Summer 2020

Automatic Keyphrase Extraction From Russian-Language Scholarly Papers in Computational Linguistics

Yves Wienecke
Portland State University

Automatic Keyphrase Extraction from Russian-Language Scholarly Papers in Computational Linguistics

by

Yves Wienecke

An undergraduate honors thesis submitted in partial fulfillment of the

requirements for the degree of

Bachelor of Science

In

University Honors

And

Computer Science

And

Russian

Thesis Advisor

William Comer, Ph.D

Portland State University

2020

AUTOMATIC KEYPHRASE EXTRACTION FROM RUSSIAN-LANGUAGE SCHOLARLY PAPERS IN COMPUTATIONAL LINGUISTICS

Yves Wienecke[1]

[1] Portland State University, Portland OR 97207, USA
whyves@icloud.com

Abstract. The automatic extraction of keyphrases from scholarly papers is a necessary step for many Natural Language Processing (NLP) tasks, including text retrieval, machine translation, and text summarization. However, due to the different grammatical and semantic intricacies of languages, this is a highly language-dependent task. Many free and open source implementations of state-of-the-art keyphrase extraction techniques exist, but they are not adapted for processing Russian text. Furthermore, the multi-linguistic character of scholarly papers in the field of Russian computational linguistics and NLP introduces additional complexity to keyphrase extraction. This paper describes a free and open source program as a proof of concept for a topic-clustering approach to the automatic extraction of keyphrases from the largest conference on Russian computational linguistics and intellectual technologies, Dialogue. The goal of this paper is to use LDA and pyLDAvis to discover the latent topics of the Dialogue conference and to extract the salient keyphrases used by the research community. The conclusion points to needed improvements to techniques for PDF text extraction, morphological normalization, and candidate keyphrase ranking.

Keywords: Automatic Keyphrase Extraction, Topic Modeling, LDA, pyLDAvis, Scholarly Papers, Russian.

1 Introduction

A challenge which is familiar to all aspiring researchers is the task of understanding the research areas and technical terminology of a scientific community. In addition to comprehending the current research directions of the community, it is important to develop a familiarity with the existing literature of a particular research area. This literature forms a foundation for understanding the state-of-the-art methods and participating in the community's discourse. The natural evolution of research over time and the varying ways of naming and defining terminology further complicate this process, because there exist multiple names for the same concept. For example, researchers have used different terms to refer to "keyphrase" throughout the years, including: "term" [1], [2], "indexing term" [1], [3], "domain-oriented multi-word term" [4], "technical term" [5], "keyword" (ключевое слово) [6], [7], [8], [9], and "keyphrase" [10], [11], [12], [13], [14], [15].

Furthermore, the naming conventions for a scientific community may differ depending on the language in which the literature is published. These complexities increase the amount of work and knowledge necessary for aspiring researchers to participate in a discourse, let alone in a cross-linguistic discourse. The aim of this paper is to apply a topic modeling approach to the automatic extraction of keyphrases in order to explore the research areas and technical terminology of a cross-linguistic research community, in particular the international "Dialogue" conference on computational linguistics and intellectual technologies[1].

The Dialogue conference, which has hosted annual gatherings since 2000, is the largest and oldest conference for Russian computational linguistics. In addition to the Dialogue conference, there are two other major conferences in this field: AINL[2] and AIST[3]. Previous research into the topical structure of these three venues shows that the AINL and AIST conferences are more centered towards computer science, whereas the Dialogue conference is more oriented towards linguistics [16]. The articles from Dialogue are available online and can be downloaded with a web scraper, but the articles from AINL and AIST are only available through the Springer library and must be manually downloaded. Also, AINL and AIST articles are required to be written in English, while Dialogue articles may be written in English or Russian [16]. According to the Dialogue website, "Scopus requires that all papers on computational linguistics be submitted in English," and articles which "require knowledge of Russian are to be submitted in Russian and should include an extended summary in English" [17]. The website explains that the shift towards publishing articles in English is for attracting a wider audience of specialists. Due to the accessibility limitations and lack of Russian-language papers from the AINL and AIST conferences, this paper focuses on papers from the Dialogue conference. Furthermore, this paper only considers papers written in Russian, because the techniques for processing a text and extracting keyphrases is largely a language-dependent process. The code for this paper is free and open source and may be accessed through GitHub[4].

2 Literature Review

2.1 Keyphrases

Some scholarly papers include a list of keywords that are relevant to the important ideas and terminology of the research. These lists include not only key words, but also key collocations, which leads to the more encompassing term "keyphrase." Keyphrases should ideally identify the most relevant and novel concepts of the research. In effect, keyphrases capture the essence of the research, forming a succinct summary of its semantic content [6]. The creation of a list of keyphrases has been compared to the creation of a "domain vocabulary" [4] or a "back-of-the-book index" [6]. Semantic-rich

[1] http://www.dialog-21.ru/
[2] https://ainlconf.ru/
[3] https://aistconf.org/
[4] https://github.com/iyves/auto-keyphrase-extraction-ru

keyphrases are versatile in the field of Natural Language Processing (NLP) for various Information Retrieval (IR) tasks, such as "text summarization, text categorization, opinion mining, document indexing" [10]. However, only a few scholarly papers include a list of keywords and the creation of such a list is a subjective process that is influenced by the biases of the authors, professional indexers[5], and readers[6] [18], [19]. Despite the importance and versatility of keyphrases, automatic keyphrase extraction is an area of research that performs worse than other areas of NLP [10]. The automatic extraction of keyphrases from scholarly papers remains an active area of research [7], [18], [20].

2.2 Automatic Keyphrase Extraction

The process of automatic keyphrase extraction typically consists of two steps: the extraction of candidate keyphrases and the ranking of candidate keyphrase [10], [21]. These steps are further broken down into sequential steps, forming a pipeline that incrementally processes raw textual data in a corpus into a list of keyphrases [6], [8]. The earlier steps are important for the overall performance of keyphrase extraction because errors from these steps will propagate to the later steps, affecting the overall quality of the keyphrases and runtime performance.

Preprocessing and Candidate Keyphrase Selection. The first step of automatic keyphrase extraction is the preprocessing of text and selection of candidate keyphrases. Documents are broken down into keyphrase-sized chunks known as tokens, and tokens are filtered based on a set of rules for determining whether the token is a candidate keyphrase. Candidate keyphrase tokens may be defined as noun phrases, acronyms, sequences of a number (*n)* of adjacent words (n-grams), and/or a combination of words that match a set of parts-of-speech (POS) patterns [9], [13]. Redundancy control mechanisms can be used to remove a candidate keyphrase if it is included in a higher-ranked keyphrase [5], [10]. Frequency thresholding can remove candidate keyphrases based on how frequently the candidates occur in the corpus [5]. It is typical in NLP to remove stop words, which are language-dependent and domain-dependent words that frequently appear in the corpus but are not keywords, such as "the," "a," and "they." Research in improving the performance of automatic keyphrase extraction often neglects the importance of document preprocessing, but effectiveness of preprocessing has a significant influence on the performance of keyphrase extraction techniques [11].

Scholarly papers are most commonly stored in a PDF format. Before preprocessing the papers, the text must be extracted from these PDF files. However, the automatic extraction of plain text from scholarly papers in PDF format is a notoriously difficult task [22]; special attention is required to deal with malformed text from corrupted files and irrelevant text from tables, equations, and footnotes [11]. In instances when the text cannot be extracted from the PDF files, i.e., when the files are scanned images, optical character resolution is necessary. Boudin notes that the techniques for dealing with extracting only the relevant text from scholarly PDF files range from "simple heuristics

[5] https://www.ebsco.com/products/research-databases/inspec
[6] http://semeval2.fbk.eu/semeval2.php?location=tasks&taskid=6

to sophisticated document logical structure detection" [11]. After text is extracted from a PDF file and the corrupted and irrelevant text is removed, it can be preprocessed and transformed into candidate keyphrases tokens. The steps used for preparing text for NLP depends on a number of factors, including the language, formality, and structure of the text. For example, the techniques used for preprocessing Russian-language scholarly papers differ from those used for English-language tweets. In contrast to tweets, scholarly papers have considerable length, which greatly increases the amount of candidate keyphrases that must be ranked. However, scholarly papers often have a fairly consistent document structure, predictable topic change (keyphrases are more prevalent at the beginning and end of papers), and keyphrases are usually correlated [10]. These features of scholarly papers can be used to reduce the length of the text and the amount of candidate keyphrases. This reduction improves the overall performance and quality of keyphrase extraction.

