Russian Journal of Linguistics
ISSN 2687-0088 (print), ISSN 2686-8024 (online)

2022 Vol. 26 No. 2 275–316
http://journals.rudn.ru/linguistics

https://doi.org/10.22363/2687-0088-30161

Research article

Computational linguistics and discourse complexology: Paradigms and research methods

Valery SOLOVYEV[1] ✉, Marina SOLNYSHKINA[1] and Danielle MCNAMARA[2]

[1]Kazan (Volga Region) Federal University, Kazan, Russia
[2]Arizona State University, Tempe, USA
✉maki.solovyev@mail.ru

Abstract
The dramatic expansion of modern linguistic research and enhanced accuracy of linguistic analysis have become a reality due to the ability of artificial neural networks not only to learn and adapt, but also carry out automate linguistic analysis, select, modify and compare texts of various types and genres. The purpose of this article and the journal issue as a whole is to present modern areas of research in computational linguistics and linguistic complexology, as well as to define a solid rationale for the new interdisciplinary field, i.e. discourse complexology. The review of trends in computational linguistics focuses on the following aspects of research: applied problems and methods, computational linguistic resources, contribution of theoretical linguistics to computational linguistics, and the use of deep learning neural networks. The special issue also addresses the problem of objective and relative text complexity and its assessment. We focus on the two main approaches to linguistic complexity assessment: "parametric approach" and machine learning. The findings of the studies published in this special issue indicate a major contribution of computational linguistics to discourse complexology, including new algorithms developed to solve discourse complexology problems. The issue outlines the research areas of linguistic complexology and provides a framework to guide its further development including a design of a complexity matrix for texts of various types and genres, refining the list of complexity predictors, validating new complexity criteria, and expanding databases for natural language.

Keywords: *computational linguistics, linguistic complexology, discourse complexology, text complexity, machine learning, natural language processing*

For citation:
Solovyev, Valery, Marina Solnyshkina & Danielle McNamara. 2022. Computational linguistics and discourse complexology: Paradigms and research methods. *Russian Journal of Linguistics* 26 (2). 275–316. https://doi.org/10.22363/2687-0088-30161

Научная статья

Компьютерная лингвистика и дискурсивная комплексология: парадигмы и методы исследований

В.Д. СОЛОВЬЕВ[1] ✉, М.И. СОЛНЫШКИНА[1], Д.С. МАКНАМАРА[2]

[1]Казанский федеральный университет, Казань, Россия
[2]Университет штата Аризона, Темпе, США
✉maki.solovyev@mail.ru

Аннотация
Важнейшей особенностью современных исследований является значительное расширение научной проблематики и повышение точности расчетов лингвистического анализа за счет способности искусственных нейронных сетей к обучению и возможности не только автоматизировать лингвистический анализ, но и решать задачи отбора, модификации и сопоставления текстов различных типов и жанров. Цель данной статьи, как и выпуска в целом, – представить некоторые направления исследований в области компьютерной лингвистики и лингвистической комплексологии, а также обосновать целесообразность выделения новой междисциплинарной области – дискурсивной комплексологии. В обзоре трендов компьютерной лингвистики делается акцент на следующих аспектах исследований: прикладные задачи, методы, компьютерные лингвистические ресурсы, вклад теоретической лингвистики в компьютерную, применение нейронных сетей глубокого обучения. Особое внимание в спецвыпуске уделено вопросам оценки объективной и относительной сложности текста. Выделяются два основных подхода к решению проблем лингвистической комплексологии: «параметрический подход» и машинное обучение, прежде всего, нейронные сети глубокого обучения. Исследования, публикуемые в специальном выпуске, показали не только высокую значимость методов компьютерной лингвистики для развития дискурсивной комплексологии, но и расширение методологических находок компьютерной лингвистики, используемых для решения новых задач, стоящих перед комплексологами. Они высветили основные проблемы, стоящие перед отечественной лингвистической комплексологией, и наметили направления дальнейших исследований: создание матрицы сложности текстов различных типов и жанров, расширение списка предикторов сложности, валидация новых критериев сложности, расширение баз данных для естественного языка.

Ключевые слова: *компьютерная лингвистика, лингвистическая комплексология, дискурсивная комплексология, сложность текста, машинное обучение, обработка естественного языка*

Для цитирования:
Solovyev V., Solnyshkina M., McNamara D. Computational linguistics and discourse complexology: Paradigms and research methods. *Russian Journal of Linguistics.* 2022. T. 26. № 2. P. 275–316. https://doi.org/10.22363/2687-0088-30161

1. Introduction

The article addresses modern trends in computational linguistics, language and discourse complexity. It also provides a brief overview of the articles in the issue.

Computational linguistics (hereinafter CL), as the name implies, is an interdisciplinary science at the intersection of linguistics and computer sciences. It explores the problems of automatic processing of linguistic information. Another commonly used name for this discipline, that is synonymous with the term "computational linguistics", is Natural Language Processing (NLP). In a number of research works these concepts are separated, considering that CL is more of a theoretical discipline, and NLP is of a more applied nature. CL began to develop in the early 1950s, almost immediately after the advent of computers. Its first task was development of machine translation, and translation of journals from Russian into English in particular. The initial stage of CL development is comprehensively presented in J. Hutchins (1999). It surely was beyond the capacity of researchers to solve the problems of machine translation very quickly, and the initial optimism turned out to be groundless, although in recent years it has become possible to obtain translations of acceptable quality. However, within 70 years of development, CL has achieved significant success in solving many urgent practical problems, which made it one of the most dynamically developing and important research areas in both linguistics and computer science. In our opinion, the best monographs on CL are (Clark et al., Indurkhya & Damerau 2010). The latest review, including also an analysis of the prospects for its development, can be found in the article by Church and Liberman (2021).

In the review of computational linguistics trends, we focus on the following aspects of research: application-oriented tasks, methods, resources, contribution of theoretical linguistics to computer linguistics, and application of deep learning neural networks. The latter appeared about 10 years ago (Schmidhuber 2015) and revolutionized research of artificial intelligence, including many areas of CL. Artificial neural networks constitute a formal model of biological networks of neurons. Their most important feature is the ability to learn; in case of an error, the neural network is modified in a certain way. Although neural networks were proposed as early as 1943, a breakthrough in their use was made only a few years ago. It is associated with the three following factors: the emergence of new, more advanced 'self-learning', unsupervised training algorithms, improved performance of computers, and Internet database increase. Advances in NLP in the late 2018 were mainly related to BERT (Devlin et al. 2018), a neural network pre-trained on a corpus of texts. Currently, BERT and its enhanced models show better performance on many NLP problems (see Lauriola, Lavelli & Aiolli 2022).

2. Applied problems and methods of computer linguistics

2.1. Application-oriented tasks of Computational Linguistics

In addition to machine translation, the main application-oriented tasks of CL include document processing, computer analysis of social networks, speech analysis and synthesis (including voice assistants), question-answering systems, and recommender systems.

The largest task is document processing including a wide range of subtasks: search, summarization, classification, sentiment analysis, information extraction, etc.

Development of search engines, obviously, is the most well-known and widely used CL task, successfully implemented in Google and Yandex search engines. A detailed introduction to the issue of information retrieval can be found in (Manning et al. 2011). The main type of search queries is a set of keywords. The two main problems of search are as follows: the need to provide fast searches in the vast amount number of texts on the Internet and to ensure that any search takes into account not the query forms only but its semantics. The main idea of a quick search is to preprocess all documents on the Internet with the creation of a so-called search index that indicates location of the query in specific documents. A semantic document search, or a semantic search, is implemented in the well-known concept of Semantic Web (Domongue et al. 2011), based on the idea of ontologies (presented below). E.g., in response to a query "*Beethoven ta ta ta tam*" Google refers to the Wikipedia article about Beethoven's 5th symphony, although the text of the article does not contain the phrase "*ta ta ta tam*". Thus, the Google search engine "understands" that "*ta ta ta tam*" and the 5th symphony are semantically related. A successful search would be simply impossible without linguistic research, which led to the development of algorithms for morphological and syntactic analysis, thesauri and ontologies for the explication of semantic relationships between entities.

The term "information retrieval" is interpreted as a search for information of a certain type in the text, i.e. entities, their relationships, facts, etc. The best developed is the algorithm of extracting named entities (Name Entity Recognition, NER), i.e. persons, organizations, geographical objects, etc. A recent survey of IT professionals from various business areas[1] indicates that the NER task is the most demanded in business applications. Researchers apply various techniques to solve this problem: ready-made dictionaries of people's names and names of geographical objects; linguistic features (use of capital letters), defined patterns of noun phrases; and machine learning methods. An overview of this area can be found in Sharnagat (2014). NER systems based on dictionaries and rules correctly extract about 90% of entities in texts, while BERT-based systems already provide about 94% of correctly extracted entities (Wang 2020), which is comparable to the level of human accuracy and demonstrates benefits of deep learning neural networks.

[1] https://gradientflow.com/2021nlpsurvey/

The task of retrieval of events and facts is challenging. The classic approach here is to create event templates that capture types and roles of the entities participating in the events. For example, the event "June 24, 2021 Microsoft presented Windows 11" is described by the following template: Activity type – sales presentation, Company – Microsoft, Product – Windows 11, Date – June 24, 2021. Templates of this type are created manually, which is labour-intensive. Efficiency of information extraction systems depends on their quality. Typically, such systems extract no more than 60% of facts (Jiang et al. 2016).

In recent years, many studies addressed the problem of text sentiment analysis (cf. Cambria 2017), i.e. identification of the so-called "tone" of texts: whether a text carries a positive or negative attitude towards the text referents. This area is important for companies to evaluate user comments on their products and services. The problem is also being solved with the help of developing specific patterns, dictionaries, and machine learning methods. The Russian dictionary RuSentiLex, (Loukachevitch & Levchik 2016), registers over 12,000 lemmas marked as positive, negative or neutral. The main problem of sentiment analysis of texts is its context-dependency as a word can be positive in certain contexts and negative in others. A possible way of addressing the problem is compiling sentiment lexicon dictionaries for specific subject areas.

Another fundamental problem is not only to assess the tone of the entire text, but define the referential aspect of the sentiment. It is especially important in applied research on customer reviews of products and services (Solovyev & Ivanov 2014). The achieved accuracy in the area, which is about 85%, was effected through BERT technology (Hoang et al. 2019).

Another important task of document processing is text summarization and text skimming (Miranda-Jiménez, Gelbukh & Sidorov 2013). Its practical importance is determined by the gigantic and increasing size of texts on the Internet. There are two approaches to solving this problem: extractive and abstract. The extractive approach –implies assessing the importance score of sentences in the text and selecting a small number of the most significant ones. It requires non-trivial mathematical methods to evaluate informational hierarchy of text parts. The abstract, approach implies a generation of original sentences that summarize the content of the source text. In recent years the task of generating text abstracts was successfully fulfilled with neural networks. An important component of summarization systems are sentence parsing algorithms. A brief overview is provided in Allahyari (2017).

Computer analysis of social networks and social media is another application-oriented task. It can have multiple objectives with monitoring social attitudes, identifying manifestations of extremism and other illegal activities, and even analyzing the spread of epidemics. E.g. at the beginning of the coronavirus pandemic researchers suggested an analysis of social media content, including the spread of misinformation (cf. Cinelli, Quattrociocchi & Galeazzi 2020). Social network analysis implies defining the content of messages and connections between

users, which enables identifying groups of users with common interests. At the same time, heterogeneity of content presents a significant challenge. In recent years, neural networks have become the main tool for social network analysis (cf. Ghani et al. 2019). Batrinca & Treleaven (2015) provide an overview of the research in the area and addresses mostly humanitarians.

Speech analysis and synthesis stand apart in CL, as they require specific software and hardware tools to work with acoustic signals. Speech recognition systems are very diverse and are classified according to many parameters: vocabulary size; speaker type (age, gender); type of speech; purpose; structural types and their selection principles (phrases, words, phonemes, diphones, allophones, etc.). The input speech flow is compared with acoustic and language models, including various features: spectral-temporal, cepstral features, amplitude-frequency, features of nonlinear dynamics. Speech recognition is challenging because words are pronounced differently by different people in different situations. Nevertheless, at the moment there are many commercial speech recognition systems, in particular those built into Windows. One of the best known is "Watson speech to text" developed by IBM (Cruz Valdez 2021).

Speech recognition is the heart of voice assistants becoming increasingly popular worldwide. A voice assistant commonly known in Russia is Alice[2] designed and developed by Yandex. Alice is integrated into the Yandex services: by a voice command it searches for information. E.g. it can find a weather forecast on Yandex.Weather, traffic data in Yandex.Maps, etc. Alice can control smart home systems and even entertain: play riddles with children, tell fairy tales and jokes. Speech recognition in voice assistants is facilitated by their ability to tune in to the voice of a certain person. State of the art review in voice assistants can be found in Nasirian, Ahmadian & Lee (2017), and one of the latest reviews of speech recognition problems is presented in Nassif (2019).

Speech synthesis is being actively used in information and reference systems, in airport, railway and office announcements. They are predominantly used in situations with a limited range of synthesized phrases. The simplest way to synthesize speech is sequencing pre-recorded elements. The quality of the synthesize speech is evaluated based on its similarity with human speech. High-quality speech synthesis systems are still a dream of many researchers and users. The latest overview of speech synthesis is presented in Tan (2021).

We also address recommender systems which are probably familiar to all Internet users. Recommender systems predict which objects (movies, music, books, news, websites) might be interesting to a particular user. For this, they collect information about users, sometimes explicitly, asking them to rate objects of interest, and more often implicitly, collecting information about users' behavior on the Internet. The following idea turned out to be productive: people who similarly estimated some objects in the past are most likely to give similar estimates to other objects in the future (Xiaoyuan & Khoshgoftaar 2009). This particular idea allows

[2] https://dialogs.yandex.ru/store

researchers to effectively extrapolate user behavior. Developing recommender systems depends mostly on linguistic resource. For example, an effective recommender system is based on synonyms dictionaries. Such systems are supposed to "understand" that "children's films" and "films for children" mean the same. For synonymy in recommender systems, see Moon (2019), a general review is presented in Patel & Patel (2020).

