую очередь, на этапах концептуального проектирования.

Следовательно, акты понимания целесообразно регистрировать на протяжении всего процесса концептуального проектирования и постоянно анализировать на предмет несоответствия выводам, сделанным ранее. Для этого предлагается использовать проектные онтологии, которые призваны отражать концептуальное пространство, в котором работает проектировщик, т. е. содержать информацию о концептах (терминах), относящихся к проекту, и семантических связях между ними. Идеи применения инструментов онтологического

сопровождения и контроля процесса проектирования частично описаны в [2], [3] и [4]. В данной же статье рассмотрим более подробно процесс извлечения «следов» семантических конструктов, под которыми мы понимаем как концепты, так и семантические связи различных типов («часть – целое», «причина – следствие», «род – вид» и т. д.), из текстовых документов (в частности, с использованием результатов автоматического синтаксического анализа), что является значимой задачей в рамках создания инструмента онтологического контроля проектных решений. Наш инструмент предполагает работу с русским и английским языками, поэтому рассмотрим примеры для обоих языков.

Инструменты автоматического обнаружения синтаксических и семантических связей в русско- и англоязычных текстах

Методы автоматического анализа естественного языка и компьютерные технологии обработки информации – это наиболее динамично развивающиеся области компьютерной лингвистики в наши дни. Тем не менее, в этих областях остаются задачи, до сих пор не нашедшие общепризнанного решения, и среди них – синтаксический анализ, наиболее сложный этап досемантической обработки текста.

На сегодняшний день известны различные синтаксические анализаторы (парсеры) – например, Stanford Parser, разработанный для анализа английского языка [5].

В области автоматической обработки русского языка на уровне синтаксиса и семантики накоплен богатый и разнообразный опыт: так, в рамках конференции «Диалог-2016» был представлен инструмент Pullenti [6], позволяющий извлекать данные из неструктурированных текстов. А также в начале 2000-х годов получил известность проект «Диалинг» (АОТ) [7], который базировался на доступном синтаксическом анализаторе с открытой документацией. На данный момент это не единственные существующие парсеры, однако рассмотрим их в качестве примера.

Stanford University Parser представляет собой вариант вероятностного парсера естественного языка. Основным языком для анализа является английский, однако, синтаксический анализатор может быть адаптирован для работы и с другими языками. Парсер может считывать текст и выводить результат анализа в различных форматах: текст с частеречной разметкой, деревья синтаксической структуры и списки грамматических отношений.

Pullenti – это SDK для информационных систем, имеющих дело с текстами на естественном языке. Имеются функционально эквивалентные библиотеки на .NET Framework 4.0, .NET Core 2.0, Java и Python. Инструмент позволяет осуществлять семантический анализ текста и выводить результаты в виде двух типов графов объектов, представленных на рис. 1.

Для наглядного примера работы инструмента рассмотрим следующее предложение: «*В разработке комплекса средств необходимо ориентироваться на теоретически обоснованное представление понятийного ядра проекта в виде проектной онтологии*».

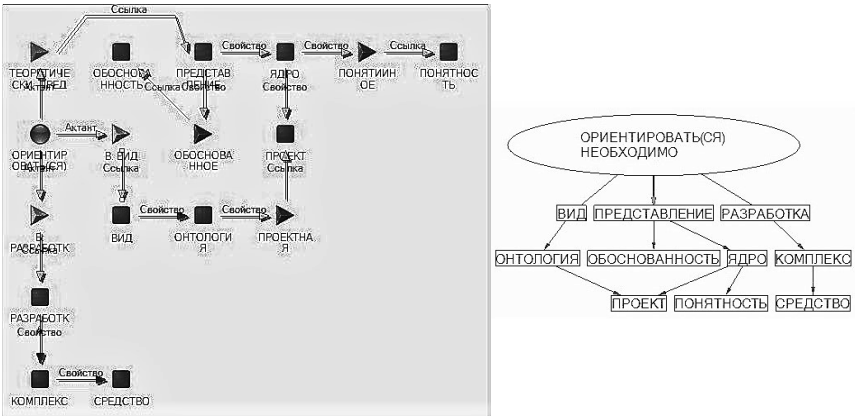


Рис. 1. Граф семантических объектов, построенный с помощью анализатора Pullenti

Технология АОТ (Автоматическая обработка текста) базируется на многоуровневом представлении естестевенного языка (морфология, синтаксис, семантика). Кроме того, программа позволяет анализировать морфологические атрибуты слов, осуществлять синтаксический разбор, строить поверхностно-семантический граф, осуществлять автоматический перевод с русского языка на аглийский, а также поиск по массивам и леммным биграммам.