In addition to considering the characteristics of scholarly papers, there are other considerations that are dependent on the language of the papers. Most research for the automatic extraction of keyphrases is adapted towards processing English text [6], [23]. However, the linguistic features of a language may require different techniques for optimal performance [23], [24]. For example, compound-rich European languages like Dutch and German contain long compound words that can be broken down through decompounding strategies [11]. English is an analytic language, i.e., syntactic relationships between words are primarily expressed through word order and helper words. Russian, however, is a language with a rich morphology [7], [25], which places less importance on word order and greater importance on the inflection of words. The spelling of a word in Russian may change depending its grammatical role (noun, adjective, verb) and its syntactical relationship to other words (case, gender, number) [23]. A study of the noun case usage in a beginning-level Russian textbook shows that, even in a corpus which overrepresents three of twelve case forms, the average case frequency of nouns is approximately 8.3% and the highest frequency of any one noun case form does not exceed 27% [26]. This means that the frequency of a single word is thinly distributed among its various case forms, resulting in a large number of tokens with relatively small frequencies. The morphological quality of Russian leads to regular homonymy in the language; the sentence "его решение задачи было неправильным" has over 100 variants of POS tagging and morphological labeling describing the grammatical role, number, gender, and case for each word [25]. Morphological normalization techniques like stemming and lemmatization deal with the inflectional variation of words by reducing words to a common stem form or dictionary form. On the one hand, stemming uses algorithms to remove the suffixes and/or prefixes of a word, producing stems that are not always real words. Stemming does not typically require access to external resources. For example, the collocation "ключевые слова извлеклись" would produce the stemmed counterpart "ключев слов извле" [8]. On the other hand, lemmatization uses external resources and the POS meta-information about words in a sentence to reduce inflected words to their dictionary equivalents. For example, the collocation "ключевые слова извлеклись" may lemmatize into "ключевой слово извлекаться" [8]. Lemmatization is shown to be beneficial for dealing with the preprocessing of morphologically rich languages such as Russian [23].

Aside from the grammatical differences between languages, there may be orthographical differences. For English and Russian, the most obvious difference is the alphabet in which each language is written; Russian text is written in the Cyrillic alphabet, while English is written in the Latin alphabet. Furthermore, Russian text may include punctuation that does not exist for English text, such as the guillemets («») and the reversed double quotes („“). The em dash (—) is used differently in Russian than in English; it is often used for direct quotes or as a replacement of the verb “быть.” These punctuation differences mean that the patterns and techniques used for breaking down text into tokens can differ between English and Russian. In some Russian texts, the letter ‘e’ is used rather than the letter ‘ë,’ but this can be controlled by replacing all occurrences of ‘ë’ with ‘e’ in the text. Also, while not relevant to this paper, Russian text may exhibit spelling and orthographic variations throughout time as a result of language reforms. It is important to consider the orthographic features of a language in order to produce correct candidate keyphrase tokens. Incorrect tokens may impact the quality of morphological normalization, which then impacts the quality of keyphrase extraction.

Candidate Keyphrase Ranking. In the second step, candidate keyphrases are ranked and filtered to remove unlikely keyphrases. Approaches in this step fall into the categories of supervised or unsupervised approaches. Both supervised and unsupervised approaches may involve judging a candidate keyphrase on a combination of extracted features. The taxonomy of extractable features is vast and will not be discussed in detail in this paper [6], [10]. Some examples include heuristical, statistical, and external features [7], [9]. Heuristical features may take into consideration the section of paper in which a candidate keyphrase appears [11], statistical features may involve calculating the term frequency–inverse document frequency (TF—IDF) of a candidate [3], [4], and external features depend on external tools like a thesaurus or Wikipedia [4]. Earlier supervised approaches train a machine learning model to classify candidates as keyphrases or not. Later approaches involve training a machine learning model to rank candidates. While supervised models are more resilient to noise in the data [11], an annotated corpus with a list of gold-standard keyphrases is often required in order to train the machine learning model.

In contrast to supervised approaches, unsupervised approaches do not require gold-standard keyphrases. Unsupervised approaches to ranking keyphrases differ substantially from supervised approaches and can be generally categorized into graph-based ranking, topic-based clustering, simultaneous learning, and language modeling approaches [10]. Graph-based ranking is the most prevalent approach [5], [9], [12], [13]. Graph-based ranking derives from Google’s PageRank algorithm for ranking websites [27]. A popular graph-based approach to keyphrase extraction is TextRank [9], [10], which is inspired by the PageRank algorithm. The intuition behind graph-based ranking approaches is similar to the intuition behind calculating popularity in a social network; candidate keyphrases recommend one another based on a number of features, such as co-occurrence relation [9]. The candidates with the highest recommendation scores are the most likely keyphrases, and the recommendations from these candidates have a greater impact on the scores of other candidates. Although graph-based approaches are

more prevalent for automatic keyphrase extraction, this paper focuses on a simplistic topic-clustering approach. Topic-clustering approaches describe the corpus by discovering its latent topics and the associated keyphrases which characterize each topic. These approaches assume that the corpus is a combination of some number of latent topics and that the ideal keyphrases are those that appear in multiple, main topics of the documents [10].

This paper takes inspiration from sophisticated topic-clustering methods that include complex keyphrase ranking techniques, but instead, uses a simplistic approach of keyphrase extraction based purely on topic modeling. Topic modeling techniques involve matrix decomposition algorithms or probabilistic machine learning models to discover the latent topics that describe the semantic space of a document, as well as the candidate keyphrases which characterize each topic. Under topic modeling, documents in a corpus are typically represented as Bags-of-Words (BoW), where each document is an unordered collection of tokens. This allows for representing the entire corpus as a matrix, where the rows are documents, the columns are tokens in the vocabulary, and the cells are frequency counts for each token in each document. The first topic modeling technique is Latent Semantic Indexing/Analysis (LSI/A) and involves weighing a candidate keyphrase by its TF—IDF score instead of its frequency count, and then using Singular Value Decomposition (SVD) or Non-Negative Matrix Factorization (NMF) to factor out the topics and associated keyphrases from the corpus [28], [29]. Hofmann takes a probabilistic approach to topic modeling with the introduction of probabilistic LSI/A (pLSI/A), which is a generative probabilistic approach that learns topic distributions and word distributions for a corpus [2]. As a generative model, pLSA assumes that a document in a corpus is a finite mixture over a pre-defined number of latent topics, with each topic being an infinite mixture over the vocabulary of the corpus [5]. In summary, pLSA assumes that a document is created by repeatedly selecting a topic associated with the document and writing down a candidate keyphrase associated with the selected topic [29]. A shortcoming of the pLSA model is that it fails to probabilistically calculate the topic distribution for each document in a corpus, which means that each document is associated with a hidden parameter and, as a consequence, there is no clear way to calculate the topic distribution of a document outside of the training dataset. Latent Dirichlet Allocation (LDA) addresses the limitations of pLSA by introducing a Dirichlet prior on the document-topic distributions. This makes LDA generalizable to unseen documents and more computationally efficient than pLSA [30]. A well-known topic-clustering approach that uses LDA, Topical PageRank (TPR) [5], is a sophisticated technique that uses TextRank on each of the topics learned by LDA in order to select the main keyphrases for a document.

2.3 Open Source Tools for Automatic Keyphrase Extraction

The creation of open source toolkits and resource repositories has grown in the recent two decades as a response to the increasing complexity of implementing state-of-the-art NLP algorithms, the resources needed for training sophisticated machine learning models, and the need for reproducing experimental results [31]. For example, the

Natural Language Toolkit[7] (nltk) is a popular open source Python library that contains methods for preprocessing text in various languages. Similarly, Boudin created an "open source Python-based keyphrase extraction toolkit" (pke), which provides an implementation of an automatic keyphrase extraction pipeline [32]. This toolkit allows for the rapid prototyping and evaluation of keyphrases extracted from state-of-the-art supervised and unsupervised approaches. However, this toolkit is not currently adapted towards Russian text and therefore can not be used for this paper.

As an alternative to pke, this paper uses an open source Python library for topic modeling, Gensim[8]. Gensim implements LSA and LDA and has been used in a previous analysis of the topical structure of English-language papers from Dialogue, AINL, and AIST [16]. In a paper [33], these researchers develop an open source toolkit, WebVectors[9], which includes code for preprocessing Russian text. This paper uses the preprocessing techniques of WebVectors alongside the Gensim implementation of LDA to process and extract candidate keyphrases from Russian-language Dialogue conference papers. Although this topic-modeling based approach to automatic keyphrase extraction is not representative of state-of-the-art keyphrase extraction techniques, it can be improved upon in future research and potentially expanded to an implementation of TPR. To encourage future research in the automatic extraction of keyphrases from Russian-language scholarly papers, the code used for this paper is free and open source.