Question-answering systems, or QA-systems, are designed to provide answers in natural language, i.e. they have a natural language interface. They search for answers in a textual database that QA systems have. Like search engines, QA systems provide a user with the ability to search for information. However, an important distinguishing feature of QA systems is that they allow a user to find information that might be implicit, e.g., a film that a user might like but it could not be found with a regular search engine. Obviously, the quality of a QA system depends on its database size, i.e. whether it contains an answer to a question at all, as well as on the technologies for processing questions and comparing them with the database information. As for processing a question, it begins with identifying the type of question and the expected response. For example, the question "Who..." suggests that the answer is to contain the name of a person. QA systems apply numerous complex CL methods and, similar to recommender systems, face the issue of synonymy (Sigdel 2020). The latest review of QA systems is published by Ojokoh and Adebisi (2018).

2.2. Methods of Computational Linguistics

All CL methods can be divided into two large classes: a class based on dictionaries and rules (templates) and a class based on machine learning. These two classes are fundamentally different in their approaches. Dictionaries and rules use accumulated knowledge about the language, as well as results of highly professional manual labor, and therefore they are extremely expensive. Machine learning is implemented on a large number of examples, presented in annotated corpora which function as training sets. The algorithm implies analyzing training sets, identifying the existing patterns and then offering solutions to the problems set. Modern machine learning systems vary in their functions and applications, although deep learning neural networks have proved to be the most efficient. At an input node of a neural network, any language data is fed in encoded forms as tokens: letters, bigrams, short high-frequency morphemes, and words.

Application of this approach depends on a large body of annotated texts at a researcher's disposal: the larger the training set, the better the neural network will learn. At the same time, annotation is quite simple and its implementation does not necessarily involve professional linguists as researchers can refer to services of native speakers.

In this article, we will focus on the basic methods of CL and refer readers to the above-mentioned monographs for a detailed review of the area (cf. Clark et al. 2013, Indurkhya & Damerau 2010).

Automatic text analysis usually begins with its pre-processing which includes text segmentation, i.e. segmentation into words and sentences. Though it may seem like a simple task, since words are separated from each other by spaces and sentences begin with a capital letter and end with a period (rarely, exclamation marks, question marks, ellipsis) followed by a space. The most typical example of the rule or pattern is the following: a period – space – capital letter. However, it is not that simple. A period can be in the middle of a sentence after the first initial, followed by a space and then a capitalized second initial. Here, the period does not explicitly indicate the division of the text into sentences. As an example, we can refer to the following sentence: "Lukashevich N.V., Levchik A.V. Creation of a lexicon of evaluative words of the Russian language RuCentilex // Proceedings of the OSTIS-2016 conference. pp. 377–382". Despite all the difficulties, the segmentation problem is considered to be practically solved. In 1989, Riley (1989) managed to achieve a 99.8% accuracy rate for splitting texts into sentences. To achieve this result, the researcher developed a complex system of rules taking into account the following features: length of the word before the dot, length of the word after the dot, presence of a word before the dot in the dictionary of abbreviations, etc.

The next step in the course of text analysis is morphological. Consider, as an example, a language with complex morphology – Russian. For the Russian language, morphological analysis is performed by a number of analyzers: MyStem, Natasha, pymorphy2, SpaCy, etc. In CL, morphological analysis, the purpose of which is to determine the morphological characteristics of a word, is based on a detailed description of inflectional paradigms. For the Russian language, a reference book of this kind is Zaliznyak (1977), which presents paradigm indices of almost 100,000 lemmas of the Russian language. The presence of such a directory made it possible to generate about 3 mln Word forms for the registered lemmas of the Russian language. Automatic text analysis finds a lemma corresponding to any word form and a complete list of morphological characteristics. The main challenge for the existing analyzers is homonymy, which the available parsers have not solved yet. And in situations when users require not all parsing options but one, analyzers produce the variant of morphological parsing of the highest frequency, still ignoring senses of the word in the context.

Another problem is parsing of the so-called "off-list" words, i.e. words not registered in the dictionary. Given that the average number of such words is about 3%, their morphological analysis requires developing special algorithms. The simplest solution foreseen is the following: based on the analysis of its flexion, the off-list word is assigned its morphological paradigm.

Syntactic parsing, or parsing, is much more complex. The result of syntactic parsing of a sentence is a dependency tree that presents a sentence structure either in the formalism of a generative grammar or in the formalism of a dependency grammar (cf. Tesnière 2015). Parsing requires a detailed description of the syntax of the language. The most successful analyzer for the Russian language is ETAP

developed by the Laboratory of Computational Linguistics of the Institute for Information Transmission Problems of the Russian Academy of Sciences as a result of over 40 years of research. Its latest version, ETAP-4, is available at (ENA, June 6, 2020)[3]. ETAP parser is based on the well-known model "Meaning ⇔ Text" (Melchuk 1974), its formalized version is described in the monograph by Apresyan (1989).

In the recent decade, parsing has also been performed by neural networks (cf. Chen & Manning 2014) trained on syntactically annotated corpora. English Penn Treebank (ENA, June 6, 2022)[4] is used for English. For the Russian language, one can use SynTagRus (ENA, June 6, 2022)[5], developed by the Laboratory of Computational Linguistics at the Institute for Information Transmission Problems RAS.

The task of semantic analysis is even more difficult. However, if we want the computer to "understand" the meaning, it is necessary to formalize semantics of words and sentences. The problem is solved in two classical ways. The first was initiated by C. Fillmore (1968), who introduced concepts of semantic cases or roles of noun phrases in a sentence. The correct establishing of semantic roles is an important step towards sentence comprehension. Fillmore's original ideas were realized in FrameNet lexical database (ENA, June 6, 2022)[6].

The second approach was implemented in an electronic thesaurus, or lexical ontology, WordNet (Fellbaum 1998) which was originally designed for the English language. Subsequently its analogues were developed for many languages. There are numerous analogues of WordNet for the Russian language, the most effective and being widely used is RuWordNet thesaurus (ENA, June 6, 2022)[7], (cf. Loukachevitch & Lashevich 2016), comprising over 130,000 words. WordNet-like thesauri explicate semantic relationships between words (concepts) including synonymy, hyponymy, hypernymy, etc., and their systemic parameters partially define their semantics. WordNet has been successfully implemented in a large number of both linguistic and computer research.

The idea of vector representation of semantics, i.e. word embeddings, has been proposed recently. Its core is constituted by the distributive hypothesis: linguistic units occurring in similar contexts have similar meanings (Sahlgren 2008). This hypothesis has been confirmed in numerous studies aimed at defining frequency vectors of words registered in large text corpora. There are multiple refinements and computer implementations of the idea, the most popular of which is word2vec (Mikolov et al. 2013) available in Gensim library (ENA, June 6, 2022)[8]. RusVectores system (Kutuzov & Kuzmenko 2017), available at (ENA, June 6,

[3] http://proling.iitp.ru/ru/etap4
[4] https://catalog.ldc.upenn.edu/LDC99T42
[5] https://universaldependencies.org/treebanks/ru_syntagrus/index.html
[6] https://framenet.icsi.berkeley.edu/fndrupal/
[7] https://ruwordnet.ru/ru
[8] https://github.com/rare-technologies/gensim

2022)[9] identifies vector semantics for Russian words. Specifically, RusVectores evaluates semantic similarity of words.

Obviously, the most important tool for research in CL, as indeed in all modern linguistics, are text corpora. The first corpus compiled in the 1960s was *Brown Corpus* which when released contained one million words. Since then, corpora size requirements have increased dramatically. For the Russian language, the most well known is the National Corpus of the Russian Language (NCRL, ENA, June 6, 2022[10]). Created in 2004, it is being constantly updated and currently includes over 600 mln words. In 2009, Google compiled and uploaded a very interesting multilingual resource, i.e. Google Books Ngram (ENA, June 6, 2022)[11], containing 500 bln words, 67 bln words of which constitute the Russian sub-corpus (cf. Michel 2011).

Another important problem is corpus annotation or tagging, which in difficult cases is done manually. The work is usually carried out by several annotators and their performance consistency is closely monitored (Pons & Aliaga 2021). Despite the fact that corpora have become an integral part of linguistic research, there have been ongoing disputes on their representativeness, balance, differential completeness, subject and genre relatedness, as well as data correctness (cf. Solovyev, Bochkarev & Akhtyamova 2020).

Thus, thanks to CL, researchers fully implement numerous services including information retrieval, automatic error correction, etc. This became possible due to fundamentally important accomplishments not only in computer science, but also in linguistics. CL uses extensive dictionaries and thesauri, detailed syntax models, and giant corpora of texts. Automatic morphological analysis in its modern form would not exist without A. A. Zaliznyak's "Dictionary of the Russian Language Grammar" (1977). Multiple studies in CL are based on manually created WordNet and RuWordNet thesauri. Computer technologies, in turn, contribute to the development of linguistics. Text corpora and statistical methods have already become commonplace; without them serious linguistic research would be impossible.

All key CL technologies are publicly available, e.g. (ENA, June 6, 2022)[12] houses programs to solve numerous basic tasks for numerous languages.

It is not really feasible to cover all the topics of CL, a vast and rapidly developing field of linguistics, in one article. Many important questions have been left beyond. We refer readers interested in the topics of co-reference resolution, disambiguation, topic modeling, etc. to the above-mentioned publications.

[9] https://rusvectores.org/ru/
[10] https://ruscorpora.ru/new/
[11] https://books.google.com/ngrams
[12] https://stanfordnlp.github.io/CoreNLP/

3. Complexity of language and text as a research problem

The core of the special issue is made up of the articles focused on text complexity assessment. At first glance, estimating language complexity based on the number of categories in its system seems to be very logical, and the task itself appears feasible. A good example of the idea can be a phonological inventory of the language, the number of morphophonological rules or verb forms. Obviously, in this case, it becomes possible to compare complexity of different languages and assign them to some objective, absolute complexity (Miestamo, Sinnemäki & Karlsson 2008). Notably, it is the "objective" complexity that is significant when mastering a non-native language. On the other hand, if a language is acquired as a native language, it does not present any difficulty for children, and from this point of view, all languages complexity is absolutely the same. Researchers admit that language and text complexity "resists measurement", and scholars working in this field face conceptual and methodological difficulties.

Significant in the light of the problems under study is the description of the relationship and interdependence of two areas of complexity studies: language, or 'lingue' complexity, i.e. linguistic complexology, on the one hand, and text or discourse, 'parole' complexity, i.e. discursive complexology, on the other.

The interpretation of the very concept of "language (lingue) complexity" changed dramatically in the 19th-20th centuries. In the 19th century, the Humboldtian theory on interdependence between the structure of a language and stage of development of people speaking this language was universally accepted (Humboldt 1999: 37). Acknowledging this concept, researchers actually acknowledge unequal status of languages and peoples. In the XXth century, the Humboldian views asserting inequality of languages and their speakers were replaced by the concept of the so-called single complexity, identical and equal for all languages of the world. The idea received two names: ALEC — "All Languages are Equally Complex" (Deutscher 2009: 243) and linguistic equi-complexity dogma (Kusters 2003: 5). Researchers who support the idea are to prove two hypotheses: (1) language complexity is constituted of sub-complexities of its elements; (2) all sub-complexities in linguistic subsystems are compensated: simplicity in area A is compensated by complexity in area B, and vice versa ("compensatory hypothesis"). Arguing the concept "All languages are equally complex", Ch. Hockett quite boldly stated: "Objective measurement is difficult, but impressionistically it would seem that the total grammatical complexity of any language, counting both the morphology and syntax, is about the same as any other. This is not surprising, since all languages have about equally complex jobs to do: and what is not done morphologically has to be done syntactically" (Hockett 1958: 180–181). Unfortunately, in the works of that period and approach, scholars discussed neither complexity criteria nor its empirical evidence. For a detailed overview of the "linguistic equi-complexity dogma", see the seminal work by J. Sampson, D. Gil, and P. Trudgill, Language Complexity as an Evolving Variable (Sampson et al. 2009).

The twenty-first century opened with a number of critical reviews of ALEC theory, on the one hand (cf. Miestamo et al. 2008), and McWhorter's provocative statement that "Creole grammars are the simplest grammars in the world" (McWhorter 2001). The very idea that all languages are equally complex has been convincingly rejected by sociolinguists, who have shown that language contact can lead to language simplification. This is shown in Afrikaans, Pidgins and Koine. Simplifying a language is possible, hence, before its simplification, the language was more complicated than after. And if a language can be more or less complex at different periods of its history, then some languages can be more complex than others (Trudgill 2012).

In the early 2000s the idea of linguistic complexity and the "dogma of equal complexity" was actively discussed at conferences and seminars (see the seminar "Language complexity as an evolving variable" organized by Max Planck Institute for Evolutionary Anthropology in 2007 in Leipzig ENA, June 6, 2022[13]), in a number of journal articles (cf. Shosted 2006, Trudgill 2004) and monographs (Dahl 2009, Kusters 2003, Miestamo et al. 2008, Sampson et al. 2009).

Publications on language complexity in Russia are predominantly reviews written by foreign scholars, although in recent years interest in the area has visibly grown. The most comprehensive are the studies conducted by A. Berdichevsky (2012) and the review of Peter Trundgill's book "Sociolinguistic Typology", 2011 by Vakhtin (2014). The problems of language complexity were also discussed at the Institute for Linguistic Research of the Russian Academy of Sciences (ILI RAS) in 2018 at the conference "Balkan Languages and Dialects: Corpus and Quantitative Studies".