Для примера приведём синтаксический разбор того же предложения (см. рис 2).

Помимо этого, разработан ряд других интересных парсеров, среди которых – парсер русского языка LPaRus, разработанный на основе лингвистических технологий компании Megaputer Intelligence [8],

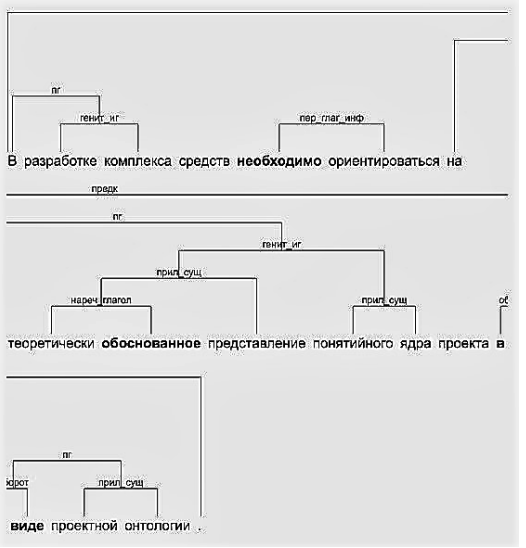


Рис. 2. Синтаксическое дерево, построенное с помощью анализатора АОТ

семантико-синтаксический анализатор SemSin [9], Томита-парсер от Yandex [10], извлекающий данные при помощи контекстно-свободных грамматик и словарей ключевых слов. Все они дают нам богатую основу для создания собственного инструмента по анализу семантики в прикладных целях.

Обнаружение следов семантических связей в результатах парсинга на примере англоязычного текста

Одним из способов выявления некоторых семантических конструктов, содержащихся в тексте, может послужить анализ его синтаксической структуры. В английском языке такой анализ осуществляет программа Stanford University Parser. Рассмотрим работу парсера на примере следующего текста, представляющего собой постановку задачи:

In order to improve the quality of conceptual activity of designers who develop systems with software we need to create an integrated set of tools, which provides generating and using a project language. The project language helps to detect and prevent semantic mistakes, particularly, by achieving necessary and sufficient understanding and recording the language for further application.

Developing the integrated set of tools, we should base on the theoretically justified representation of the conceptual core in form of project ontology. Also, we aim at a higher degree of automation of designer's actions, particularly by using multi-agent support.

Вначале парсер строит структуру каждого предложения текста в виде деревьев (часть такого разбора представлена на рис. 3), а затем перечисляет все типы связей, обнаруженные в обработанных высказываниях.

Parse

```
(ROOT
  (S
    (SBAR (IN In) (NN order)
      (S
        (VP (TO to)
          (VP (VB improve)
            (NP
              (NP (DT the) (NN quality))
              (PP (IN of)
                (NP
                  (NP (JJ conceptual) (NN activity))
                  (PP (IN of)
                    (NP
                      (NP (NNS designers))
                      (SBAR
                        (WHNP (WP who))
                        (S
                          (VP (VBP develop)
                            (NP (NNS systems))
                            (PP (IN with)
                              (NP (NN software))))))))))))))))
    (NP (PRP we))
    (VP (VBP need)
      (S
        (VP (TO to)
          (VP (VB create)
            (NP
              (NP (DT an) (VBN integrated) (NN set))
              (PP (IN of)
                (NP (NNS tools)))
              (SBAR
                (WHNP (WDT which))
                (S
                  (VP (VBZ provides)
                    (S
                      (VP (VBG generating)
                        (CC and)
                        (VBG using)
                        (NP (DT a) (NN project) (NN language))))))))))))
    (. .)))
```

Рис. 3. Синтаксическое дерево, построенное с помощью Standford Parser

При этом оказывается, что некоторые связи заключают в себе определённую семантику. Так, в ходе анализа был обнаружен ряд

синтаксических отношений, которые могут быть использованы для фильтрации концептов и выявления семантических связей между ними.

Первый такой тип синтаксической связи обозначается как *adjectival modifier* с тэгом *amod*. Данный тип связи служит для указания на определение объекта – например, *amod(mistakes-10, semantic-9), amod(core-19, conceptual-18)*. Следовательно, он помогает выявлять свойства или атрибуты концептов проектной онтологии.