2.4 Automatic Keyphrase Extraction with Gensim and pyLDAvis

Before using the Gensim implementation of LDA, the corpus must first be transformed into a BoW representation. Typically, a Gensim dictionary is created, mapping each candidate keyphrase token in the vocabulary to a unique token identifier. Gensim allows for frequency thresholding with the *no_above* and *no_below* hyperparameters. The *no_above* hyperparameter is a float that controls the upper threshold for the maximum percentage of documents in the corpus in which a candidate keyphrase may appear to be considered. The *no_below* hyperparameter is a lower threshold, which is an integer that controls the minimum number of documents in which a candidate keyphrase must appear to be considered. The dictionary is then used to build the BoW representation of the corpus. A BoW corpus is a sparse vector that contains the frequency counts of candidate keyphrase token ids for each document and will be used to train the LDA model. The number of latent topics that the LDA model learns, *num_topics*, is another hyperparameter that must be optimized for a particular corpus. After training, the Gensim library provides methods for displaying the main topics discovered by a trained LDA model and the most probable candidate keyphrases which characterize each latent topic. Additionally, a Python library for generating interactive topic modeling visualizations, pyLDAvis[10], can be used to interactively explore the LDA

[7] https://www.nltk.org/
[8] https://radimrehurek.com/gensim/
[9] https://github.com/akutuzov/webvectors
[10] https://github.com/bmabey/pyLDAvis

topics and rank the candidate keyphrases. This library is a Python port for the LDAvis R package[11].

Hyperparameter optimization is an important step for controlling the candidate keyphrases selected by the LDA model. The quality of a topic model after fine-tuning hyperparameters can be judged by extrinsic or intrinsic measures. Extrinsic measures involve indirectly measuring the quality of the topics by evaluating the performance of the topic model on other NLP tasks, such as word sense disambiguation or document retrieval. Intrinsic measures are more convenient for this research, as they do not involve other NLP tasks; intrinsic measures are statistical and semantic measures of topic model quality. The creators of LDA propose a statistical measure, perplexity or held-out likelihood [30], which captures the degree to which a topic model is a comprehensive representation of the semantic space of the corpus. Perplexity is used to determine the optimal number of latent topics for a model to learn [16], [34]. However, perplexity has been shown to correlate poorly and sometimes negatively with generating keyphrases that characterize human-interpretable topics [35]. This is particularly problematic for the automatic extraction of keyphrases, because LDA tends to produce topics which are less specific than the matrix factorization techniques, which further makes the topics more difficult for humans to interpret [29], [35].

In response to the drawbacks of perplexity, coherence metrics have been proposed for measuring the semantic quality of topics [35], [36]. Whereas perplexity measures the comprehensiveness or generality of a topic, coherence captures how well candidate keyphrases describe a topic. Coherence metrics typically involve a calculation of either keyphrase co-occurrence for a topic in a corpus or the distance between candidate keyphrases vectors in semantic space [29]. The Gensim library exposes methods for evaluating topic models on perplexity and an array of coherence metrics. In addition to the log perplexity calculated by Gensim, this paper considers the Cv coherence score of LDA topic models. Cv coherence is an improvement on earlier coherence metrics, such as the one proposed by Mimno [35] and the normalized Pointwise Mutual Information (NPMI) coherence proposed by Bouma [37]. A study of various coherence metics found that the Cv coherence metric was the best predictor for human interpretability [38]. This paper uses perplexity for the optimization of the *num_topics* hyperparameter and Cv coherence for the optimization of the *no_below* and *no_above* hyperparameters.

After optimizing the hyperparameters and using LDA to select candidate keyphrases, the pyLDAvis tool is used to calculate the marginal topical distribution of the topics learned by LDA and rank the candidate keyphrases which characterize each topic. The marginal topical distribution can be understood as the size of a topic and is used in this paper to describe a topic's distinctiveness or specificity. LDAvis introduces a relevancy metric for ranking the terms associated with a topic. This metric is a combination of a term's lift and topic-specific probability. A term's lift is the ratio of the term's "probability within a topic to its marginal probability across the corpus" [39], and has a similar function as TF—IDF in reducing the ranking of terms which are frequent in the entire corpus. The weight hyperparameter (λ) for the relevancy metric controls the extent to which a term's topic-specific probability affects its

[11] https://github.com/cpsievert/LDAvis

ranking. Gensim's methods for generating keyphrases is purely based on topic probability, which is equivalent to setting the weight parameter to 1. Conversely, setting the weight parameter to 0 in pyLDAvis considers terms solely on lift, which can be noisy, as it raises the ranking of rare terms that may only appear in a single document [39]. The optimal λ value for producing the most human-interpretable topics is 0.6 [39]. This paper ranks keyphrases both by lift and human-interpretability in order to better understand the topics in the Dialogue conference, as well as some specific keyphrases associated with each topic.

3 Research Question

What are the salient keyphrases for Russian papers published in the Dialogue journal?

4 Methods

The code for this paper consists of a pipeline with six steps: the creation of the corpus, preprocessing, hyperparameter optimization, candidate keyphrase selection, keyphrase ranking, and post-processing. The modular quality of this pipeline allows for improving or changing a single stage without requiring modifications in the other stages. The pipeline extracts all conference papers and online articles from the Dialogue website, partitions papers by language, preprocesses each paper, performs topic modeling on the corpus, and generates a list of topics and keyphrases.

4.1 Corpus Creation

The first step in the pipeline involves scraping the Dialogue website and downloading all articles published in the online digest. Articles from the proceedings begin from 2000 and continue annually until the year of this paper, 2020. These articles are primarily accessible in a PDF format, but some older articles are available only in text format. The articles in PDF format are directly downloadable from the article's URL, but the articles in text format require extra web scraping in order to isolate and extract only the relevant text of the article from the Dialogue website.

For a given range of years, all published articles from a conference and all online articles are downloaded locally. Each article is accessed through the article URL and saved as a file with the article title as the filename. To avoid conflict with filesystem rules for naming files, article titles are transformed into URL slugs through the Django slugify method[12] and any titles longer than 100 characters are truncated.

All articles from the dialogue conference are written in either English or Russian. This paper focuses only on the articles written in Russian, so articles must be partitioned by language. First, text from articles in PDF format are extracted through the pdftotext command line tool[13]. Only text present between the first 80 pixels and 650 pixels are

[12] https://docs.djangoproject.com/en/3.0/_modules/django/utils/text/#slugify
[13] https://www.xpdfreader.com/pdftotext-man.html

extracted from the article, which removes extraneous text from the header and footer of the document. This is a simplistic heuristic which is not representative of state-of-the-art document logical structure detection techniques for removing irrelevant text from tables and examples. Next, the text from each article is passed through the langid language identification tool[14] to classify the document as an English-language or Russian-language document. Finally, the extracted text is saved into a separate target folder according to the language classification. **Fig. 1** shows the language distribution of dialogue articles throughout the years. There is a notable shift in the language distribution in 2011, which may be a reflection of Scopus requirements for papers in the field of computational linguistics to be written in English [17]. This language distribution is very similar to the language distribution visualized in the previous research on the topical structure of Dialogue, AIST, and AINL [16]. It is notable to mention that, as a consequence of the simplistic heuristic used to extract text from the PDFs, some papers written in English are misclassified as Russian-language papers if they contain a significant amount of Russian-language examples.

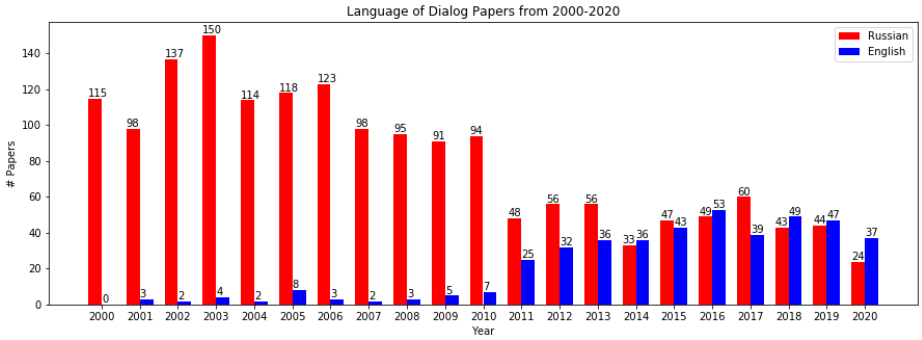

Fig. 1. The language distribution of articles published from the Dialogue conference.

4.2 Preprocessing

The preprocessing step expects the corpus to be represented as a list of text files, with each file consisting of the raw text extracted from a single Dialogue article. After preprocessing, the corpus will be represented as a list of document tokens, where each document in the corpus is a list of lowercase, lemmatized candidate keyphrase tokens. The document tokens for each conference year is saved in a binary format as a pickle file, which allows for loading the preprocessed corpus at a later time in Python. The pickle format preserves the data structure and character set encoding of the corpus.