Local and global complexity

The development of linguistic complexology led to the identification of two types of complexity: global, i.e. the complexity of the language (or dialect) as a whole, and local complexity, i.e. complexity of a particular level of language or domain (Miestamo 2008). And if the assessment of global complexity of a language, according to researchers, is a very ambitious and probably hopeless task which H. Deutscher compares with "chasing wild geese" (Deutscher 2009), then the measurement of local complexity is considered as a feasible task, which imples the compiling of a list and evaluating complexity predictors at various language levels. The list of predictors of *phonological complexity* traditionally includes phoneme inventory, frequency of marked [14] phonemes, tonal differences, suprasegmental patterns, phonotactic restrictions, and maximal consonant clusters (Nichols 2009, Shosted 2006). When evaluating *morphological complexity*, classical "inconvenience factors" (Braunmüller's term 1990: 627) are the size of inflectional morphology of a language (or language variety), specificity of

[13] https://www.eva.mpg.de/fileadmin/content_files/linguistics/pdf/ComplexityWS_Webpage_2007.pdf

[14] Phonemes that are rarely found in the languages of the world are considered marked (Berdichevsky 2012).

allomorph and morphophonemic processes, etc. (Dammel & Kürschner 2008, Kusters 2003). *Syntactic complexity* assessment is based on the accumulated data of syntax rules and follows the principle "the more, the more difficult", as well as language ability to generate recursions and clauses within a syntactic whole (Ortega 2003, Givón 2009, Karlsson 2009). *Semantic and lexical complexity* is estimated based on the number of ambiguous language units, the difference between inclusive and exclusive pronouns, lexical diversity, etc. (Fenk-Oczlon & Fenk 2008, Nichols 2009). *The pragmatic* or "hidden" complexity built on the law of economy is the complexity of inferences necessary to comprehend texts. Latent complexity languages allow for minimalist, very simple surface structures in which grammatical categories inferences are far from being trivial. The idea is exemplified by languages of Southeast Asia, which have achieved a particularly high degree of latent complexity. The latter is observed in the omission of pronouns and consequent multiple co-references in relative clauses, absence of relational markers, "bare" nouns lacking determiners and as such enabling a wide range of interpretations (Bisang 2009).

Research has indicated that high levels of local complexity at one level in a language do not necessarily entail low local complexity at another level, as predicted by the "dogma of equal complexity". For example, the analysis of metrics of morphological and phonological complexity in 34 languages carried out by R. Shosted did not reveal any expected statistically significant correlation (Shosted 2006). And the individual "balancing effects" (trade-offs) between local complexities observed by G. Fenk-Ozlog and A. Fenk, unfortunately, are also insufficient to validate the "dogma of equal complexity" of languages. G. Fenk-Ozlog and A. Fenk, in particular, found that in English the tendency towards phonological complexity and monosyllabicity is associated with a tendency towards homonymy and polysemy, towards a fixed word order and idiomatic speech (Fenk-Oczlon & Fenk 2008: 63). D. Gil has convincingly argued that isolating languages do not necessarily compensate for simple morphology with more complex syntax (Gil 2008).

Factors (or predictors) of language complexity are usually divided into internal and external ones. The number of elements and categories in the language, redundancy and irregularity of language categories are viewed as *the internal factors* of complexity. The modern paradigm developed the so-called "list approach" to assess internal complexity. The latter implies compiling a list of linguistic phenomena, the presence of which in a language increases its complexity. In fact, the lists of intrinsic complexity predictors are lists of the local complexity described above. For example, the complexity predictors list compiled by J. Nichols contains over 18 parameters and includes phonological, morphological, syntactic and lexical features (Nichols 2009). A language is considered more complex if it has more marked phonemes, tones, syntactic rules, grammatically expressed semantic and / or pragmatic differences, morphophonemic rules, more cases of addition, allomorph, agreement, etc. Scholars working in the area are interested, for

example, in the number of grammatical categories in the language (Shosted 2006), the number of phonemic oppositions (McWhorter 2008), the length of the "minimal description" of the language system (Dahl 2009). McWhorter (2001) compares word order, i.e. the position of the verb in the Germanic languages, proving that English syntax has a lower degree of complexity than Swedish and German. The reason for the claim is the loss of the V2 (verb-second) rule in English, according to which the personal verb in Swedish and German takes the second place in the sentence.

Language elements and functions with "duplicate" information or overspecification are viewed as "redundant" internal predictors of complexity, and therefore optional elements in a discourse (McWhorter 2008). P. Trudgill calls such elements "historical baggage" (Trudgill 1999: 149), V. M. Zhirmunsky – "hypercharacterization" (Zhirmunsky 1976), McWhorter – "ornamental elaboration", or "baroque accretion[s]" (McWhorter 2001). Syntagmatic redundancy is exemplified in indirect nomination and "semantic agreement". Language paradigmatic redundancy is manifested in synthetic grammatical categories, such as agreement and obviative markers (see McWhorter 2001).

The irregularity or "opacity" of form and word-formation processes as an internal factor in language complexity (see Mühlhäusler 1974) manifests itself in irregular affixes (prefixes pa- in 'pasynok' (stepson), su- in 'symrak' (twilight), niz- in 'nizvodit' (reduce), suffixes -tash in 'patrontash' (bandolier), -ichok in 'novichok' (novice), -arnik in 'kustarnik' (bush)) (see Kazak 2012).

External factors that determine language complexity are culture, language age and language contacts. Older languages serving well-developed multi-level cultures are considered to be more complex because they accumulated "mature language features" (cf. Dahl 2009, Deutscher 2010, Parkvall 2008). At the same time, intensive contacts between linguistic communities have a significant impact on the complexity of languages. At the beginning of this century, P. Trudgill stated that "small, isolated, low-contact communities with tight social networks" develop more complex languages than high-contact communities (Trudgill 2004: 306). However, in his later work, the researcher clarifies that the dynamics of interacting languages complexity is determined by the duration of contacts and the age of speakers mastering the superstratum: language simplification occurs during short-term contacts of communities, when adults learn a foreign (second) language. Language complication can take place in cases where the contacts are long-term and the second language is mastered not by adults, but children (Trudgill 2011). To prove the influence of language contacts on language complexity, B. Kortman and B. Smrechani (2004) compare the ways of implementing 76 morphosyntactic parameters, including the number of pronouns, noun phrases patterns, tense and aspect, modal verbs, verb morphology, adverbs, ways of expressing negations, agreement, word order, etc. in 46 variants of the English language. Researchers divide all variants of the English language into three large groups: (1) native to their speakers and performing all functions in the language community; (2) languages

that function as the second official language of the state, and (3) creole languages based on English. The study confirmed that the third group of languages, i.e. English-based creoles, are the least complex, native English (first) language varieties are the most complex, and second-language English varieties exhibit intermediate complexity (Kortmann & Szmrecsanyi 2004).

In the most general terms, *analytical methods* for assessing complexity are divided into *absolute* (theoretical-oriented and treated as "objective") and *relative* (user-oriented and thus "subjective[15]") (Crossley et al. 2008.). The absolute approach is popular in linguistic typology and is used to assess language complexity, while sociolinguistics and psycholinguistics use a relative approach. P. Trudgill defines relative difficulty as the difficulty which adults experience while learning a foreign language (Trudgill 2011: 371).

Text complexity as a construct is also modeled in discourse studies, linguistic personology, psycholinguistics and neurolinguistics. The area of these studies also includes relative complexity (or difficulty) of a text for different categories of recipients in different communicative environments, as well as absolute and relative (comparative) complexity of texts generated by different authors (see McNamara et al. 1996, Solnyshkina 2015).

4. Summary of articles in the issue

The current issue contains a detailed review and discussion of the best practices of text and discourse complexity assessment, as well as methods ranging from purely linguistic to complex interdisciplinary including multiple hard- and software tools.

One of the methods, i.e. eye-tracking, is viewed in the area as an objective way of assessing text complexity for different categories of readers. Research implementing eye-tracking techniques to evaluate Russian texts complexity remains spares. The basic task here is to select text parameters and oculomotor activity, as well as to identify methods of measuring text complexity perception. The features typically selected to measure text complexity are average word length and word frequency; as for parameters of oculomotor activity, it is preferably assessed with relative speed of reading a word, duration of fixations, and the number of fixations. Text readability is estimated in the number of words read per minute. Eye-tracking is the focus of articles contributed by Laposhina and co-authors and Bonch-Osmolovskaya and co-authors. Laposhina and co-authors show that the number of fixations on a word correlates with its length, while the duration of fixations correlates with its frequency. The research of Bonch-Osmolovskaya and co-authors is aimed at elementary discursive units (EDU) defined as "the quantum of oral discourse, a minimum element of discourse

[15] Attributing this type of complexity as subjective is not universally accepted, since it is quite objective for all participants in communication. It would be more appropriate to define this type of complexity as "individual".

dynamics" (cf. Podlesskaya, Kibrik 2009: 309). Eye-tracking techniques allow to indicate that the structure of EDE affects text readability.

Methods of neural networks are implemented to assess texts complexity in the articles by Cortalescu et al., Sharoff, Morozov et al., and Ivanov et al. They also share the object of research, i.e. texts for studying Russian as a foreign language. An accurate assessment of their complexity enables better text selection for various educational environments. E.g. implementation of BERT model mentioned above provides a high degree of accuracy of text complexity assessment, i.e. 91–92%.

While using neural networks, researchers face an important research problem, i.e. which text features affect neural network results. A possible approach here is to use neural network to measure correlation coefficients of numerous text features with text complexity. An extensive study of collections of texts of various genres in English and Russian, taking into account dozens of linguistic features, has made it possible to identify a number of non-obvious effects. For example, research shows that more prepositions are used in more complex texts in Russian but in simpler texts in English. Obviously, this is due to the difference in the typological structures of languages. Notably, however, genre has a much larger effect on text complexity across all languages as compared to differences between languages.

A broad review of multiple methods applied in the area is provided in the work of M.I. Solnyshkina and co-authors. The paper covers six historic paradigms of discourse complexology: formative, classical, closed tests, structural-cognitive period, the period of natural language processing, and the period of artificial intelligence.

An important distinguishing feature of the articles in this special issue and its contribution to discourse complexology is constituted by its diverse and extensive data: several hundred linguistic features, different languages, different text corpora, different genres. Text complexity is assessed on several levels: lexical, morphological, syntactic, and discourse. Multifaceted studies prove to explicate the nature of text complexity. The publications in the current issue also provide information on the corpora and dictionaries being compiled.

One of the most important parameters of text complexity is abstractness. The more abstract words a text contains, the more difficult it is. The latter makes it relevant to compile dictionaries of abstract/concrete words and means of estimating text abstractness. English dictionaries of abstract/concrete words were published at the turn of the century, and the Russian language was lately viewed as "under resourced" since no dictionary identifying the degree of words abstractness was available. Solovyov and co-authors present a detailed methodology of composing a dictionary of abstractness for the Russian language. The article also describes the areas of dictionary application.

Linguistic complexity is an interdisciplinary problem, an object of computational linguistics, philosophy, applied linguistics, psychology, neurolinguistics, etc. In the 21st century, complexity studies acquired concepts and terminology, developed and verified a wide range of linguistic parameters of

complexity. The main achievement of the new paradigm was the validation of cognitive predictors of complexity enabling the assessment of discourse complexity. This success, as well as an interdisciplinary approach to the problem, made it possible to integrate studies of discourse complexity into a separate area, i.e. discourse complexology. Complexity issues are not an "end in itself", since the research results are relevant both for linguistic analysis and for predicting comprehension in a wide range of pragmalinguistic situations.

One of these situations is cognitive analysis of mistakes made in a foreign language learning which is the object of research conducted by Lyashevskaya and Yanda and colleagues. Both studies focus on the interrelationship between text complexity of texts and cognitive resources necessary to comprehend a text. Lyashevskaya et al. established that the number of mistakes made by a student is correlated with morphological complexity of his/her discourse. Yanda et al. present a computer system designed to analyze and adequately explain mistakes of a learner of Russian as a foreign language.

5. Conclusion

The recent successes of computational linguistics have largely ensured accomplishments in discourse complexology and allowed scientists not only to automate a number of linguistic analysis operations, but also create user-friendly text profilers. Tools such as ReaderBench, Coh-Metrix, and RuMOR (cf. the current issue) are capable of solving both research and practical tasks: selecting texts for target audiences, editing and shortening texts, analyzing cognitive causes of errors, and even suggesting verbal strategies. The algorithms of automatic text profilers are based on classical and machine learning methods, including deep learning neural networks, one of the latest systems of which is BERT. At present, and this is well shown in a number of articles of the special issue, researchers are successfully combining methods of machine learning and the so-called "parametric approach".

However, the most important feature of modern research is a vast expansion of research problems and accuracy increase resulting from the abilities of artificial neural networks to learn and modify. Artificial intelligence breakthroughs are attributable to the three main factors: new advanced self-learning algorithms, high computer speeds, and a significant increase in training data. Modern databases, as well as dictionaries and tools for the Russian language developed in recent years, allowed the authors of the special issue to address and successfully solve a number of problems of text complexity.

A solid foundation for success in discourse complexity were findings of cognitive scientists at the beginning of our century which completely changed complexology paradigm. If the main achievement of the XXth century complexology was the idea that "different types of texts are complex in different ways", the discourse complexology of the XXIst century proposed and verified complexity predictors for various types of texts and developed toolkits for assessing relative complexity of texts in various communicative situations. With cognitive

methods in its arsenal, complexology acquired two additional variables: linguistic personality of the reader and reading environment.

The new research paradigm of linguistic complexology is manifested in those articles of the special issue which are aimed at defining new criteria for text complexity: expert evaluation, comprehension tests and reading speed tests have been replaced by new methods, which allow scholars to identify discourse units affecting text comprehension.

The studies published in the special issue also highlighted the main problems facing Russian linguistic complexology: creating a complexity matrix for texts of various types and genres, expanding the list of complexity predictors, validating new complexity criteria, and expanding databases for the Russian language.