Следующий тип синтаксической связи – *nominal subject* (*nsubj*) – определяет сказуемое и подлежащее предложения. Примеры: *nsubj(develop-13, designers-11), nsubj(helps-4, language-3)*. Подлежащие играют значимую роль в тексте, и поэтому зачастую они могут представлять собой потенциальные концепты для занесения в проектную онтологию. Также существует отдельный тип связи, объединяющий подлежащее и сказуемое в страдательном залоге – *passive nominal subject* (*nsubjpass*). Данная связь таким же образом выявляет субъекты высказываний.

В некоторых случаях отношение *nominal subject* указывает на местоимение в качестве субъекта – например, *nsubj(provides-27, which-26), nsubj(develop-13, who-12)*. Очевидно, что подобные местоимения должны быть связаны с каким-либо объектом, который, в свою очередь, можно рассматривать в роли концепта. В используемом парсере также присутствует тип связи, называемый *referent* (*ref*). *Referent* отвечает за связь между местоимением в придаточном определительном предложении и определяемым словом. Так, вышеуказанным примерам соответствуют отношения *ref(set-23, which-26)* и *ref(designers-11, who-12)*.

Также при выявлении концептов необходимо обратить внимание на синтаксическую связь *compound*. Этот тип связи описывает так называемые сложные слова, которые в английском языке, в отличие от русского или немецкого, записываются в форме нескольких отдельных слов. Например, двухосновное существительное «пресс-конференция» в русском языке пишется через дефис, в немецком – слитно (Pressekonferenz), а в английском – раздельно (press conference). Вследствие этого, в парсере каждая основа сложного слова анализируется как самостоятельная единица, но в результате они объединяются связью *compound*. *Compound* не всегда указывает на сложное существительное, на его месте также может быть числительное или фразовый глагол. Чтобы исключить ошибочное определение концепта, необходимо принимать во внимание все син-

таксические связи, в которые вовлечена рассматриваемая лексема. Например, в анализируемом тексте была обнаружена связь *compound(language-3, project-2)*. Кроме того, лексема *language* имеет связь *nominal subject* с глаголом *help* и, следовательно, является потенциальным концептом. Определив две данных связи, мы можем сделать вывод о том, что предполагаемым концептом является не отдельное слово, а словосочетание *project language*, имеющее более узкое значение.

Существует также тип связи, который позволяет выявлять действия, осуществляемые над объектом, и называется *direct object* (*dobj*) – *прямой объект*. Примеры связи данного типа: *dobj(detect-6, mistakes-10), dobj(generating-28, language-33), dobj(recording-21, language-23)*. В случае если то же самое действие зафиксировано в связи *nominal subject*, то оно выступает связующим звеном между субъектом и объектом и может образовать отношение между двумя концептами.

Другой тип синтаксической связи *multi-word expression* (*mwe*) выражает лексемы, состоящие из нескольких частей. Это могут быть составные предлоги и союзы, например, *mwe(In-1, order-2)* в высказывании *In order to improve the quality.* Данная синтаксическая связь позволяет идентифицировать тэги, состоящие из нескольких слов, что является необходимой функцией для автоматического анализа текста.

Ниже представлен список основных синтаксических связей в проанализированном тексте и их возможных семантических конструктов, которые могут быть их них извлечены:

Тэг	Расшифровка	Примеры из текста	Использование для установления семантической связи
amod	adjectival modifier	amod(activity-9, conceptual-8) amod(set-23, integrated-22) amod(mistakes-10, semantic-9) amod(understanding-19, necessary-16) amod(application-26, further-25) amod(representation-15, justified-14) amod(core-19, conceptual-18) amod(degree-7, higher-6) amod(support-19, multi-agent-18)	Фильтрация концепта, его свойства

Продолжение таблицы

com-pound	noun or other part of speech compound modifier	compound(language-3, project-2) compound(ontology-24, project-23)	Фильтрация концепта
dobj	direct object	dobj(improve-4, quality-6) dobj(develop-13, systems-14) dobj(create-20, set-23) dobj(generating-28, language-33) dobj(detect-6, mistakes-10) dobj(recording-21, language-23) dobj(using-17, support-19)	Действие над объектом
mwe	multi-word expression modifier	mwe(In-1, order-2)	Идентификация тегов, состоящих из нескольких слов
nsubj	nominal subject	nsubj(develop-13, who-12) nsubj(need-18, we-17) nsubj(provides-27, which-26) nsubj(aim-3, we-2) nsubj(develop-13, designers-11) nsubj(provides-27, set-23) nsubj(helps-4, language-3) nsubj:xsubj(detect-6, language-3) nsubj:xsubj(prevent-8, language-3)	Фильтрация концепта
ref	referent	ref(designers-11, who-12) ref(set-23, which-26)	Фильтрация концепта