First, the references section is removed from each document. This paper uses simple heuristics to identify and remove the references section and irrelevant text. For this paper, the references section is defined as any text following the final line which contains ‘Литература’ or ‘Список литературы’. Any paragraphs that are shorter than 100 characters are assumed to be unusual for a body paragraph and are removed. While the

[14] https://github.com/saffsd/langid.py

pdftotext tool automatically removes hyphenation for words split at newline boundaries, it does not remove hyphenation from words split at new page boundaries. This paper fills the gap by removing any words at the end of a paragraph and ending in a dash, and any words at the beginning of a paragraph that does not begin with a capital letter. In-text citations for the Dialogue journal are typically text inside either parentheses or brackets. All parenthesized and bracketed text are removed to reduce noise from this irrelevant information. **Fig. 2** shows the text before and after these transformations.

```
Raw text, before any preprocessing:


7. Заключение
В данной работе представлен опыт создания
и развития лингво-семантических представлений
в интеллектуальных информационных системах,
разработанных на основе аппарата расширенных
семантических сетей (РСС). Аппарат РСС обеспечивает мощные изобразительные возможности для
описания всех уровней естественного языка, включая уровень глубинно-семантических представлений, и межъязыко
вых соответствий. Конкретные

After removing the references section, short paragraphs, parenthesized/bracketed text, and text at pagebreak
s:


7. Заключение В данной работе представлен опыт создания и развития лингво-семантических представлений в инте
ллектуальных информационных системах, разработанных на основе аппарата расширенных семантических сетей . Апп
арат РСС обеспечивает мощные изобразительные возможности для описания всех уровней естественного языка, вклю
чая уровень глубинно-семантических представлений, и межъязыковых соответствий. Конкретные лингвистические пр
```

Fig. 2. Preprocessing the corpus; removing the references section, short paragraphs, in-text citations, and text at page delimiters.

Next, documents are tokenized and preprocessed using the methods developed by Web-Vectores. Each document is split into paragraphs and passed through the *rus_preprocessing_udpipe* processing method, resulting in a representation of each document as a list of paragraph tokens. This method first passes the text through a UDPipe model to transform the text to a CoNLL-U format. Words and punctuation are isolated as lowercase, lemmatized tokens. Punctuation and newline tokens are kept to preserve information about sentence and paragraph boundaries. Any tokens which are not a sentence delimiter, paragraph delimiter, and do not have at least one Cyrillic letter are removed. These tokens are primarily punctuation marks or English words. **Fig. 3** shows the result after conducting these preprocessing steps.

```
After tokenization:


['x', '.', 'заключение', 'в', 'данный', 'работа', 'представлять', 'опыт', 'создание', 'и', 'развитие', 'линг
во-семантический', 'представление', 'в', 'интеллектуальный', 'информационный', 'система', ',', 'разрабатыват
ь', 'на', 'основа', 'аппарат', 'расширить', 'семантический', 'сеть', '.', 'аппарат::рсс', 'обеспечивать', 'м
ощный', 'изобразительный', 'возможность', 'для', 'описание', 'весь', 'уровень', 'естественный', 'язык', ',',
'включая', 'уровень', 'глубинный', '-', 'семантический', 'представление', ',', 'и', 'межъязыковый', 'соответ
ствие', '.', 'конкретный', 'лингвистический', 'процессор', ',', 'который', 'быть', 'давать', 'на', 'основа',
After removing tokens which do not have at least 1 alphabetical character and are not sentence delimiters:


['.', 'заключение', 'в', 'данный', 'работа', 'представлять', 'опыт', 'создание', 'и', 'развитие', 'лингво-семантический',
'представление', 'в', 'интеллектуальный', 'информационный', 'система', 'разрабатывать', 'на', 'основа', 'аппарат', 'расши
рить', 'семантический', 'сеть', '.', 'обеспечивать', 'мощный', 'изобразительный', 'возможность', 'для', 'описание', 'весь
', 'уровень', 'естественный', 'язык', 'включая', 'уровень', 'глубинный', 'семантический', 'представление', 'и', 'межъязык
овый', 'соответствие', '.', 'конкретный', 'лингвистический', 'процессор', 'который', 'быть', 'давать', 'на', 'основа', 'э
```

Fig. 3. Preprocessing the corpus; tokenization, lemmatization, filtering punctuation, and filtering non-Cyrillic tokens.

Bigrams and trigrams are added to the document tokens through the Gensim collocation detection class[15]. Bigrams that appear at least three times in a document are considered, as well as all trigrams. Stop words do not contribute to the word count of n-grams; bigrams and trigrams which contain a stop word have more than the expected amount of tokens. Then, numbers and tokens shorter than three characters or longer than 100 characters are removed. This filters out punctuation and any unusually long tokens. **Fig. 4** shows the result after these steps. Finally, all stop words are removed from the corpus and the list of paragraph tokens is flattened out into a list of document tokens. At the end of preprocessing, each document in the corpus is represented as a list of candidate keyphrase tokens. **Fig. 5** is the state of one document from the corpus after preprocessing. The preprocessed text for articles from each year are saved into a binary pickle format, allowing Python to load the text as input for the following step. It is noted that some PDF files are in an unknown format, which causes the pdftotext tool to extract malformed text. While it would be best to remove documents with malformed text or perform OCR to correctly extract the text, this paper removes only malformed tokens and keeps valid tokens. **Fig. 6** shows the input and output after preprocessing two sentences with malformed text.

```
After adding bigrams and trigrams:


['заключение', 'в', 'данный', 'работа', 'представлять', 'опыт', 'создание', 'и', 'развитие', 'лингво-семантический
', 'представление', 'в', 'интеллектуальный', 'информационный', 'система', 'разрабатывать', 'на', 'основа', 'аппарат
', 'расширить', 'семантический', 'сеть']
['обеспечивать', 'мощный', 'изобразительный', 'возможность', 'для', 'описание', 'весь', 'уровень', 'естественный яз
ык', 'включая', 'уровень', 'глубинный', 'семантический', 'представление', 'и', 'межъязыковый', 'соответствие']

After filtering short tokens, long tokens, and stop words:


['данный', 'представлять', 'опыт', 'создание', 'развитие', 'лингво-семантический', 'представление', 'интеллектуальн
ый', 'информационный', 'разрабатывать', 'основа', 'аппарат', 'расширить', 'семантический', 'сеть']
['обеспечивать', 'мощный', 'изобразительный', 'возможность', 'описание', 'уровень', 'естественный язык', 'включая',
'уровень', 'глубинный', 'семантический', 'представление', 'межъязыковый', 'соответствие']
```

Fig. 4. Preprocessing the corpus; adding ngrams and removing stop words.

```
Final preprocessed text:

 ['данный', 'представлять', 'опыт', 'создание', 'развитие', 'лингво-семантический', 'представление', 'интеллектуаль
ный', 'информационный', 'разрабатывать', 'основа', 'аппарат', 'расширить', 'семантический', 'сеть', 'обеспечивать',
'мощный', 'изобразительный', 'возможность', 'описание', 'уровень', 'естественный язык', 'включая', 'уровень', 'глуб
инный', 'семантический', 'представление', 'межъязыковый', 'соответствие', 'конкретный', 'лингвистический', 'процесс
```

Fig. 5. Preprocessing the corpus; final preprocessed text.

```
=
4. Заключение
Предлагаемый метод позволяет в единообразной форме описывать=
этимологию широкого множества слов различных языков с помощью=
отдельных информационных модулей=-=многократно используемых=
шаблонов.=При этом устраняется дублирование хранимой информацииI=
сокращается вероятность возникновения ошибокI=а для исправления=
замеченных ошибок достаточно однократного редактирования шаблона=-=
исправленная информация автоматически появляется во всех статьяхI=где=
этот шаблон используется.==
=
', 'возможность', 'помечаться', 'соответствующий', 'предлагать метод', 'позволять', 'единообразный', 'этимология
', 'широкий', 'множество', 'различный', 'язык', 'отдельный', 'информационный', 'устранять', 'дублирование', 'хра
нить', 'сокращаться', 'вероятность', 'возникновение', 'заметить', 'ошибка', 'достаточно', 'однократный', 'редакт
ирование', 'исправить', 'информация', 'автоматически', 'появляться']
```

[15] https://radimrehurek.com/gensim/models/phrases.html

Fig. 6. Preprocessing the corpus; poorly parsed .pdf file.