RU

1. Введение

Статья посвящена современным трендам компьютерной лингвистики и проблематике сложности языка и дискурса. В ней также дается краткий обзор статей выпуска.

Компьютерная лингвистика (далее КЛ) является междисциплинарной наукой на стыке лингвистики и компьютерных наук. Она исследует проблемы автоматической обработки информации в языковой форме. Другое часто используемое название этой дисциплины, фактически синонимичное термину «компьютерная лингвистика», – обработка естественного языка (Natural Language Processing, NLP). Иногда эти понятия разграничивают, считая, что КЛ – в большей степени теоретическая дисциплина, а NLP – более прикладная. КЛ начала развиваться в начале 1950-х гг., почти сразу после появления компьютеров. Первой ее задачей была разработка машинного перевода, в частности перевода научных журналов с русского языка на английский. О начальном этапе развития КЛ можно прочитать в работе (Hutchins 1999). Безусловно, первоначальный оптимизм по поводу быстрого решения проблемы машинного перевода оказался необоснованным, и лишь в последние годы удалось получить переводы приемлемого качества. Однако в КЛ за 70 лет развития достигнуты серьезные успехи в решении многих актуальных практических задач, что сделало ее одним из самых динамично развивающихся и важных разделов как лингвистики, так и компьютерных наук. На наш взгляд, лучшими монографиями по КЛ являются (Clark et al. 2013, Indurkhya & Damerau 2010). Последний обзор, включающий также анализ перспектив ее развития, можно найти в статье (Church & Liberman 2021).

Появившееся примерно 10 лет назад глубокое обучение нейронных сетей (Schmidhuber 2015) обеспечило настоящую революцию в области искусственного интеллекта и в том числе во многих разделах КЛ. Искусственные нейронные сети представляют собой формальную модель биологических се-

тей нейронов. Важнейшей их особенностью является способность к обучению, в случае ошибки нейронная сеть определенным образом модифицируется. Хотя нейронные сети были предложены еще в 1943 г., лишь несколько лет назад был совершен прорыв в их использовании. Он связан с тремя факторами: появлением новых, более совершенных алгоритмов самообучения, повышением быстродействия компьютеров, увеличением накопленного в интернете объема данных для обучения. В области NLP к прорыву привело появление в конце 2018 г. модели BERT (Devlin et al. 2018) – нейронной сети, предобученной на корпусе текстов. В настоящее время BERT и ее усовершенствованные варианты показывают лучшие результаты в решении многих задач NLP (новейший обзор см. (Lauriola et al. 2022)).

В обзоре трендов компьютерной лингвистики делается акцент на следующих аспектах исследований: прикладные задачи, методы, компьютерные лингвистические ресурсы, вклад теоретической лингвистики в компьютерную, применение нейронных сетей глубокого обучения.

2. Прикладные задачи и методы компьютерной лингвистики

2.1. Прикладные задачи компьютерной лингвистики

Кроме машинного перевода можно выделить следующие основные классы прикладных задач, лежащих в русле КЛ: обработка документов, компьютерный анализ социальных сетей, анализ и синтез речи (в том числе голосовые помощники), вопросно-ответные системы, рекомендательные системы. Наиболее объемной является задача обработки документов, включающая в себя большой спектр подзадач: поиск, суммаризация, классификация, анализ тональности, извлечение информации и т.д.

Поиск, очевидно, следует рассматривать как наиболее известную задачу КЛ, успешно реализованную в поисковиках Google, «Яндекс» и повсеместно используемую. Обстоятельное введение в проблематику информационного поиска можно найти в (Маннинг и др. 2011). Основной вид поисковых запросов – набор ключевых слов. Двумя главными проблемами поиска являются: необходимость обеспечить быстрый поиск в гигантском количестве текстов в интернете и обеспечить поиск с учетом семантики запроса, а не просто совпадения слов в запросе и документе. Быстрый поиск предполагает предобработку всех документов в интернете и создание так называемого поискового индекса, указывающего, в каких конкретно документах находится искомое слово. Поиск документов по семантике, или семантический поиск, реализован в рамках хорошо известной концепции Семантической паутины, или Semantic Web (Domingue et al. 2011), в основе которой лежит идея онтологий, о которых речь пойдет ниже. Пример семантического поиска: Google в ответ на запрос *Бетховен та та та там* первой выдает ссылку на статью в «Википедии» о 5-й симфонии Бетховена, хотя в тексте статьи не содержится фраза *та та та там*. Таким образом, поисковик Google «понимает», что *та та та там* и

5-я симфония семантически связаны. Успешный поиск был бы просто невозможен без лингвистических исследований, которые привели к созданию алгоритмов морфологического и синтаксического анализа, тезаурусов и онтологий для экспликации семантических связей между сущностями.

Термин «извлечение информации» трактуется как поиск в тексте информации определенного вида: сущностей, их отношений, фактов и т.д. Наиболее проработанной является задача извлечения именованных сущностей (Name Entity Recognition, NER), т.е. имена персон, организаций, географических объектов и т.д. Недавний опрос IT-профессионалов из различных сфер бизнеса (ENA, June 6, 2022)[16] показал, что задача NER является наиболее востребованной в бизнес-приложениях. Для решения этой задачи применяются различные техники: использования готовых словарей имен людей, названий географических объектов; лингвистических признаков (использование заглавных букв), подготовленных паттернов именных групп; методов машинного обучения. Обзор этой области можно найти в (Sharnagat 2014). Системы NER, основанные на словарях и правилах, правильно извлекают около 90% сущностей в текстах. BERT-основанные системы обеспечивают уже около 94% правильно извлекаемых сущностей (Wang 2020), что сопоставимо с уровнем точности человека и демонстрирует преимущества нейронных сетей с глубоким обучением. Значительно сложнее задача извлечения событий и фактов. Классический подход здесь состоит в создании шаблонов событий, в которых фиксируются типы и роли сущностей, участвующих в событиях. Например, событие «24 июня 2021 г. Майкрософт презентовала Windows 11» описывается следующим шаблоном: Тип активности – коммерческая презентация, Компания – Майкрософт, Продукт – Windows 11, Дата – 24 июня 2021 г. Шаблоны такого вида создаются вручную, что является весьма трудоемким делом. От их качества зависит эффективность системы извлечения информации. Обычно такие системы извлекают лишь около 60% фактов (Jiang et al. 2016).

В последние годы много работ посвящено сентимент-анализу текстов (Cambria 2017). Под этим понимается определение тональности текстов: выражено ли в тексте позитивное или негативное отношение к описываемым объектам. Эта область важна компаниям для оценки комментариев пользователей об их товарах и услугах. Для решения этой задачи также используются паттерны, словари, методы машинного обучения. Для русского языка создан словарь RuSentiLex (Loukachevitch & Levchik 2016), включающий более 12 тыс. слов и словосочетаний, маркированных как позитивные, негативные или нейтральные. Главная проблема сентимент-анализа текстов – это зависимость тональности слова от контекста. Слово в одних контекстах может иметь позитивную окраску, а в других – негативную. Возможным решением данной проблемы можно рассматривать построение словарей сентимент-лексикона для специфических предметных областей. Еще одна фундаментальная проблема – не просто оценить тональность всего текста в целом, а установить, к

[16] https://gradientflow.com/2021nlpsurvey/

какому аспекту ситуации относится оценочное высказывание. Это особенно важно в прикладных исследованиях отзывов пользователей о товарах и услугах (Solovyev & Ivanov 2014). Лучший в настоящее время результат – около 85% по стандартным метрикам точности и полноты – достигнут с применением технологии BERT (Hoang et al. 2019).

Еще одной важнейшей задачей обработки документов является суммаризация или саммаризация текстов (Miranda-Jiménez et al. 2013) – автоматическое построение краткого изложения (абстракта) содержания текста (или текстов). Ее практическая важность определяется гигантским и все возрастающим объемом текстов в интернете. Существует два подхода к решению этой задачи: экстрактивный и абстрактивный. Первый подход – экстрактивный – состоит в оценке информационной значимости предложений в тексте и выделении небольшого числа наиболее значимых. Он требует нетривиальных математических методов оценки информационной значимости фрагментов текста. Второй – абстрактивный – состоит в генерации оригинальных предложений, суммирующих все содержание исходного текста. Для генерации абстрактов, т.е. аннотаций текстов, в последние годы успешно применяются нейронные сети. В качестве одного из наиболее важных компонентов системы суммаризации включают алгоритмы синтаксического анализа предложений. Краткий обзор представлен в (Allahyari 2017).

Следующей задачей, которую мы здесь рассмотрим, является компьютерный анализ социальных сетей (social network, social media). Анализ контента социальных сетей преследует много различных целей. Это и мониторинг настроений в обществе, и выявление проявлений экстремизма и иной противозаконной деятельности, и даже анализ распространения эпидемий. Анализ контента социальных сетей, связанного с пандемией ковида, в том числе с распространением дезинформации, появился уже в начале эпидемии (Cinelli et al. 2020). В ходе анализа социальных сетей определяются как собственно содержание сообщений, так и связи между пользователями, что позволяет выявлять группы пользователей с общими интересами. При этом существенную трудность представляет разнородность контента. В последние годы основным инструментом анализа социальных сетей стали нейронные сети (Ghani et al. 2019). В работе (Batrinca & Treleaven 2015) представлен обзор данной области исследований, специально ориентированный на гуманитариев.

Несколько особняком в КЛ стоят анализ и синтез речи, требующие специфических программно-аппаратных средств работы с акустическими сигналами. Системы распознавания речи очень разнообразны и классифицируются по многим параметрам: размеру словаря; типу (возрасту, полу) диктора; типу речи; назначению; типу структурной единицы и принципам ее выделения (фразы, слова, фонемы, дифоны, аллофоны и др.). Входной речевой поток сопоставляется с акустическими и языковыми моделями, включаю-

щими разнообразные признаки: спектрально-временные, кепстральные, амплитудно-частотные, признаки нелинейной динамики. Распознавание речи признается сложной задачей, поскольку слова произносятся разными людьми и в разных ситуациях по-разному. Тем не менее на настоящий момент существует множество коммерческих систем распознавания речи, в частности встроенных в Windows. Хорошо известна система Watson speech to text, разработанная IBM (Cruz Valdez 2021). На распознавании речи строится работа все более широко используемых голосовых помощников. В России широко известной среди них является разработка «Яндекса» – Алиса (ENA, June 6, 2022)[17]. Алиса интегрирована с сервисами «Яндекса»: по голосовой команде она ищет информацию в одноименном браузере, узнает погоду на Яндекс.Погоде, данные о трафике – в Яндекс.Картах и т.д. Алиса может управлять системами умного дома и даже развлекать: играть с детьми в загадки, рассказывать сказки и анекдоты. Распознавание речи в голосовых помощниках облегчается тем, что им достаточно настроиться на голос определенного человека. Обзор современного состояния проблематики голосовых помощников можно найти в (Nasirian et al. 2017), а по общим проблемам распознавания речи – в (Nassif 2019).

Синтез речи уже активно применяется в информационно-справочных системах, в объявлениях об отправлении поездов, в приглашениях к стойке в аэропортах, к определенному окну в госучреждениях и т.д. Во всех случаях это ситуации с ограниченным спектром синтезируемых фраз. Наиболее простым способом синтеза речи является ее компоновка из заранее записанных фрагментов. Качество синтеза оценивается по сходству синтезированной речи с речью человека. В целом к настоящему времени не удалось создать высококачественные системы синтеза речи. Новейший обзор по синтезу речи представлен в (Tan 2021).

Перейдем к рекомендательным системам, с которыми сталкивалось, вероятно, большинство пользователей интернета. Рекомендательные системы предсказывают, какие объекты (фильмы, музыка, книги, новости, веб-сайты) будут интересны конкретному пользователю. Для этого они собирают информацию о пользователях, иногда в явном виде, просят их дать оценку объектам интереса, а чаще – в неявном виде, собирая информацию о поведении пользователей в интернете. Продуктивной оказалась следующая идея: люди, одинаково оценивавшие какие-либо объекты в прошлом, вероятнее всего, будут давать похожие оценки другим объектам и в будущем (Xiaoyuan & Khoshgoftaar 2009). Именно эта идея позволяет эффективно экстраполировать поведение пользователей. При разработке рекомендательных систем возникают чисто лингвистические проблемы, например учет синонимии. Такие системы должны понимать, что «детский фильм» и «фильмы для детей» – это одно и то же. По проблеме синонимии в рекомендательных системах см. работу (Moon 2019), а общий обзор представлен в (Patel & Patel 2020).

[17] https://dialogs.yandex.ru/store

Вопросно-ответные системы, или QA-системы, призваны обеспечивать ответы на естественном языке на вопросы пользователей, т.е. обладать естественно-языковым интерфейсом. Речь идет о поиске ответов в текстовой базе данных, которой располагают QA-системы. QA-системы, как и поисковики, предоставляют пользователю возможность искать информацию. Однако важным отличительным свойством QA-систем является то, что они позволяют найти такую информацию, о которой пользователь мог и не подозревать, например, соответствующие его вкусам, но не известные ему фильмы, которые он бы не смог найти с помощью поисковика. Очевидно, что качество QA-системы зависит от того, насколько полна база данных, т.е. есть ли в ней вообще ответ на поставленный вопрос, а также от технологий обработки вопросов и сопоставления их с информацией в базе данных. Обработка вопроса начинается с определения типа вопроса и ожидаемого ответа. Например, вопрос «Кто …» предполагает, что в ответе должно быть имя человека. Далее применяются сложные методы КЛ. QA-системы, аналогично рекомендательным системам, также сталкиваются с проблемой синонимии (Sigdel 2020). Обзор проблематики QA-систем можно найти в (Ojokoh & Adebisi 2018).