Автоматизированное извлечение из текста пар семантически связанных фразовых единств

С учётом вышеизложенных идей были разработаны инструменты для автоматизированного извлечения из текстов на английском языке пар лексических единиц (фразовых единств), связанных отношениями типа «часть – целое» и «причина – следствие». Выбор данных типов отношений обусловлен спецификой задачи, в рамках которой создавались инструменты по анализу текста, – задачи онтологического сопровождения процесса проектирования автоматизированных систем. Понимание отношений типа «часть –

целое», присутствующих в концептуальном пространстве разрабатываемого проекта, позволяют в дальнейшем перейти к пониманию архитектуры будущей системы, а понимание отношений типа «причина – следствие» позволяет строить её логико-алгоритмические прототипы.

Помимо результатов парсинга, для выявления связей также использовались наборы лексических маркеров (тегов) различных типов, которые свидетельствуют о наличии связи того или иного типа в тексте. Такими тегами могут выступать предлоги, союзы, глаголы или глагольные конструкции. В качестве примера приведём наборы тегов, относящихся к связи типа «причина – следствие», которые были получены на основе анализа различных лингвистических исследований, посвящённых способам выражения каузального типа связи в английском языке:

1) Маркеры причины: *because, as, since, inasmuch as, as long as, now that* и т. д.

2) Маркеры следствия: *for this reason, so, therefore, consequently, as a result, accordingly, for this reason, hence, thus, nevertheless, that's why, the result is* и т. д.

Для выявления связей типа «часть – целое» в англоязычных текстах также применялись синтаксические паттерны, описанные в работе группы учёных из США [11].

Внешний вид разработанных средств представлен на рис. 4. В верхней части интерфейса расположен текст, который необходимо проанализировать, – в исходном варианте (левый блок) и с подсвеченными тегами (правый блок). В нижней части отображаются результаты анализа – пары связанных единиц (левый блок) и группы единиц с общим «целым» (правый блок).

Аналогичным образом выглядит инструмент по анализу связей типа «причина – следствие».

Формирование групп с общим «целым» (или общей «причиной») является важным этапом, который позволяет в дальнейшем встроить данные инструменты в систему онтологического сопровождения проекта.

Полученные пары и группы связанных фразовых единств, по решению проектировщика, могут быть занесены в онтологию проекта и впоследствии использоваться для построения схем архитектуры и прототипов системы, а также для анализа проектных документов и рассуждений проектировщика, зафиксированных на более поздних этапах проектирования, на согласованность и корректность.

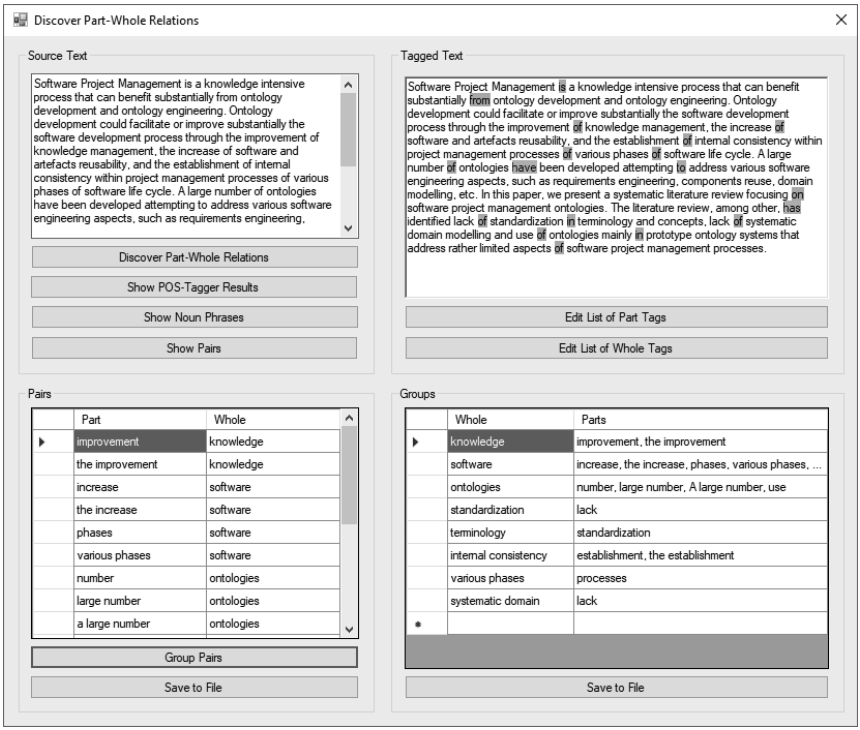


Рис. 4. Интерфейс инструмента по выявлению связей типа «часть – целое»

Разработанные инструменты имеют невысокую точность, но, тем не менее, могут существенно облегчить проектировщику задачу построения проектной онтологии.