4.3 Hyperparameter Optimization

After preprocessing the corpus, topic modeling may be conducted to discover the latent topics of the corpus and the candidate keyphrases which characterize each topic. To select the best topics, there are a few hyperparameters which should be optimized, as these parameters have a dramatic impact on the quality of the LDA modes. LDA models are trained on the entire corpus using varying hyperparameter combinations to find the optimal hyperparameter values. The resulting LDA models are evaluated based on the log perplexity and Cv coherence scores calculated by the Gensim library. LDA models with a better perplexity have log perplexity scores that are closer to zero. This indicates that the topics learned by these models are more representative of the entire corpus, rather than a few documents in the corpus. Models with a higher Cv coherence score indicate a better coherence. Better coherence scores correlate with generating candidate keyphrases that characterize the topics in a manner that is most conducive to human interpretation.

The first hyperparameter to optimize is *num_topics*, the number of latent topics for the LDA model to learn. A larger number of topics allows for the discovery of more topics which are likely more specific, but too many topics may produce results which are less interpretable for humans. Perplexity is often used for the optimization of *num_topics*, but perplexity may be considered alongside the Cv coherence score to get a different perspective for judging the semantic quality of the keyphrases which characterize each latent topic. A fixed value of 0.75 for *no_above* and 1.0 for *no_below* was arbitrarily chosen in order to focus on the impact of varying *num_topics*. As seen in **Fig. 7**, the perplexity decreases at a steady rate of about 0.2 as *num_topics* increases from six to nine topics before rapidly decreasing. This suggests that the optimal *num_topics* falls somewhere under nine latent topics. The coherence scores sharply drop between *num_topics* of five and six, which suggests that the optimal *num_topics* is five latent topics. In **Fig. 8**, three of the topics generated from nine *num_topics* are identical. This topic, which is also present for 6 *num_topics*, contains a rare, outdated keyphrase, “речевать,” and a phrase fragment that is a very unlikely keyphrase. As a result, it is difficult to understand how each of the keyphrases are related. The third topic for five *num_topics* also has some suspicious keyphrases, but the coherence scores suggest that these keyphrases are more correlated than the keyphrases for the other topic. Therefore, this paper uses five *num_topics* as the optimal number of latent topics.

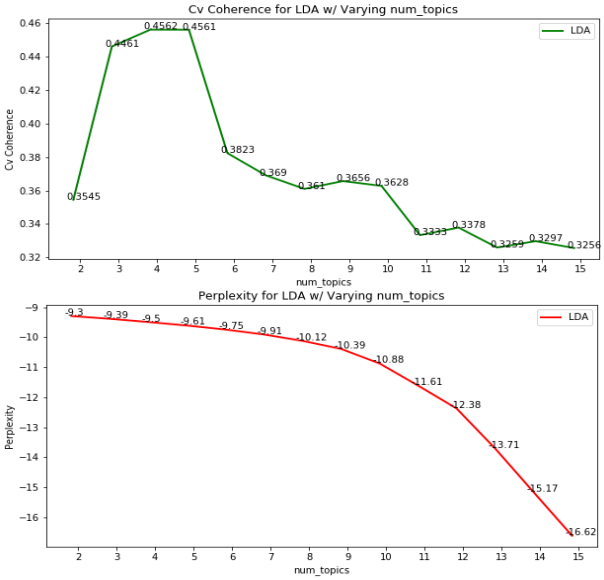

Fig. 7. Optimization for *num_topics; no_above*=0.75 and *no_below*=1; average of 3 repetitions.

Topic #	Top keywords
1	задача, качество, число, информация, разметка
2	клауза, абы, именной, антецедент, предлог
3	глаод, лср, нареч, предлог с винута пункт, обозначение лср
4	век, устный, реплика, дискурсивный, коммуникация
5	выражать, отрицание, семантика, семантический, сочетание

Topic #	Top keywords
1	дискурсивный, реплика, собеседник, устный, коммуникация
2	задача, информация, разметка, использование, метод
3	проступок, речевать, преобразиться, преподавание иностранный язык, приводить к то что студент
4	выражать, семантика, семантический, событие, вводный
5	синтаксический, статья, показывать, оказываться, разный
6	параметр, запрос, отличаться, документ, временный

Topic #	Top keywords
1	выражать, разный, событие, указывать, единица
2	проступок, речевать, преобразиться, преподавание иностранный язык, приводить к то что студент
3	отрицание, семантический, интерпретация, предикат, коннектор
4	проступок, речевать, преобразиться, преподавание иностранный язык, приводить к то что студент
5	число, признак, качество, оценка, показывать
6	разметка, информация, связь, поиск, задача
7	проступок, речевать, преобразиться, преподавание иностранный язык, приводить к то что студент
8	префиксальный, тридцать, словоформый, именительный, грамматический описание
9	параметр, запрос, отличаться, просодический, подкорпус

Fig. 8. Candidate keyphrases for *num_topics* of 5, 6, and 9; highest coherence of 3 repetitions.

Hyperparameter optimization for LDA is usually limited to *num*_topics, but the Gensim library exposes two dictionary-related hyperparameters that can be used for frequency thresholding, *no_below* and *no_above*. The first hyperparameter sets a lower limit on the minimum number documents in which a candidate keyphrase must appear, while the second hyperparameter establishes an upper bound on the maximum percentage of documents in which a candidate may appear. **Fig. 9** displays the coherence and perplexity values for varying combinations of *no_below* and *no_above* with five *num_topics*. Additionally, the candidate keyphrases generated by the three models with the highest coherence for each value of *no_below* is displayed in **Fig. 10Fig. 10**.

Although the coherence score for a *no_below* value of 1 is the highest on multiple instances, the perplexity is consistently significantly worse compared to the other *no_below* values. The perplexity for *no_below* values of 2 and 3 are similar, with the *no_below* value of 3 performing better, but the coherence scores for the *no_below* value of 2 are generally higher. A look at the candidate keyphrases generated by these models shows that the keyphrases generated for the first topic for the *no*_below value of 1 contain non-existent words and a fragment of a sentence that is very unlikely to be a keyphrase. Given the lower perplexity for the *no_below* value of 1 and the unusually high coherence for the *no_above* value of 0.8, these candidate keyphrases are likely very rare terms that frequently occur in a single document. This noise is not present in the keyphrases generated by the other models, which suggests that the optimal *no_below* and *no_above* combination for this corpus is the combination of a *no_below* value of 2 and a *no_above* value of 0.75.

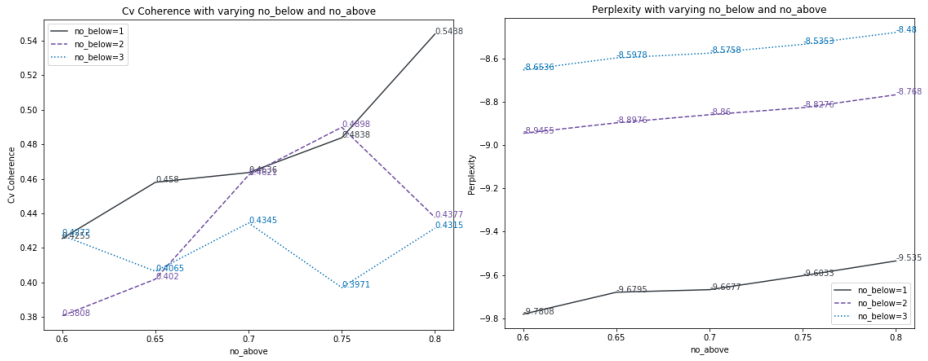

Fig. 9. Optimization for *no_above* and *no_below*; *num_topics*=5, averaged over 3 repetitions.

Topics for num_topics=5; no_below=1; no_above=0.8

	Topic #	Top keywords
0	1	глаод, лср, нареч, обозначение лср, предлог с винута пункт, местинский, лсп
1	2	значение, пример, выражать, отрицание, семантика, событие, вводный
2	3	пауза, диктор, просодический, строка, незавершенность, запись, акцент
3	4	тип, пример, значение, результат, семантический, признак, класс
4	5	разметка, результат, метод, задача, поиск, запрос, алгоритм

Topics for num_topics=5; no_below=2; no_above=0.75

	Topic #	Top keywords
0	1	общий, показывать, разный, приводить, единица, число, выражать
1	2	пауза, запись, диктор, просодический, строка, акцент, незавершенность
2	3	разметка, метод, задача, поиск, информация, запрос, база
3	4	век, реплика, собеседник, дискурсивный, событие, акт, адресат
4	5	семантический, семантика, отрицание, класс, предикат, сочетание, обозначать

Topics for num_topics=5; no_below=3; no_above=0.7

	Topic #	Top keywords
0	1	разметка, число, признак, структура, статья, показывать, правило
1	2	диктор, просодический, акцент, дискурсивный, испытуемый, рема, незавершенность
2	3	метод, запрос, поиск, ресурс, информация, пользователь, задача
3	4	выражать, семантический, семантика, отрицание, сочетание, обозначать, интерпретация
4	5	вводный, событие, указывать, отмечать, собеседник, век, характерный

Fig. 10. Candidate keyphrases for the *no_below* and *no_above* combinations: (1, 0.8), (2, 0.75), and (3, 0.7).