2.2. Методы компьютерной лингвистики

Все методы КЛ можно разделить на два больших класса: основанные на словарях и правилах (шаблонах) и основанные на машинном обучении. Эти два класса принципиально различаются по подходам. В основе словарей и правил лежат знания о языке, аккумулированные лингвистами. Это высокопрофессиональный ручной труд и поэтому весьма дорогостоящий. Машинное обучение предполагает наличие большого числа примеров, обычно в виде размеченных корпусов (обучающего множества), проанализировав которые и выявив их закономерности, компьютер сможет находить решение и при анализе новых данных. Существуют различные способы машинного обучения, однако наибольшие успехи в последнее время демонстрируют нейронные сети глубокого обучения. Языковые данные подаются на вход нейронной сети в закодированном виде в формате токенов: букв, биграмм, коротких высокочастотных морфем и слов. Сложностью в применении этого подхода является необходимость разметки большого корпуса текстов под решаемую задачу: чем больше обучающее множество, тем лучше обучится нейронная сеть. При этом разметка носит достаточно простой характер и для ее выполнения не обязательно привлечение профессиональных лингвистов, можно ограничиться просто носителями языка.

Остановимся на базовых методах КЛ, отсылая за детальным изложением вопроса к вышеупомянутым монографиям (Clark et al. 2013, Indurkhya & Damerau 2010).

Автоматический анализ текста обычно начинается с его предобработки, включающей сегментацию текста, т.е. его разбиение на слова и предложения.

Может показаться, что это несложные задачи, поскольку слова отделяются друг от друга пробелами, а предложения начинаются с заглавной буквы и заканчиваются точкой (редко – восклицательным или вопросительным знаками, многоточием) с последующим пробелом. Это простейший пример правила или шаблона: «точка – пробел – заглавная буква». Однако точка может стоять в середине предложения после первого инициала, за ней будет пробел и затем второй инициал с заглавной буквой. Здесь точка явно не указывает на разделение текста на предложения. В качестве примера можно привести такое предложение: «Лукашевич Н.В., Левчик А.В. Создание лексикона оценочных слов русского языка РуСентилекс // Труды конференции OSTIS-2016. С. 377–382». Тем не менее, несмотря на указанные сложности, проблема сегментации считается практически решенной. Еще в 1989 г. в (Riley 1989) была достигнута точность 99,8% в решении задачи разбиения текста на предложения. Для достижения такого результата потребовалась сложная система правил. В ней учитывались такие признаки, как длина слова перед точкой, длина слова после точки, наличие слова перед точкой в словаре аббревиатур и ряд других.

Следующий шаг в ходе анализа текста – морфологический. Рассмотрим в качестве примера язык со сложной морфологией – русский. Для русского языка морфологический анализ выполняется многими анализаторами: MyStem, Natasha, pymorphy2, SpaCy и др. В КЛ морфологический анализ, цель которого состоит в определении морфологических характеристик слова, основан на детальном описании парадигм словоизменения. Для русского языка справочник создан такого рода создан (Зализняк 1977), в котором представлены индексы парадигм почти 100 тыс. слов (лемм) русского языка. Наличие такого справочника позволило сгенерировать около 3 миллионов словоформ для зафиксированных лемм русского языка. Автоматический анализ текста находит соответствующую любой словоформе лемму и полный перечень морфологических характеристик. Главной сложностью, с которой существующие анализаторы пока не справляются полностью, является омонимия форм. Базовое решение состоит в том, что анализатор выдает все варианты разборов. Однако во многих задачах требуется указать единственное решение. В этом случае анализаторы выдают наиболее частотный вариант морфологического разбора, не учитывая значение слова в контексте. Еще одна проблема – это проблема разбора «несловарных» слов, т.е. слов, отсутствующих в словаре. Для их морфологического анализа, учитывая, что количество таких слов в среднем составляет около 3%, приходится разрабатывать специальные алгоритмы. В простейшем случае анализируется окончание несловарной единицы и ей приписывается типичная для этого окончания парадигма словоизменения.

Синтаксический анализ, или парсинг, намного более сложен. Результатом синтаксического парсинга предложения является дерево зависимостей,

отражающее структуру предложения либо в формализме генеративной грамматики, либо в формализме грамматики зависимостей (dependency grammar (Tesnière 2015)). Для успешного синтаксического разбора необходимо детальное описание синтаксиса языка. Для русского языка наиболее успешным признан анализатор проекта ЭТАП, разрабатываемый более 40 лет в Лаборатории компьютерной лингвистики Института проблем передачи информации РАН. Его последняя версия – ЭТАП-4 доступна по адресу (ENA, June 6, 2022)[18]. В основу синтаксического анализатора проекта ЭТАП положена хорошо известная модель «Смысл ⇔ Текст» (Мельчук 1974), ее формализованный вариант изложен в монографии (Апресян 1989). В последнее десятилетие конкурирующим стал подход на основе нейронных сетей (Chen & Manning 2014). Для обучения нейронных сетей используются базы данных предложений с их синтаксическим разбором. Для английского языка это, например, English Penn Treebank (ENA, June 6, 2022)[19]. Для русского языка можно использовать SynTagRus (ENA, June 6, 2022)[20], созданный в Лаборатории компьютерной лингвистики ИППИ РАН.

Еще более сложной следует признать задачу семантического анализа. Однако, если мы хотим, чтобы компьютер хотя бы в какой-то степени «понимал» смысл, необходимо, некоторым образом, формализовать семантику слов и предложений. Классическими в решении данной проблемы являются два направления. Первое направление инициировано Ч. Филлмором (Fillmore 1968), который ввел понятия семантических падежей или ролей именных групп в предложении. Правильное установление семантических ролей – важный шаг к пониманию предложения. Исходные идеи Ч. Филлмора были воплощены в компьютерной лексической базе данных FrameNet (ENA, June 6, 2022)[21].

Второе направление – это создание электронного тезауруса (лексической онтологии) WordNet (Fellbaum 1998) для английского языка и его аналогов – для многих других языков. Для русского языка было предпринято несколько попыток создания аналога WordNet, наиболее удачным из которых и широко используемым в настоящее время признан тезаурус RuWordNet (ENA, June 6, 2022 [22](Loukachevitch & Lashevich 2016)), содержащий более 130 тыс. слов. В WordNet-подобных тезаурусах эксплицированы семантические отношения между словами (понятиями), в том числе синонимия, гипонимия, гиперонимия и ряд других. Данные системные параметры в определенной степени уже определяют часть семантики слов. WordNet успешно использовался в большом числе как лингвистических, так и компьютерных исследований.

[18] http://proling.iitp.ru/ru/etap4
[19] https://catalog.ldc.upenn.edu/LDC99T42
[20] https://universaldependencies.org/treebanks/ru_syntagrus/index.html
[21] https://framenet.icsi.berkeley.edu/fndrupal/
[22] https://ruwordnet.ru/ru,

В последние годы была предложена идея векторного представления семантики слов (word embeddings), в основу которой положена дистрибутивная гипотеза: лингвистические единицы, встречающиеся в аналогичных контекстах, имеют близкие значения (Sahlgren 2008). Данная гипотеза подтверждена в ряде работ, в рамках которых созданы и исследованы векторы частот слов, зафиксированных в большом корпусе текстов в контексте изучаемых слов. Существует целый ряд уточнений и компьютерных реализаций этой идеи, однако используется преимущественно word2vec (Mikolov et al. 2013), доступная в библиотеке Gensim (ENA, June 6, 2022) [23] и пользующаяся наибольшей популярностью. Для русского языка существует система RusVectores (Kutuzov & Kuzmenko 2017), доступная по адресу: (ENA, June 6, 2022)[24] и выполняющая ряд операций со словами на основе их векторной семантики. RusVectores, например, может рассчитывать семантическую близость слов.

Разумеется, важнейшим инструментом исследований в КЛ, да и всей лингвистики в целом, являются корпусы текстов. Первым корпусом был созданный в 1960-е гг. *Brown Corpus*, содержащий на момент создания один миллион слов. С тех пор требования по объему корпусов стали неизмеримо выше. Для русского языка наиболее известен Национальный корпус русского языка (НКРЯ, ENA, June 6, 2022[25]). Созданный в 2004 г., он постоянно пополняется и в настоящий момент включает более 600 млн слов. В 2009 г. Google создал очень интересный многоязычный ресурс – Google Books Ngram (ENA, June 6, 2022)[26], содержащий 500 млрд слов, в том числе 67 млрд слов для русского языка (подробнее о данном ресурсе см. Michel 2011). Важной проблемой остается разметка корпусов, которая в сложных случаях осуществляется вручную. При этом важным является привлечение нескольких аннотаторов и контроль согласованности их разметок (Pons & Aliaga 2021). Несмотря на то, что корпусы стали неотъемлемым элементом лингвистических исследований, споры о репрезентативности, сбалансированности, дифференциальной полноте, предметной и жанровой отнесенности, корректности данных продолжаются. Обсуждение этих вопросов для корпуса Google Books Ngram можно найти в (Solovyev et al. 2020).

Подводя итог этому разделу статьи, отметим, что благодаря КЛ мы имеем такие уже ставшие привычными сервисы, как информационный поиск, автоматическая коррекция ошибок и многие другие. Это стало возможным благодаря принципиально важным достижениям не только в компьютерных науках, но и в лингвистике. В КЛ используются обширные словари и тезаурусы, детально проработанные модели синтаксиса, гигантские корпусы текстов. Автоматический морфологический анализ в современном виде

[23] https://github.com/rare-technologies/gensim
[24] https://rusvectores.org/ru/
[25] https://ruscorpora.ru/new/
[26] https://books.google.com/ngrams

просто не существовал бы без «Грамматического словаря русского языка» А.А. Зализняка (1977). Многие исследования в КЛ основаны на созданных вручную тезаурусах WordNet и RuWordNet. Компьютерные технологии, в свою очередь, вносят вклад в развитие лингвистики. Использование корпусов текстов, статистических методов стало уже общим местом, без этого проведение серьезных лингвистических исследований становится невозможным. Все ключевые технологии КЛ являются общедоступными. Программы для решения основных задач для ряда языков, но не для русского, доступны здесь (ENA, June 6, 2022)[27].

В одной статье, разумеется, невозможно дать исчерпывающее представление о столь обширной и быстро развивающейся области науки о языке, как компьютерная лингвистика. Многие важные вопросы остались незатронутыми. К ним можно отнести следующие: разрешение кореференции, снятие омонимии, тематическое моделирование и др., для знакомства с которыми следует обратиться к специальной литературе или указанным выше монографиям.

3. Сложность языка и текста как научная проблема

Ядром спецвыпуска является группа статей, посвященных оценке сложности текстов.

Оценка сложности языка в зависимости от количества имеющихся в его системе категорий представляется, на первый взгляд, весьма логичной, а сама задача – выполнимой. Иллюстрацией в данном случае могут служить, например, фонологический инвентарь языка, количество морфофонологических правил или форм глагола. Очевидной в данном случае становится возможность сравнительной оценки сложности разных языков и присвоения им некоторой объективной, абсолютной сложности (Miestamo et al. 2008). Добавим, что именно «объективная» сложность значима при освоении неродного языка. С другой стороны, если язык изучается как родной, он не представляет для детей сложности, и с этой точки зрения сложность всех языков абсолютно одинакова. Исследователи признаются, что сложность языка и текста «сопротивляется измерению», а ученые, работающие в этой области, сталкиваются с концептуальными и методологическими трудностями.

Значимым в свете изучаемой проблематики представляется описание взаимосвязи и взаимозависимости двух направлений изучения сложности: сложности языка (lingue), или языковой (лингвистической) комплексологии, с одной стороны, и сложности текста (parole) или дискурса (discourse complexity), или дискурсивной комплексологии, – с другой.

Трактовка самого понятия «сложность языка (lingue)» кардинально менялась в течение XIX–XX вв. В XIX в. общепринятым было выдвинутое В. Гумбольдтом положение о том, что различия в структуре языка и, следовательно, сложности определяют развитие говорящих на этом языке людей (Humboldt

[27] https://stanfordnlp.github.io/CoreNLP/

1999: 37). Признавая данное положение, ученые фактически соглашались с концепцией неравного статуса языков и народов. В XX в. на смену гумбольдианским взглядам, утверждающим неравные позиции языков и их носителей, пришла концепция единой, неизменной для всех языков мира сложности, получившая два названия: ALEC («All Languages are Equally Complex», букв. «Все языки одинаково сложны») (Deutscher 2009: 243) и linguistic equi-complexity dogma – букв. лингвистическая догма равной сложности (Kusters 2003: 5). В работах ученых, поддерживающих данную концепцию, доказательству подлежали две гипотезы: (1) сложность языка складывается из под-сложностей (sub-complexities) его элементов; (2) все под-сложности в лингвистических подсистемах компенсированы: простота в области А компенсируется сложностью в области В, и наоборот («компенсаторная гипотеза»). Аргументируя концепцию «Все языки одинаково сложны», Ч. Хоккет весьма смело заявил: «Объективное измерение сложности затруднено, но субъективно понятно, что общая грамматическая сложность любого языка, включая его морфологию и синтаксис, примерно одинакова. Это неудивительно, поскольку все языки выполняют одни и те же функции: что не может быть сделано «морфологически», должно быть сделано «синтаксически» (Hockett 1958: 180–181). К сожалению, в работах данного направления и периода традиционно не обсуждались критерии оценки сложности, а эмпирические доказательства попросту отсутствуют. Подробный обзор точек зрения о «догме равной сложности» представлен в основополагающей работе Дж. Сэмпсона, Д. Гила и П. Традгилла «Сложность языка как эволюционирующая переменная» (Sampson et al. 2009).

Начало XXI в. ознаменовалось появлением ряда критических обзоров теории равной сложности всех языков, с одной стороны (см. Miestamo, Sinnemäki & Karlsson 2008), и провокационным заявлением Дж. Маквортера о том, что «креольские грамматики – самые простые грамматики в мире» (McWhorter 2001). Сама же идея о том, что все языки одинаково сложны, была доказательно отвергнута социолингвистами, которые продемонстрировали, что языковой контакт может привести к упрощению языка. Это показано на примере африкаанс, пиджинов и койне. Если признать возможность упрощения языка, то отсюда неизбежно следует, что до упрощения язык был сложнее, чем после. И если язык может быть более или менее сложным на разных этапах своей истории, то очевидно, что одни языки могут быть более сложными, чем другие (Trudgill 2012).