Заключение

Таким образом, можно сделать вывод, что некоторые синтаксические связи слов в тексте могут заключать в себе определённую семантику.

На основе синтаксических связей предлагается автоматизированно выделять концепты, их свойства, а также отношения концептов друг с другом, заносить их в проектную онтологию, а также сравнивать семантическую структуру различных высказываний, чтобы проверить их на возможную противоречивость или рассогласован-

ность. Безусловно, это не единственный метод извлечения подобных семантических конструктов, однако он может быть эффективно использован в качестве дополнения к морфологическому анализу, анализу текста по ключевым словам, использованию специальных терминологических словарей и некоторых других.

ЛИТЕРАТУРА

1. Chaos report 2015. [Электронный ресурс]. – Режим доступа: https://www.standishgroup.com/sample_research_files/CHAOSReport2015-Final.pdf. – (Дата обращения: 18.10.2018).

2. Sosnin P., Pushkareva A., Negoda V. Ontological Support of Design Thinking in Developments of Software Intensive Systems. In: Abraham A., Kovalev S., Tarassov V., Snasel V., Vasileva M., Sukhanov A. (eds) Proceedings of the Second International Scientific Conference “Intelligent Information Technologies for Industry” (IITI’17). IITI 2017. Advances in Intelligent Systems and Computing, vol 679. Springer, Cham.

3. Sosnin P., Pushkareva A. Ontological Controlling the Lexical Items in Conceptual Solution of Project Tasks. In: Gervasi O. et al. (eds) Computational Science and Its Applications – ICCSA 2017. ICCSA 2017. Lecture Notes in Computer Science, vol 10409. Springer, Cham.

4. P. Sosnin, A. Kulikova. Ontology-Based Way of Formulating the Statements of Project Tasks in Designing a System with Software. In: Proceedings of the 18th International Conference on Computational Science and Applications (ICCSA 2018), pp. 25–30.

5. Danqi Chen and Christopher D Manning. 2014. A Fast and Accurate Dependency Parser using Neural Networks. Proceedings of EMNLP 2014.

6. Pullenti – SDK извлечения именованных сущностей из неструктурированных текстов (Puller of Entities). [Электронный ресурс]. – Режим доступа: http://www.pullenti.ru. – (Дата обращения: 18.10.2018).

7. Автоматическая обработка текста. [Электронный ресурс]. – Режим доступа: http://aot.ru/. – (Дата обращения: 18.10.2018).

8. Киселёв М. В. Федосеева Д. В. Синтаксический парсер русского языка LPaRus компании Megaputer Intelligence // Компьютерная лингвистика и интеллектуальные технологии: по материалам международной конференции «Диалог 2017». Москва, 31 мая – 3 июня 2017.

9. Боярский К.К., Каневский Е.А. Семантико-синтаксический парсер SemSin // Научно-технический вестник информационных технологий, механики и оптики. 2015. Т. 15. № 5. С. 869–876.

10. Томита-парсер. [Электронный ресурс]. – Режим доступа: https://tech.yandex.ru/tomita/. – (Дата обращения: 18.10.2018).

11. Roxana Girju, Adriana Badulescu, Dan Moldovan. Automatic Discovery of Part-Whole Relations [Электронный ресурс]. – Режим доступа: http://www.hlt.utdallas.edu/~adriana/Publications/PartWhole_CL2006.pdf – (Дата обращения: 26.11.2018).

СОДЕРЖАНИЕ

ТРУДЫ
МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ
ПО КОМПЬЮТЕРНОЙ И КОГНИТИВНОЙ
ЛИНГВИСТИКЕ

TEL-2018

Том 1

В авторской редакции

Подписано в печать: 18.12.2018 г.
Формат 60×84 1/16. Бумага офсетная.
Гарнитура «Таймс». Усл.-печ. л. 25,46.
Тираж 150 экз. Заказ 18В-365-1

Отпечатано с готового оригинал-макета ИП Ольшевский В.П.
420140, Казань, Минская, 30-150

ISBN 978-5-9690-0477-1