4.4 Topic Modeling and Candidate Ranking

Once the optimal hyperparameter values have been determined, LDA can be run on various sub-corpora to extract and rank candidate keyphrases from the Dialogue papers. Separate LDA models are trained on articles from each year and then on the entire corpus. For analysis of the topics generated from this model, the top seven tokens associated with each latent topic is displayed, as well as the four documents that are most related to these topics. After training the LDA models, the pyLDAvis tool is used to interactively visualize the topics and associated candidate keyphrases. For analysis, λ is set to 0.6 to rank candidates by human-interpretability, and 0 to rank candidates by lift. For the final set of extracted keyphrases, the top five keyphrases with λ set to 0.6 are displayed to describe the topics, and the top ten keyphrases with λ set to 0 are displayed as the salient keyphrases for the corpus. The pyLDAvis tool also calculates the marginal topic distribution of each topic, which gives an idea of how large or general a

topic is; topics with a smaller marginal topic distribution are representative of a smaller percentage of total tokens in the corpus.

4.5 Post-Processing

After the completion of automatic keyphrase extraction, an analysis of the results may call for iterating back to the preprocessing step to further optimize the quality of keyphrases. The stop word list is a language-dependent and domain-dependent list of words that are frequent to the corpus but are not desirable keyphrases. For automatic keyphrase extraction with LDA, it can be beneficial to add dominating keyphrases to the stop word list if these keyphrases commonly reappear in all topics and are ranked highly. For example, it is expected that linguistic terms are extracted from the Dialogue corpus, but certain linguistic terms, scholarly terms and abbreviations may not be particularly interesting keyphrases, for example: "во-первых [=first]," "таблица [=table]," and "лексема [=lexeme]." The process of determining whether a term should be added as a stop word is a subjective process that depends on the goals of the research.

The values of the hyperparameters for Gensim and pyLDAvis may also be modified in the post-processing step. The *num_topics*, *no_below*, *no_above*, and λ hyperparameters control the generality of topics and specificity of keyphrases. Depending on the goals of the research, these hyperparameters can be modified to extract different kinds of keyphrases. Modifying the *num_topics*, *no_below*, *no_above* parameters will impact the topics and candidate keyphrases produced by LDA. Adjusting the λ parameter will rank candidate keyphrases differently and can place more importance on either the lift or the topic probability of a candidate keyphrase.

5 Results

This section displays the results of automatic keyphrase extraction from the Dialog corpus and sub-corpora. The LDA models were trained with five *num_topics*, a *no_below* value of 2, and a *no_above* value of 0.75. Then, pyLDAvis was used to calculate the marginal topical distribution of the learned topics and to rank candidate keyphrases. A λ value of 0.6 and 0 was used to rank candidate keyphrases both in a way that characterizes the topics in the most human-interpretable manner, and in a way that produces specific keyphrases. **Fig. 11** plots the perplexity and coherence of the topics discovered for each sub-corpus by year. The coherence score for each year normally ranges from around 0.33 to 0.4, with an average of approximately 0.3651. Notably, the score for 2013 is unusually high and the score for 2017 is unusually low. The perplexity scores before 2011 average at around -7.90, while the scores from 2011 and later average at -7.69.

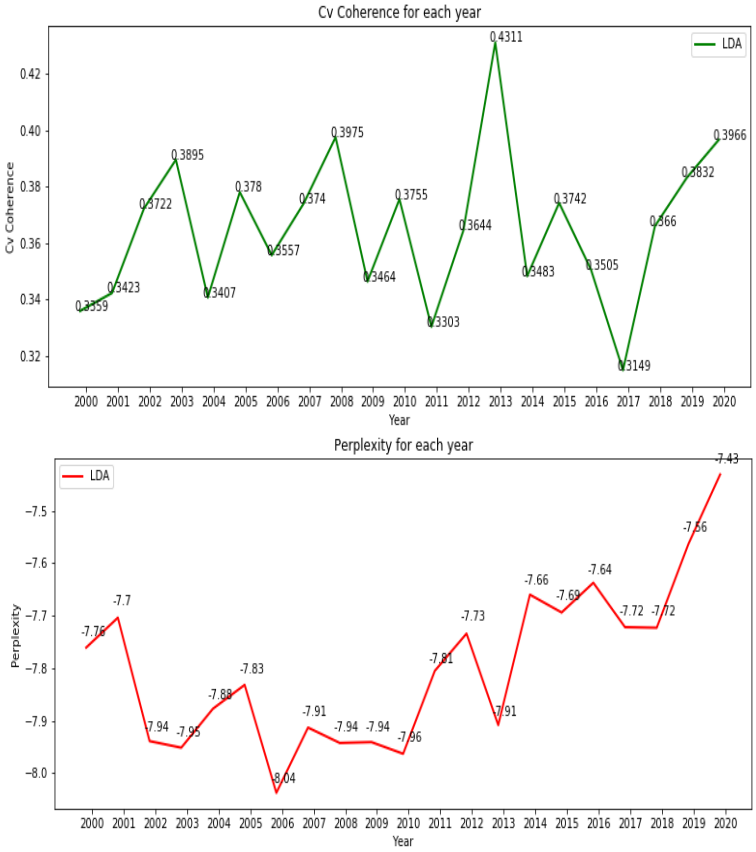

Fig. 11. Perplexity and coherence of sub-corpora by year.

For each annual sub-corpus, pyLDAvis is used to explore the latent topics and generated keyphrases. **Fig. 12** shows the pyLDAvis visualization for keyphrases from 2006. The five latent topics are reduced to circles plotted in two dimensions with Principal Component Analysis (PCA), providing an idea of how similar or different topics are from one another. The size of a topic circle is proportional to its marginal topic distribution. Next to the PCA visualization is a list of thirty candidate keyphrases for a selected topic. These candidates are ranked by the relevancy score, which can be controlled from sliding λ to a value from 0 to 1 (inclusive).

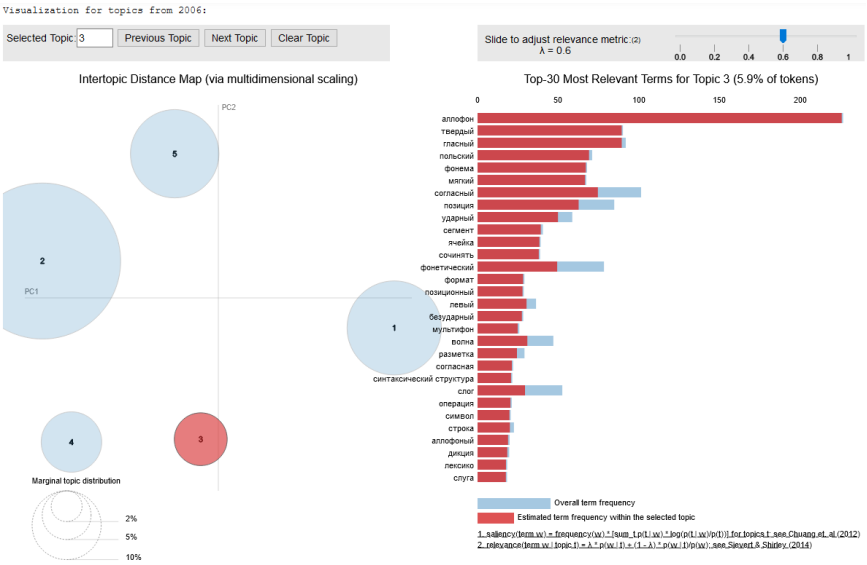

Fig. 12. Visualization of keyphrases for 2006 via pyLDAvis.

Through pyLDAvis, the relevancy weight parameter, λ, can be modified to change the ranking of keyphrases. A weight value of 0.6 is optimal for ranking keyphrases in a manner that describes a topic for human interpretation. However, a weight value of 0 ranks keyphrases solely on their lift, thereby generating keyphrases that are most frequent and exclusive to a topic. Although the lift metric can be noisy and rank extremely rare keyphrases highly, the top keywords are more specific. **Table 1.** displays the keyphrases generated for an LDA model trained on all Russian-language articles from the entire Dialogue conference, with a relevancy weight of 0.6 to understand the topic and a weight of 0 to generate specific keyphrases.

Table 1. Keyphrases and related documents for the entire corpus.