В начале 2000-х гг. идея о лингвистической сложности и «догме равной сложности» начала активно обсуждаться на конференциях и семинарах (см. семинар «Сложность языка как развивающаяся переменная», организованный Институтом эволюционной антропологии им. Макса Планка в 2007 г. в Лейпциге ENA, June 6, 2022[28]), в ряде журнальных статей (Shosted 2006,

[28] https://www.eva.mpg.de/fileadmin/content_files/linguistics/pdf/ComplexityWS_Webpage_2007.pdf

Trudgill 2004) и монографий (Даль 2009, Kusters 2003, Miestamo et al. 2008, Sampson et al. 2009).

В России публикации по сложности языка до сих пор малочисленны и преимущественно представлены обзорами, выполненными зарубежными учеными, однако в последнее время некоторый интерес к данной проблеме начал возрастать. Из наиболее значимых следует указать на статью А. Бердичевского (2012) и рецензию на книгу Питера Трандгилла «Sociolinguistic Typology», опубликованную в 2011 г. (Вахтин 2014). Проблемы сложности языка обсуждались в Институте лингвистических исследований Российской академии наук (ИЛИ РАН) в 2018 г. на конференции «Балканские языки и диалекты: корпусные и квантитативные исследования».

Локальная и глобальная сложность

Развитие лингвистической комплексологии привело к выделению двух типов сложности: глобальной, т.е. сложности языка (или диалекта) в целом, и локальной сложности, т.е. сложности отдельного уровня языка или домена (Miestamo 2008). И если оценка глобальной сложности языка, по мнению ученых, является весьма амбициозной и, вероятно, безнадежной задачей, сравниваемой Г. Дойчером с «погоней за дикими гусями» (Deutscher 2009), то измерение локальной сложности рассматривается учеными как вполне выполнимая задача, состоящая в составлении перечня и оценке предикторов сложности, объективируемых на различных уровнях языка. Список предикторов *фонологической сложности* традиционно включает объем инвентаря фонем, частоту встречаемости маркированных[29] фонем, тональные различия, супрасегментные модели, фонотактические ограничения и максимальные кластеры согласных (Nichols 2009, Shosted 2006). При оценке *морфологической сложности* классическими «факторами неудобств» (термин Браунмюллера 1990: 627) признаны объем флективной морфологии языка (или языковой разновидности), специфика алломорфии и морфофонемных процессов и др. (Dammel & Kürschner 2008, Kusters 2003). Расчет *синтаксической сложности* осуществляется на основе данных о количестве предписываемых синтаксисом языка правил по принципу «чем больше, тем сложнее», а также способности языка порождать рекурсии и клаузы внутри синтаксического целого (Ortega 2003, Givón 2009, Karlsson 2009). *Семантическая и лексическая сложность* трактуется на основе следующих параметров: количества неоднозначных единиц языка, различия инклюзивных и эксклюзивных местоимений, лексического многообразия и др. (Fenk-Oczlon & Fenk 2008, Nichols 2009). *Прагматическая*, или «скрытая», сложность, имеющая в своей основе закон экономии, есть сложность умозаключений, необходимых для восприятия текстов на данном языке. Языки со скрытой сложностью допускают минималистские, весьма простые поверхностные структуры, интерпретация грамматических категорий в которых требует нетривиальных умозаключений. В качестве примера исследователи приводят языки Юго-Восточной

[29] Маркированными считаются фонемы, редко встречающиеся в языках мира (Бердичевский 2012).

Азии, достигшие особенно высокой степени скрытой сложности, в частности за счет опущения местоимений, множественной кореференции в относительных предложениях, отсутствия маркеров отношений и «голых», без модификаторов, существительных с широким диапазоном интерпретаций (Bisang 2009).

Исследования показали, что высокие уровни локальной сложности одного уровня в языке необязательно влекут за собой низкую локальную сложность другого уровня, как это прогнозируется «догмой равной сложности». Например, анализ метрик морфологической и фонологической сложности в 34 языках, осуществленных Р. Шостедом, не выявил ожидаемой статистически значимой корреляции (Shosted 2006). А наблюдаемые Г. Фенк-Озлог и А. Фенком отдельные «балансирующие эффекты» (trade-offs) между локальными сложностями, к сожалению, также недостаточны, чтобы валидировать «догму равной сложности» языков. Г. Фенк-Озлог и А. Фенк, в частности, выявили, что в английском языке тенденция к фонологической сложности и односложности связана с тенденцией к омонимии и многозначности, к твердому порядку слов и идиоматичности речи (Fenk-Oczlon & Fenk 2008: 63). Д. Гил убедительно доказал, что изолирующие языки не обязательно компенсируют простую морфологию более сложным синтаксисом (Gil 2008).

Факторы (или предикторы) сложности языка принято делить на внутренние и внешние. *Внутренними факторами* сложности признаются количество элементов и категорий в языке, избыточность и нерегулярность языковых категорий. При оценке внутренней сложности в современных исследованиях весьма распространенным является так называемый «списочный подход», при котором ученые составляют список языковых явлений, присутствие которых в языке увеличивает степень его сложности, т.е. фактически списки предикторов внутренней сложности суть списки локальной сложности, описанной выше. Например, список предикторов сложности, составленный Дж. Николз, содержит более 18 параметров и включает фонологические, морфологические, синтаксические и лексические параметры (Nichols 2009). Язык считается более сложным, если в нем больше маркированных фонем, тонов, синтаксических правил, грамматически выраженных семантических и/или прагматических различий, морфофонемных правил, больше случаев дополнения, алломорфии, согласования и др. Ученых, работающих в рамках данного направления, интересует, например, количество грамматических категорий в языке (Shosted 2006), число фонематических оппозиций (McWhorter 2008), длина «минимального описания» системы языка (Даль 2009). Для иллюстрации упрощения языка при утрате предиктора Макуортер (2001) сравнивает порядок слов, т.е. позицию глагола в германских языках, доказывая, что синтаксис английского языка имеет более низкую степень сложности, чем шведский и немецкий. Причина положения состоит в утрате английским языком правила V2 (verb-second), в соответствии с которым личный глагол в шведском и немецком занимает второе место в предложении.

В качестве «избыточных» внутренних предикторов сложности признаются элементы и функции в системе языка, которые несут «дублирующую» информацию или «излишнюю спецификацию», букв. overspecification, и поэтому являются коммуникативно необязательными элементами (McWhorter 2008). П. Традгилл именует такого рода элементы «историческим багажом», букв. historical baggage (Trudgill 1999: 149), В.М. Жирмунский – «гиперхарактеризацией» (Жирмунский 1976), Макуортер – «декоративным украшением», букв. ornamental elaboration, или «барочными образованиями», букв. baroque accretion[s] (McWhorter 2001). В качестве иллюстрации синтагматической избыточности традиционно называют косвенную (непрямую) номинацию и «семантическое согласование». Иллюстрацией парадигматической избыточности в языке выступает синтетическое выражение грамматических категорий, например маркирование при согласовании (Избыточность в грамматическом строе языка) и маркирование обвиатива (см. McWhorter 2001).

Нерегулярность или «непрозрачность» формо- и словообразовательных процессов как внутренний фактор сложности языка (см. Mühlhäusler 1974) реализуется в нерегулярных аффиксах, встречающихся в отдельных словах (приставки па- (пасынок), су- (сумрак), низ- (низводить), суффиксы -таш (патронташ), -ичок (новичок), -арник (кустарник) (см. Казак 2012).

Внешними факторами, детерминирующими сложность языка, признаются культура, возраст языка и языковые контакты. Считается, что старые языки, обслуживающие хорошо развитые многоуровневые культуры, являются более сложными, поскольку аккумулировали «зрелые языковые черты», букв. mature language features (термин О. Даля (2009) (Deutscher 2010, Parkvall 2008). Вместе с тем существенное влияние на сложность языков оказывают интенсивные контакты между языковыми сообществами. В начале нашего столетия П. Традгилл заявил, что «небольшие, изолированные сообщества с низким уровнем контактов, имеющие тесные социальные сети», развивают более сложные языки, чем сообщества с высоким уровнем контактов (Trudgill 2004: 306). Однако в своей более поздней работе исследователь уточняет, что динамика развития сложности языков при их взаимодействии определяется длительностью контактов и возрастом носителей, осваивающих суперстрат: упрощение языка имеет место при кратковременных контактах сообществ, когда иностранный (второй) язык усваивают взрослые. Усложнение языка может иметь место в тех случаях, когда контакт долговременный, а второй язык осваивается не взрослыми, а детьми (Trudgill 2011). Для доказательства влияния языковых контактов на сложность языка Б. Кортман и Б. Смречаньи (2004) сравнивают способы реализации 76 морфосинтаксических параметров, включая количество местоимений, модели именных групп, время и вид, модальные глаголы, морфологию глагола, наречия, способы выражения отрицаний, согласование, порядок слов и др., в 46 вариантах английского языка. Ученые делят все варианты английского языка на три большие группы:

(1) родные для их носителей и выполняющие все функции в языковом сообществе; (2) языки, функционирующие как второй официальный язык государства, и (3) креольские языки, имеющие в основе английский. Исследование подтвердило, что третья группа языков, т.е. креольские языки, имеющие в основе английский язык, наименее сложны, разновидности английского как родного (первого) языка являются наиболее сложными, а разновидности английского языка, используемого носителями в качестве второго языка, демонстрируют промежуточную сложность (Kortmann & Szmrecsanyi 2004).

В самых общих чертах *аналитические методы* оценки сложности делятся на *абсолютные* (теоретико-ориентированные и трактуемые как «объективные») и *относительные* (ориентированные на пользователя и, таким образом, «субъективные[30]») (Crossley et al. 2008). Абсолютный подход популярен в лингвистической типологии и используется для оценки сложности языка, в то время как в социолингвистике и психолингвистике используется относительный подход. П. Традгилл определяет относительную сложность как трудность изучения иностранного языка взрослыми (Trudgill 2011: 371). Сложность текста как конструкт также моделируется в дискурсологии, лингвистической персонологии, в психолингвистике и нейролингвистике. При этом изучается относительная сложность (трудность) текста для разных категорий реципиентов в различных условиях коммуникации, а также абсолютная и относительная (сравнительная) сложность текстов, генерируемых различными авторами (см. McNamara et al. 1996, Солнышкина 2015).

4. Краткий обзор статей выпуска

Современный подход к оценке сложности текстов характеризуется использованием как комплекса лингвистических методов исследования, так и достаточно сложного аппаратного и программного инструментария. Основные идеи весьма полно представлены в настоящем выпуске. Важным способом объективной оценки сложности текста для читающего является методика отслеживания движения глаз, осуществляемого с помощью специального оборудования – систем айтрекинга. Для русского языка исследования в этом направлении только начинаются. В качестве базовой ученые вудвигают задачу выбора параметров текста и глазодвигательной активности, а также меры сложности восприятия текста. Обычно в качестве параметров текста выбираются средняя длина слов и средняя частотность, а в качестве параметров глазодвигательной активности: относительная скорость чтения слова, длительность фиксаций и количество фиксаций. Мерой читабельности текста является скорость чтения вслух в словах в минуту. Айтрекингу посвящены статьи А.Н. Лапошиной с соавторами и А.А. Бонч-Осмоловской с соавторами.

[30] Характеристика этого типа сложности как субъективной может быть принята условно, поскольку она является вполне объективной для всех участников коммуникации. Более подходящим являлось бы определение этого типа сложности как «индивидуальной».

В первой из вышеуказанных работ показано, что число фиксаций на слове коррелирует с его длиной, а длительность фиксаций – с частотностью. Вторая статья посвящена более сложным элементам текста – элементарным дискурсивным единицам (ЭДЕ), трактуемой как «квант устного дискурса, минимальный шаг, при помощи которого говорящий продвигает дискурс вперед» (Подлесская, Кибрик 2009: 309). Структура ЭДЕ также влияет на читабельность текста и это фиксируется с помощью айтрекинга.

Оценке сложности текстов с помощью наиболее современных методов глубокого обучения нейронных сетей посвящены работы Д. Корталеску с соавторами, С.А. Шарова, Д.А. Морозова с соавторами и В.В. Иванова с А.В. Абрамовым. Объект исследования – тексты, предназначенные для изучающих русский язык как иностранный. Точная оценка их сложности позволит правильно выбирать тексты в той или иной образовательной ситуации. Как отмечалось в первом разделе статьи, в качестве инструмента исследований используется, в первую очередь, модель BERT. Ее применение позволяет достичь высокой точности в определении сложности этого типа текстов – 91–92%.

Применение нейронных сетей предполагает успешное решение важной исследовательской лингвистической проблемы, а именно, определение признаков текстов, влияющих на решение нейронной сети. Один из возможных подходом состоит в том, чтобы вычислить коэффициенты корреляции ряда лингвистических признаков текста с оценками сложности текста нейронной сетью. Исследование на обширном материале коллекций текстов разных жанров на английском и русском языках с учетом десятков языковых признаков позволило обнаружить ряд неочевидных эффектов. Например, оказалось, что большее число предлогов ассоциируется с более сложными текстами в русском и с более простыми текстами в английском. Очевидно, это связано с различием в типологической структуре языков. Впрочем, на взаимосвязь языковых признаков текста с его сложность даже в большой мере влияет жанр текста.

Широкий обзор применения иных средств компьютерной лингвистики в проблематике сложности текстов дан в работе М.И. Солнышкиной с соавторами. В этой работе описана динамика развития и предложена периодизация в виде 6 парадигм дискурсивной комплексологии: формирующей, классической, периода закрытых текстов, структурно-когнитивного периода, периода обработки естественного языка, периода искусственного интеллекта.

Важной отличительной особенностью статей данного спецвыпуска и его вклада в дискурсивную комплексологию является учет огромного числа разнообразных данных: несколько сот языковых признаков, разные языки, разные корпусы текстов, разные жанры. Сложность текста рассматривается на нескольких уровнях: лексическом, морфологическом, синтаксическом, дискурсивном. Столь многоплановые исследования позволяют глубже понять природу самого понятия сложность текста. В статьях выпуска используются

не только уже существующие готовые корпусы текстов и словари, но описывается создание новых.

Степень абстрактности также рассматривается в качестве важнейшего параметра сложности текста. Чем больше абстрактных слов текст содержит, тем он сложнее. Это означает необходимость создания словарей абстрактной/конкретной лексики и средств расчета степени абстрактности текста. Ранее словари абстрактных/конкретных слов были созданы для английского и некоторых других языков, но не для русского. В статье В.Д. Соловьева с соавторами подробно описывается методология создания такого словаря для русского языка. Показано, как этот словарь может быть использован и в других исследованиях, кроме проблематики сложности.

Лингвистическая сложность представляет собой междисциплинарную проблему, которая изучается не только компьютерной лингвистикой, но также в рамках нескольких научных направлений: философии, прикладной лингвистики, психологии, нейролингвистики. В XXI в. проблематика сложности обрела собственный терминологический аппарат, разработала и верифицировала широкий спектр лингвистических параметров сложности, а основным достижение новой парадигмы стала валидация когнитивных предикторов сложности, поднявшая проблематику текста на новый уровень – уровень дискурса. Этот успех, а также междисциплинарный подход к проблеме позволили интегрировать исследования сложности дискурса в отдельную область – дискурсивную комплексологию. Проблематики сложности – не «вещь в себе», поскольку результаты исследований релевантны как для лингвистического анализа текста, так и для прогнозирования успешности восприятия информации в широком спектре прагмалингвистических ситуаций.

Одной из таких ситуаций является когнитивный анализ ошибок, допускаемых при изучении иностранного языка. Этой проблематике посвящены работы О.Н. Ляшевской с соавторами и Л. Янды с соавторами. В них исследования выходят на уровень взаимосвязей между сложностью текстов и когнитивными ресурсами, необходимыми для их понимания. В первой работе получен следующий интересный результат: чем сложнее используемые обучающимся аффиксы, тем меньше он допускает ошибок в текстах. Во второй работе описана компьютерная система, предназначенная для анализа и адекватного объяснения ошибок изучающего русский язык как иностранный.

5. Заключение

Успехи компьютерной лингвистики последних лет во многом обеспечили достижения дискурсивной комплексологии и позволили ученым не только автоматизировать ряд операций лингвистического анализа, но и создать удобные для пользователей профайлеры текстов. Такие инструменты, как ReaderBench, Coh-Metrix и RuMOR (подробно описанные в статьях данного выпуска) способны решать как исследовательские, так и практиче-

ские задачи: осуществлять подбор текстов для целевой аудитории, редактировать и сокращать тексты, производить анализ когнитивных причин возникновения ошибок и даже предлагать стратегии вербального поведения. Алгоритмы, используемые разработчиками при создании инструментов автоматического анализа текстов, имеют в своей основе классические методы и методы машинного обучения, включая нейронные сети глубокого обучения и одну из новейших систем – систему BERT. В настоящее время, и это хорошо показано в ряде статей спецвыпуска, ученые успешно совмещают методы машинного обучения и «параметрического подхода».

Однако важнейшей особенностью современных исследований является значительное расширение научной проблематики и повышение точности расчетов за счет способности искусственных нейронных сети к обучению и модификации. Прорыв в области искусственного интеллекта был обусловлен тремя основными факторами: появлением новых, более совершенных алгоритмов самообучения, повышением скорости работы компьютеров, многократным увеличением объема данных для обучения. Современные базы данных, а также разработанные в последние годы словари и инструменты для русского языка позволили авторам спецвыпуска обратиться и успешно решить целый ряд проблем в области сложности текста.

Еще одним фундаментом успеха в области сложности текста послужили открытия ученых когнитологов, сделанные в начале нашего века и навсегда поменявшие научную парадигму комплексологии. Если основным достижением комплексологии текста XX в. являлся вывод о том, что «разные типы текстов сложны по-разному», то дискурсивная комплексология XXI в. не только сумела предложить и верифицировать предикторы сложности для различных типов текстов, но разработала инструментарий для оценки относительной сложности текста в различных коммуникативных ситуациях. С обращением к когнитивным наукам комплексология обрела две дополнительные переменные: языковую личность читателя и коммуникативную ситуацию процесса чтения.

Новая исследовательская парадигма лингвистической комплексологии также отражена в тех работах спецвыпуска, которые посвящены поиску новых критериев сложности текста: на смену экспертной оценке, тестам на понимание и скорости чтения пришли новые методы, позволяющие выявлять дискурсивные единицы, влияющие на сложность восприятия текста.

Исследования, публикуемые в специальном выпуске высветили и основные проблемы, стоящие перед отечественной лингвистической комплексологией: создание матрицы сложности текстов различных типов и жанров, расширение списка предикторов сложности, валидация новых критериев сложности, расширение баз данных для русского языка.

Благодарность

Работа выполнена за счет средств Программы стратегического академического лидерства Казанского (Приволжского) федерального университета (ПРИОРИТЕТ-2030).

Acknowledgments

This paper has been supported by the Kazan Federal University Strategic Academic Leadership Program (PRIORITY-2030).

REFERENCES / СПИСОК ЛИТЕРАТУРЫ

Апресян Ю.Д., Богуславский И.М., Иомдин Л.Л., Лазурский А.В., Перцов Н.В., Санников В.З., Цинман Л.Л. Лингвистическое обеспечение системы ЭТАП-2. М.: Наука, 1989. [Apresyan, Yurii D., Igor M. Boguslavskii, Leonid L. Iomdin, Aleksandr V. Lazurskii, Nikolai V. Pertsov, Vladimir Z. Sannikov, Leonid L. Tsinman. 1989. *Lingvisticheskoe obespechenie sistems ETAP-2* (*Linguistic support of the system STAGE-2*). Moscow: Nauka. (In Russ.)].

Бердичевский А. Языковая сложность // Вопросы языкознания. 2012. № 5. С. 101–124. [Berdichevskii, Aleksandr. 2012. Yazykovaya slozhnost' (Language complexity). *Voprosy yazykoznaniya* 5. 101–124.] (In Russ.)

Вахтин, Н. Рец. на кн.: Peter Trudgil. Sociolinguistic Typology: Social Determinants of Linguistic Complexity // *Антропологический форум.* 2014. № 2. С. 301–309. [Vakhtin, Nikolai. 2014. Review of Peter Trudgil. Sociolinguistic Typology: Social Determinants of Linguistic Complexity. *Antropologicheskii Forum* 2. 301–309. (In Russ.)].

Даль Э. Возникновение и сохранение языковой сложности. М.: ЛКИ, 2009. [Dahl, Osten. 1976. Vozniknovenie i sokhranenie yazykovoi slozhnosti (The emergence and persistence of language complexity). Moscow: LKI. (In Russ.)].

Жирмунский В.М. Общее и германское языкознание: Избранные труды. Л.: Наука, 1976. [Zhirmunskii, Viktor M. 1976. Obshchee i germanskoe yazykoznanie: Izbrannye trudy (General and Germanic Linguistics: Selected works). Leningrad: Nauka. (In Russ.)].

Зализняк А.А. Грамматический словарь русского языка. М.: Русский язык, 1977. [Zaliznyak, Andrei A. 1977. Grammaticheskii slovar' russkogo yazyka (Grammatical dictionary of the Russian language). Moscow. (In Russ.)].

Избыточность в грамматическом строе языка / под ред. М.Д. Воейковой. СПб.: Наука, 2010. [Voeikova, Mariya D. (ed.). 2010. *Izbytochnost' v grammaticheskom stroe yazyka (Redundancy in the Grammatical Structure of the Language).* Saint Petersburg: Nauka. (In Russ.)].

Казак М.Ю. Морфемика и словообразования современного русского языка. Теория. Белгород: ИД «Белгород», 2012. [Kazak, Mariya Yu. 2012. Morfemika i slovoobrazovaniya sovremennogo russkogo yazyka. Teoriya (Morphemics and word formation of the modern Russian language. Theory). Belgorod: ID «Belgorod». (In Russ.)].

Кибрик А.А., Подлесская В.И. (ред.). Рассказы о сновидениях. Корпусное исследование устного русского дискурса. М.: Языки славянских культур, 2009. [Kibrik, A. A. & V. I. Podlesskaya (eds.). 2009. *Night Dream Stories: A Corpus Study of Russian Spoken Discourse.* Moscow: Yazyki slavyanskikh kul'tur. (In Russ.)].

Маннинг К.Д., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2011. [Manning, Kristofer D., Prabkhakar Ragkhavan & Khinrich Shyuttse. 2011.

Vvedenie v informatsionnyi poisk (Introduction to Information Search). Moscow: Vil'yams. (In Russ.)].

Мельчук И.А. Опыт теории лингвистических моделей «Смысл ⇔ Текст». М., 1974. [Mel'chuk, Igor' A. 1974. Opyt teorii lingvisticheskikh modelei «Smysl ⇔ Tekst» (The experience of the theory of linguistic models «Meaning ⇔Text»). Moscow. (In Russ.)].

Подлесская В.И., Кибрик А.А. Дискурсивные маркеры в структуре устного рассказа: Опыт корпусного исследования // *Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегод. Междунар. конф. «Диалог».* 2009. Вып. 8 (15). С. 390–396. [Podlesskaya, V.I. & Kibrik A.A. 2009. Diskursivnye markery v strukture ustnogo rasskaza: Opyt korpusnogo issledovaniya (Discursive mrkers in the structure of oral narrative: The Experience of Corpus Research). In *Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Proceedings of the Annual international conference Dialogue* 8(15). 390–396].

Солнышкина М.И., Кисельников А.С. Сложность текста: Этапы изучения в отечественном прикладном языкознании // *Вестник Томского государственного университета. Филология.* 2015. № 6. С. 86–99. [Solnyshkina, M.I., Kise'nikov, A.S. 2015. Slozhnost' teksta: Ehtapy izucheniya v otechestvennom prikladnom yazykoznanii (Text complexity: Stages of study in domestic applied linguistics). *Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya* 6. 86–99].

Allahyari, Mehdi, Seyedamin Pouriyeh, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez & Krys Kochut. 2017. Text summarization techniques: A brief survey. *arXiv* 1707.02268, URL: https://arxiv.org/pdf/1707.02268.pdf. (accessed 20.01.2022).

Batrinca, Bogdan & Philip Treleaven. 2015. Social media analytics: A survey of techniques, tools and platforms. *AI & Soc* 30 (1). 89–116. https://doi.org/10.1007/s00146-014-0549-4

Bisang, Walter. 2009. On the evolution of complexity: Sometimes less is more in East and mainland Southeast Asia. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 34–49. Oxford, New York: Oxford University Press.

Braunmüller, Kurt. 1990. Komplexe flexionssysteme – (k)ein problem für die natürlichkeitstheorie? *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung* 43. 625–635.

Cambria, Erik, Dipankar Das, Sivaji Bandyopadhyay & Antonio Feraco (eds.). 2017. *A Practical Guide to Sentiment Analysis*. Cham, Switzerland: Springer International Publishing.

Chen, Danqi & Christopher D. Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 740–750. https://doi.org/10.3115/v1/D14-1082

Church, Kenneth & Mark Liberman. 2021. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence* 4. 625341. https://doi.org/10.3389/frai.2021.625341

Cinelli, Matteo, Walter Quattrociocchi, Alessandro Galeazzi, Carlo Michele Valensise, Emanuele Brugnoli, Ana Lucia Schmidt, Paola Zola, Fabiana Zollo & Antonio Scala. 2020. The COVID-19 social media infodemic. *Sci Rep* 10. 16598. https://doi.org/10.1038/s41598-020-73510-5

Clark, Alexander, Chris Fox & Shalom Lappin (eds.). 2013. *The Handbook of Computational Linguistics and Natural Language Processing*. John Wiley & Sons.

Crossley, S.A., Greenfield, J. & McNamara, D. S. 2008. Assessing Text Readability Using Cognitively Based Indices. *TESOL Quarterly*, 42 (3), 475–493.

Dammel, Antje & Sebastian Kürschner. 2008. Complexity in nominal plural allomorphy. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 243–262. Amsterdam, Philadelphia: Benjamins.

Deutscher, Guy. 2009. «Overall complexity»: A wild goose chase? In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 243–251. Oxford: Oxford University Press.

Deutscher, Guy. 2010. *Through the Language Glass: Why the World Looks Different in Other Languages*. New York: Metropolitan Books.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv 1810.04805v2*. URL: https://arxiv.org/pdf/1810.04805.pdf. (accessed 20.01.2022).

Domingue, John, Dieter Fensel & James A. Hendler (eds.). 2011. *Handbook of Semantic Web Technologies*. Springer Science & Business Media.

Fellbaum, Christiane (ed.). 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Fenk-Oczlon, Gertraud & August Fenk. 2008. Complexity trade-offs between the subsystems of language. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 43–65. Amsterdam, Philadelphia: Benjamins.

Fillmore, Charles J. 1968. The case for case. In Emmon W. Bach & Robert T. Harms (eds.), *Universals in Linguistic Theory*, 1–88. New York, NY: Holt, Rinehart & Winston.

Ghani, Norjihan A., Suraya Hamida, Ibrahim AbakerTargio Hashemb & Ejaz Ahmedc. 2019. Social media big data analytics: A survey. *Computers in Human Behavior* 101. 417–428. https://doi.org/10.1016/j.chb.2018.08.039

Gil, David. 2008. How complex are isolating languages? In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 109–131. Amsterdam, Philadelphia: Benjamins.