Yr	#	%	Top 5 Keyphrases (λ = 0.6)	Top 10 Keyphrases (λ = 0)
All	1	9.5	дискурсивный, реплика, устный, коммуникация, акт	дискурсивный, реплика, акт, коммуникативный, адресат, пропозиция, акцент, рема, испытуемый, иллокутивный
	2	24	выражать, семантика, отрицание, семантический, событие	выражать, движение, петя, тоска, отрицание, толкование, уезжать, сравнить, эффект, итак
	3	41.9	показывать, число, оценка, разный, вариант	оценка, вариант, показатель, частотный, раздел, частотность, допускать, высоко, гипотеза, коннектор
	4	8.6	запрос, разметка, обучение, документ, поиск	запрос, обучение, документ, интернет, сайт, выдача, электронный, обучать, сеть, поисковый
	5	15.9	задача, алгоритм, метод, связь, признак	тег, аннотация, классификатор, извлечение, тестирование, модуль, интерфейс, эксперт, атрибут, граф

The marginal topic distributions indicate the specificity of the topics. The unusually high marginal distribution of topic three suggests that the keyphrases are more representative of scholarly language in computational linguistics as a whole rather than a particular research area. Judging from the top five keyphrases ranked by pyLDAvis with a relevancy weight of 0.6, the remaining four topics can be roughly described as discourse and communication, semantics and word sense, IR, and machine learning. Setting the relevancy weight to 0 allows for extracting specific keyphrases for these topics. In the second topic, the keyphrase "петя" is a name that is very frequently used in the examples for a single document. The keyphrase "итак" is most likely a stop word, which suggests that additional post processing may need to be done to improve the performance of the LDA models.

For a deeper analysis of the keyphrases, LDA was induced on documents in each sub-corpora to estimate the topic distribution of the documents, with the documents having the highest topic distributions being displayed. **Table 2.** and **Table 3.** show the results of the topics models with unusual coherence and perplexity scores. In contrast to see how hyperparameters affect keyphrases, **Table 4.** and **Table 5.** show the results of the topic models with the highest coherence scores and perplexity scores that are not unusual for their relative years. The remaining yearly result data, showing the latent topics, marginal topic distributions, and top five human-interpretable keyphrases, and top ten specific keyphrases for each topic, can be found in the appendix.

Table 2. Keyphrases and related documents for 2013.

Yr	#	%	Top 7 Keyphrases ($\lambda = 0.6$)	Documents with Top 4 Highest Topic Distribution
2013	1	17	век, движение, пушкин, анализатор, пупок, жанр, щека	[2013_жестикуляционные-про - 69.51], [2013_жестикуляционные-про - 65.26], [2013_грамматический-слова - 55.49], [2013_александр-евгеньевич - 55.48]
	2	40.3	признак, описание, лексический, тело, характеристика, элемент, семантический	[2013_типологическая-база- - 93.63], [2013_больше-единицы-русск - 91.03], [2013_визуализация-данных- - 82.16], [2013_семантические-роли-и - 78.94]
	3	26.4	анекдот, фраза, адресат, действие, выражать, отрицание, ленин	[2013_вместе-или-раздельно - 80.35], [2013_неотрицаемые-предика - 69.92], [2013_кто-ищет-всегда-ли-н - 69.82], [2013_юбилеи-и-годовщины-в - 66.35]
	4	4.1	запрос, ошибка, исправление, опечатка, тестовый, документ, классификатор	[2013_автоматическое-испра - 76.14], [2013_влияние-различных-ти - 66.03], [2013_дорожка-по-оценке-ма - 40.95], [2013_correcting-collocati - 28.68]
	5	12.2	частотный, грамматический, метод, словоформа, распознавание, украинский, распределение	[2013_распознавание-двуязы - 76.33], [2013_классификация-отзыво - 72.09], [2013_использование-метода - 59.88], [2013_multi-functional-web - 56.33]

Table 3. Keyphrases and related documents for 2017.

Yr	#	%	Top 7 Keyphrases (λ = 0.6)	Documents with Top 4 Highest Topic Distribution
2017	1	4.5	файл, предикатив, идиолект, элемент, пространственный, текстовый, сигнал	[2017_автоматизация-постро - 82.82], [2017_русскоязычная-вопрос - 32.82], [2017_русские-предикативы- - 25.74], [2017_ruskell-13-онлайн-ин - 17.50]
	2	25.1	частота, событие, тип, именной, устный, гипотеза, наблюдать	[2017_корпус-кетских-и-эве - 90.33], [2017_automatic-generation - 84.68], [2017_интонация-перечислен - 70.32], [2017_идиоматика-пьянства- - 65.37]
	3	23.1	пример, отрицание, семантический, якобы, семантика, коннектор, единица	[2017_semantic-halo-of-a-m - 95.80], [2017_tensed-and-non-tense - 84.49], [2017_реверсивный-перевод- - 70.01], [2017_ксенопоказатели-по-д - 69.49]
	4	15.9	признак, связь, правило, набор, метод, алгоритм, пара	[2017_автоматическое-опред - 99.75], [2017_expanding-hierarchic - 80.96], [2017_коартикуляционные-из - 59.17], [2017_синтаксический-парсе - 52.93]
	5	31.4	различный, использование, проводить, показывать, количество, связывать, оказываться	[2017_lingcorpora-создание - 99.68], [2017_автоматическое-опред - 99.66], [2017_ruskell-13-онлайн-ин - 82.26], [2017_анализ-методов-класт - 76.43]

Table 4. Keyphrases and related documents for 2008.

Yr	#	%	Top 7 Keyphrases (λ = 0.6)	Documents with Top 4 Highest Topic Distribution
2008	1	8.3	пауза, падение, акцентный схема, подъем, акцент, предикация, запись, тон	[2008_арифметика-от-яндекс - 78.72], [2008_база-данных-интонаци - 52.79], [2008_интонация-незавершён - 50.35], [2008_способы-взаимодейств - 46.63]
	2	11.9	агент, фраза, состояние, клауза, адресат, протокол, придаточный, эмоция	[2008_глаголы-погружения-с - 62.15], [2008_дейксис-в-отсутствие - 54.84], [2008_домашние-слова-в-асп - 53.17], [2008_опыт-выборочного-под - 47.02]
	3	11	устный, диалектный, взаимодействие, разновидность, современный, жанр, языковой, вариант	[2008_вариантность-в-русск - 61.24], [2008_эволюция-форм-речево - 53.44], [2008_веб-пространство-и-м - 40.14], [2008_текстовый-диалектоло - 38.44]
	4	48	задача, информация, словосочетание, метод, структура, определять, связь, схема	[2008_комплексная-технолог - 99.38], [2008_портал-знаний-по-ком - 98.43], [2008_особенности-извлечен - 93.23], [2008_классификационная-сх - 92.69]
	5	20.8	поведение, эмоциональный, вещь, семантический, семантика, разный, тело, толкование	[2008_благородный-наивно-я - 91.21], [2008_идея-одноименности-в - 70.14], [2008_конструкция-с-творит - 67.84], [2008_я-не-был-меня-не-был - 66.63]

Table 5. Keyphrases and related documents for 2020.

Yr	#	%	Top 7 Keyphrases (λ = 0.6)	Documents with Top 4 Highest Topic Distribution

2019	1	17.1	вопросительный, позиция, косвенный, предположение, согласование, интерпретация, именной	[2019_driving-us-crazy-wit - 79.69], [2019_корпусная-грамматика - 60.84], [2019_pragmatics-in-the-in - 56.12], [2019_слово-это-в-частном- - 54.88]
	2	5.5	импликатура, дискурсивный, отменять, значение, валентность, неопределенность, параллельный	[2019_семантические-типы-и - 38.67], [2019_русское-что-то-как-д - 35.90], [2019_some-features-of-the - 35.01], [2019_derivative-meanings- - 33.13]
	3	39.3	пример, значение, приходить, факт, встречаться, приводить, условие	[2019_adding-to-the-treasu - 99.92], [2019_conceptualization-of - 99.91], [2019_news-headline-genera - 99.07], [2019_headline-generation- - 98.94]
	4	28.6	разметка, результат, задача, качество, сущность, данный, показывать	[2019_evolution-of-dialect - 88.98], [2019_an-interactive-dicti - 87.88], [2019_named-entites-in-cyb - 84.67], [2019_analysis-of-prosodic - 79.03]
	5	9.5	коррекция, строка, монолог, пересказчик, дискурс, фрагмент, комментатор	[2019_самоисправления-гово - 62.90], [2019_unified-multichannel - 43.91], [2019_просодия-и-грамматик - 36.17], [2019_поиск-в-мультиканаль - 30.96]

6 Discussion

The coherence and perplexity scores of the sub-corpora shown in **Fig. 11** give insight into the quality of topics and keyphrases that are extracted. The first two years of the conference have an unusually low perplexity compared to the following few years. This suggests that the research field of the Dialog conference broadened in the third year, which is not unusual for a conference as it becomes more popular and established over time. The lower average perplexity scores since 2011 reflect the sharp drop in the number of Russian-language articles in the Dialogue conference since that time. The reduction in corpus size and steady lowering of perplexity since 2011 indicates that the research space of Russian-language articles for this conference is becoming increasingly narrower in recent times. Although hyperparameter optimization is done for the entire Dialogue corpus in this paper, the differences in sub-corpora size and perplexity indicates that it may be better to optimize hyperparameters separately for documents before 2011 and since 2011.