Givón, Thomas. 2009. *The Genesis of Syntactic Complexity: Diachrony, Ontogeny, Neuro-Cognition, Evolution*. Amsterdam, Philadelphia: Benjamins.

Hoang, Mickel, Oskar Alija Bihorac & Jacobo Rouces. 2019. Aspect-based sentiment analysis using BERT. In Mareike Hartmann & Barbara Plank (eds.), *Proceedings of the 22nd Nordic conference on computational linguistics*, 187–196. Turku, Finland: Linköping University Electronic Press Publ.

Hockett, Charles F. 1958. *A Course in Modern Linguistics.* New York: Macmillan.

Humboldt, Wilhelm von. 1999. *On Language: On the Diversity of Human Language Construction and its Influence on the Mental Development of the Human Species*. Cambridge, U.K. New York: Cambridge University Press.

Hutchins, John. 1999. Retrospect and prospect in computer-based translation. In *Proceedings of MT Summit VII «MT in the Great Translation Era»*. 30–44. Tokyo: AAMT.

Indurkhya, Nitin & Fred J. Damerau (eds.). 2010. *Handbook of Natural Language Processing*. CRC Press.

Jiang, Ridong, Rafael E. Banchs & Haizhou Li. 2016. Evaluating and combining name entity recognition systems. In Nancy Chen, Rafael E. Banchs, Xiangyu Duan, Min Zhang & Haizhou Li (eds.), *Proceedings of NEWS 2016. The Sixth named entities workshop*, 21–27. Berlin, Germany.

Karlsson, Fred. 2009. Origin and maintenance of clausal embedding complexity. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 192–202. Oxford: Oxford University Press.

Kortmann, Bernd & Benedikt Szmrecsanyi. 2004. Global synopsis: Morphological and syntactic variation in English. In Bernd Kortmann, Edgar Schneider Werner, Clive Upton,

Kate Burridge & Rajend Mesthrie(eds.), *A Handbook of varieties of English*, 1142–1202. Berlin, New York: Mouton de Gruyter.

Kusters, Wouter. 2003. *Linguistic Complexity: The Influence of Social Change on Verbal Inflection.* Utrecht: LOT.

Kutuzov, Andrey & Elizaveta Kuzmenko. 2017. WebVectors: A toolkit for building web interfaces for vector semantic models. In Wil M. P. van der Aalst, Dmitry I. Ignatov, Michael Khachay, Sergei O. Kuznetsov, Victor Lempitsky, Irina A. Lomazova, Natalia Loukachevitch, Amedeo Napoli, Alexander Panchenko, Panos M. Pardalos, Andrey V. Savchenko &Stanley Wasserman (eds.), *Analysis of Images, Social Networks and Texts,* 155–161. Moscow: AIST.

Lauriola, Ivano, Alberto Lavelli & Fabio Aiolli. 2022. An introduction to deep learning in natural language processing: Models, techniques, and tools. *Neurocomputing* 470. 443–456. https://doi.org/10.1016/j.neucom.2021.05.103

Loukachevitch, Natalia V. & Anatolii Levchik. 2016. Creating a general Russian sentiment lexicon. In *Proceedings of Language Resources and Evaluation Conference LREC-2016*.

Loukachevitch, Natalia V. & G. Lashevich. 2016. Multiword expressions in Russian Thesauri RuThes and RuWordNet. In *Proceedings of the AINL FRUCT*. 66–71. Saint-Petersburg.

McNamara, Danielle S., Elieen Kintsch, Nancy Butler Songer & Walter Kintsch. 1996. Are Good Texts Always Better? Interactions of Text Coherence, Background Knowledge, and Levels of Understanding in Learning from Text. *Cognition and Instruction*, 14 (1), 1–43

McWhorter, John. 2001. The world's simplest grammars are creole grammars. *Linguistic Typology* 6. 125–166. https://doi.org/10.1515/LITY.2001.001

McWhorter, John. 2008. Why does a language undress? Strange cases in Indonesia. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 167–190. Amsterdam, Philadelphia: Benjamins.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian veres, Matthew K. Gray, The Google books team, Joseph P. Pickett & Dale Hoiberg. 2011. Quantitative analysis of culture using millions of digitized books. *Science* 331 (6014). 176–182. https://doi.org/10.1126/science.1199644

Miestamo, Matti, Kaius Sinnemäki & Fred Karlsson (eds.). 2008. *Language Complexity: Typology, Contact, Change.* Amsterdam, Philadelphia: John Benjamins.

Miestamo, Matti. 2008. Grammatical complexity in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 23–42. Amsterdam, Philadelphia: Benjamins.

Mikolov, Thomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv 1301.3781*. URL: https://arxiv.org/abs/1301.3781 (accessed 20.01.2022).

Miranda-Jiménez, Sabino, Alexander Gelbukh & Grigori Sidorov. 2013. Summarizing conceptual graphs for automatic summarization task. In *Conceptual Structures for STEM Research and* Education. 245–253. *Lecture Notes in Computer Science* 7735.

Moon, Chang Bae, Jong Yeol Lee, Dong-Seong Kim & Byeong Man Kim. 2020. Multimedia content recommendation in social networks using mood tags and synonyms. *Multimedia Systems* 26 (6). 1–18. https://doi.org/10.1007/s00530-019-00632-w

Mühlhäusler, Peter. 1974. *Pidginization and Simplification of Language*. Canberra: Dept. of Linguistics, Research School of Pacific Studies, Australian National University.

Nasirian, Farzaneh, Mohsen Ahmadian & One-Ki D. Lee. 2017. *AI-based Voice Assistant Systems: Evaluating from the Interaction and Trust Perspectives*. Twenty-third Americas Conference on Information Systems, Boston.

Nassif, Ali Bou, Ismail Shahin, Imtinan Attili, Mohammad Azzeh & Khaled Shaalan. 2019. Speech recognition using deep neural networks: A systematic review. *IEEE access* 7. 19143–19165. https://doi.org/10.1109/ACCESS.2019.2896880

Nichols, Johanna. 2009. Linguistic complexity: A comprehensive definition and survey. In Geoffrey Sampson, David Gil & Peter Trudgill (eds.), *Language complexity as an evolving variable*, 64–79. Oxford: Oxford University Press.

Ojokoh, Bolanle & Emmanuel Adebisi. 2018. A review of question answering systems. *Journal of Web Engineering* 17 (8). 717–758. https://doi.org/10.13052/jwe1540-9589.1785

Ortega, Lourdes. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24. 492–518.

Parkvall, Mikael. 2008. The simplicity of creoles in a cross-linguistic perspective. In Matti Miestamo, Kaius Sinnemäki & Fred Karlsson (eds.), *Language complexity: Typology, contact, change*, 265–285. Amsterdam, Philadelphia: Benjamins.

Patel, Krupa & Hiren B. Patel. 2020. A state-of-the-art survey on recommendation system and prospective extensions. *Computers and Electronics in Agriculture* 178. 105779. https://doi.org/10.1016/j.compag.2020.105779

Pons Bordería, Salvador & Pascual Aliaga E. 2021. Inter-annotator agreement in spoken language annotation: Applying uα-family coefficients to discourse segmentation. *Russian Journal of Linguistics 25*(2). 478–506. https://doi.org/10.22363/2687-0088-2021-25-2-478-506

Riley, Michael D. 1989. Some applications of tree-based modelling to speech and language indexing. In *Proceedings of the DARPA Speech and Natural Language Workshop*. 339–352. San Mateo, CA.

Sahlgren, Magnus. 2008. The Distributional Hypothesis. From context to meaning. In distributional models of the lexicon in linguistics and cognitive science (special issue of the Italian Journal of Linguistics). *Rivista di Linguistica* 20 (1). 33–53.

Sampson, Geoffrey, David Gil & Peter Trudgill. 2009. *Language Complexity as an Evolving Variable.* Oxford linguistics. Oxford, New York: Oxford University Press.

Schmidhuber, Jürgen. 2015. Deep learning in neural networks: An overview. *Neural Networks* 61. 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Sharnagat, Rahul. 2014. *Named Entity Recognition: A Literature Survey*. Center for Indian Language Technology.

Shosted, Ryan K. 2006. Correlating complexity: A typological approach. *Linguistic Typology* 10 (1). 1–40.

Sigdel, Bijay, Gongqi Lin, Yuan Miao & Khandakar Ahmed. 2020. Testing QA systems' ability in processing synonym commonsense knowledge. *IEEE* [Special issue]. *24th International Conference Information Visualisation (IV).* 317–321. https://doi.org/10.1109/IV51561.2020.00059

Solovyev, Valery & Vladimir Ivanov. 2014. Dictionary-based problem phrase extraction from user reviews. In Petr Sojka, Aleš Horák, Ivan Kopeček & Karel Pala (eds.), *Text, speech and dialogue*, 225–232. Springer.

Solovyev, Valery D., Vladimir V. Bochkarev & Svetlana S. Akhtyamova. 2020. Google Books Ngram: Problems of representativeness and data reliability. *Communications in Computer and Information Science* 1223. 147–162. https://doi.org/10.1007/978-3-030-51913-1_10

Su, Xiaoyuan & Taghi M. Khoshgoftaar. 2009. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*. 1–19. https://doi.org/10.1155/2009/421425

Tan, Xu, Tao Qin, Frank Soong & Tie-Yan Liu. 2021. A survey on neural speech synthesis. *arXiv 2106.15561*. URL: https://arxiv.org/pdf/2106.15561.pdf (accessed 20.01.2022).

Tesnière, Lucien. 2015. *Elements of Structural Syntax*. Amsterdam: John Benjamins Publishing Company.
Trudgill, Peter. 1999. Language contact and the function of linguistic gender. *Poznan Studies in Contemporary Linguistics* 35. 133–152.
Trudgill, Peter. 2004. Linguistic and Social Typology: The Austronesian migrations and phoneme inventories. *Linguistic Typology* 8(3). 305–320.
Trudgill, Peter. 2011. *Sociolinguistic Typology: Social Determinants of Linguistic Complexity.* Oxford: Oxford University Press (reprinted 2012).
Trudgill, Peter. 2012. On the sociolinguistic typology of linguistic complexity loss. In Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts & Paul Trilsbeek (eds.), *Language documentation & conservation special publication No. 3 (August 2012): Potentials of language documentation: Methods, analyses, and utilization*, 90–95.
Valdez, Cruz & Monika Louize. 2021. *Voice Authentication Using Python's Machine Learning and IBM Watson Speech to Text*. Universitat Politècnica de Catalunya.
Wang, Yu, Yining Sun, Zuchang Ma, Lisheng Gao, Yang Xu & Ting Sun. 2020. Application of pre-training models in named entity recognition. In *2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC)*. 23–26. Hangzhou, China.

Article history:
Received: 20 October 2021
Accepted: 06 February 2022

Bionotes:

Valery D. SOLOVYEV is Doctor Habil. of Physical and Mathematical Sciences, Professor, Chief Researcher of "Text Analytics" Research Lab, Institute of Philology and Intercultural Communication of Kazan Federal University, Kazan, Russia. He is a member of the Presidium of the Interregional Association for Cognitive Research, author of four monographs and more than 60 publications on the text complexity.

Contact information:
Kazan Federal University
18 Kremlevskaya str., Kazan, 420008, Russia
e-mail: maki.solovyev@mail.ru
ORCID: 0000-0003-4692-2564
Scopus ID: http://www.scopus.com/authid/detail.url?authorId=26665013000

Marina I. SOLNYSHKINA is Doctor Habil. of Philology, Professor of the Department of Theory and Practice of Teaching Foreign Languages, Head and Chief Researcher of "Text Analytics" Research Lab, Institute of Philology and Intercultural Communication of Kazan Federal University, Kazan, Russia. She is the author of over 60 publications on text complexity.

Contact information:
Kazan Federal University
18 Kremlevskaya str., Kazan, 420008, Russia
e-mail: mesoln@yandex.ru
ORCID: 0000-0003-1885-3039

Danielle S. MCNAMARA, Ph.D., is Professor of Psychology in the Psychology Department and Senior Scientist at Arizona State University. She is an international expert in the fields of cognitive science, comprehension, natural language processing, and intelligent systems.

Contact information:
Arizona State University Payne Hall, TEMPE Campus, Suite 108, Mailcode 1104, the USA
e-mail: Danielle.McNamara@asu.edu
ORCID: 0000-0001-5869-1420

Сведения об авторах:
Валерий Дмитриевич СОЛОВЬЕВ – доктор физико-математических наук, профессор, главный научный сотрудник НИЛ «Текстовая аналитика» Института филологии и межкультурной коммуникации Казанского федерального университета, Казань, Россия. Член президиума Межрегиональной ассоциации когнитивных исследований. Автор четырех монографий и более 60 публикаций по сложности текста.

Контактная информация:
Казанский федеральный университет
Россия, 420008, Казань, ул. Кремлевская, д. 18
e-mail: maki.solovyev@mail.ru
ORCID: 0000-0003-4692-2564

Марина Ивановна СОЛНЫШКИНА – доктор филологических наук, профессор, профессор кафедры теории и практики преподавания иностранных языков, заведующий и главный научный сотрудник НИЛ «Текстовая аналитика» Института филологии и межкультурной коммуникации Казанского федерального университета, Казань, Россия. Автор более 60 публикаций по сложности текста.

Контактная информация:
Казанский федеральный университет
Россия, 420008, Казань, ул. Кремлевская, д. 18
e-mail: mesoln@yandex.ru
ORCID: 0000-0003-1885-3039

Даниэль С. МАКНАМАРА – доктор наук, профессор кафедры психологии Университета штата Аризона, психолингвист, международный эксперт в области когнитивистики, понимания, обработки естественного языка и интеллектуальных систем.

Контактная информация:
Arizona State University Payne Hall, TEMPE Campus, Suite 108, Mailcode 1104, the USA
e-mail: Danielle.McNamara@asu.edu
ORCID: 0000-0001-5869-1420