The coherence for 2013 is unusually high but the perplexity is unusually low, which suggest that the keyphrases for this sub-corpus are from a very small amount of documents. 2017 has a normal perplexity, but the coherence is unusually low, signaling that the learned topics are difficult to interpret from the associated keyphrases. 2020 has an unusually low perplexity, but it has the second highest coherence, which means that the topics may be easy to interpret, but the diversity of topics is much lower than in previous years. The models for 2008 and 2019 have perplexity scores that are close to the average for their time periods, and they have the highest coherence which falls under the expected range.

Table 2. displays the keyphrases from 2013. The LDA model trained on this sub-corpus had an unusually high coherence and an unusually low perplexity. This is reflected in the keyphrases generated for the first and third topics. While "век [=century]", "движение [=movement]", "анализатор [=analyzer]", "пушкин [=pushkin]", and "жанр [=genre]" are keyphrases that may suggest that the topic describes genre analysis over time, it is not immediately clear how the keyphrases describing body parts, "пупок [=bellybutton]" and "щека [=neck]," are related. When looking at the documents with the top highest topic distributions, two of the documents with the highest topic distribution have expanded versions included in the sub-corpus, which artificially raises the rankings of their keyphrases. While it may seem that the keyphrase "щека" may originate from the top two documents, which involve gestures in Russian, it's very unlikely that "пупок" also comes from these documents. In fact, the top two documents are not the sources for either of these two keyphrases, rather the keyphrases are extremely rare and likely receive a higher rating due to their lift. "пушкин" only appears once in an example in the third highest document, while "пупок" and "щека" appear only once and twice in examples in the fifth highest document. As for the third topic, the seemingly random keyphrase "ленин [=lenin]" is frequently used in the fourth highest document, which analyzes anniversaries in Russian jokes.

Table 3. displays the keyphrases from 2017. This model has a normal perplexity, but the coherence is unusually low. The first topic has a relatively small marginal distribution and it is difficult to understand the underlying concept that is characterized by the keyphrases. Each keyphrase appears very frequently in only two of the highest documents, forming a chain that narrowly links the documents together. For example, the first keyphrase appears frequently in the first and second document, while the second keyphrase appears frequently in the third and fourth document, and the third keyphrase is frequent only in the third document.

Table 4. and **Table 5.** display the keyphrases for 2008 and 2019. The topics with the highest marginal topic distributions generate keyphrases that may be common in a scholarly text but are not as specific as the keyphrases generated for topics with the smallest marginal distributions. For example, "задача [=task]," "информация [=information]," and "словосочетание [collocation]" are very common computer science and linguistics terms and are generated from a topic with nearly half of the marginal distribution for 2008, whereas "пауза [=pause]," "падение [=falling]," and "акцентный схема [=accent paradigm]" are generated from a topic with a relatively small marginal distribution, but are more specific to a certain research area in linguistics.

The extracted keyphrases from the yearly sub-corpora and entire Dialogue corpus give insight into some of the research areas and salient keyphrases used in Russian computational linguistics. Due to the sequential nature of the methods, the quality of earlier steps in the pipeline have an impact on the later steps; it is a snowball effect that can greatly impact the overall quality of automatic keyphrase extraction. This snowball effect is present in this paper due to limited access to state-of-the-art techniques for text extraction from scholarly PDFs and keyphrase extraction implementations adapted towards Russian-language text. Some errors throughout the methods have impacted the overall quality of the topics learned by LDA and the keyphrases which characterize these topics.

The first error comes from the extraction of text from PDFs with the pdftotext tool. The tool was unable to successfully extract the text for a small amount of articles, resulting in the malformed text seen in **Fig. 6**. Furthermore, the pdftotext tool extracted all body text from the page and joined together words that were split from end-of-line hyphenation, but it included text from tables, in-text examples, figures, and sometimes footnotes. In some articles, the footnotes did not appear in the footer of the article, but rather in the body paragraph with a demarcation line to signify that the section was not part of the body text. This led to unpredictable behavior, with text sometimes disappearing or appearing in the wrong order. Additionally, the pdftotext tool failed to remove footnote numbers from the end of words in the body paragraph. Other than this, end-of-line hyphenation was sometimes not removed at page breaks. While this paper attempts to remove these split words entirely, removal of words may have negative ramifications for morphological normalization of the sentence as a whole. Layout-aware PDF text extraction tools and OCR can be used to improve the quality of text extracted from the articles, as well as filter out irrelevant sections.

The next error occurs from the language misclassification from the langid library and can be mitigated by improving the PDF text extraction. Articles which contain a significant number of Russian words in the examples are mistakenly classified as Russian articles, even if the main text is in English. This may be because English occurring in Russian text is more probable than the reverse. This means that the presence of a significant amount of Russian may lead to a misclassification, but it is not clear precisely how sensitive the tool is to the presence of Russian in English text. This is particularly significant for international conferences such as Dialogue because many of the papers published since 2011 include an abstract and title page written in both Russian and English. This paper filters out any English tokens, but the inclusion of misclassified articles has an impact on the word frequency and document frequency of all keyphrase tokens.

Improper lemmatization causes some errors as well. The preprocessing methods provided by the WebVectores paper incorrectly attempt to lemmatize abbreviated words and acronyms. This results in keyphrase tokens which include words that do not exist in Russian. For example “рис.” and “яндекс.новости” turned into “рисяча” and “яндексяча новости,” and the abbreviation for *Национальный корпус русского языка* НКРЯ was lemmatized as “нкрь.” Nevertheless, the methods did not appear to lemmatize all novel keywords for a research community, which results in the candidate keyphrases ‘импликатура’ and ‘импликатур’, and ‘самоисправление’ and ‘самоисправления’.

During candidate keyphrase ranking with pyLDAvia, there is no redundancy control for keyphrases. This means that unigrams which appear in higher-ranking multi-word keyphrases are not removed. For example, the lower ranking keyphrase “дискурсивный” is not removed, even though it appears in the higher ranking keyphrases “элементарный дискурсивный единица” and “дискурсивный маркер.” Also, there is no control of the appearance of identical Russian keyphrases in varying POS inflectional forms. For instance, it is not uncommon to see nouns alongside derived adjectives, such as “самоисправление” with “самоисправлений,” and “семантика” with “семантический.”

Finally, this paper takes a topic-modeling based approach to the automatic extraction of keyphrases. Although LDA models are used in sophisticated topic-clustering approaches which include graph-based ranking techniques, topic modeling alone is not representative of state-of-the-art keyphrase extraction approaches. Most of the open source and free Python implementations for automatic keyphrase extraction are adapted towards English-language text, thereby limiting the convenience of rapid prototyping automatic keyphrase from Russian-language text.

7 Conclusion

Although the preprocessing step of this paper contains a number of errors, the results generated in the keyphrase ranking step are still useful. Depending on the goals of the research, the optimization of hyperparameters can be manipulated to include more specific or more general keyphrases. Furthermore, the marginal distribution of the topics generated by a topic modeler used to judge whether a topic is more specific to a particular research area or to the research community as a whole.

While this paper used limited methods for preprocessing Russian scholarly text and automatically extracting keyphrases, it serves as a proof of concept for applying simplistic topic modeling techniques to explore the salient technical terms used in a research community. To improve the results of this paper, more sophisticated methods should be introduced into the pipeline. For improving the quality of text extracted from PDFs and removing irrelevant text from examples and tables, layout-aware text extraction software, such as such as LA-PDFText[16], may be used. Better morphological normalization techniques, redundancy control mechanisms, and additional post-processing can be used to improve the quality of the candidate keyphrases. More sophisticated topic-clustering techniques which use LDA can be used to improve the ranking of candidate keyphrases, but there is also active research into the usage of neural networks and deep learning for better automatic keyphrase extraction [18], [20], [21].

In this paper, an LDA and pyLDAvis based topic-clustering approach to automatic keyphrase extraction was used to discover the salient keyphrases used in the largest Russian computational linguistics and Russian NLP conference, Dialogue. The results show that the keyphrases in this conference generally fall under four latent topics, discourse and communication, semantics and word sense, IR, and machine learning. Topics with high marginal distributions may be representative of the technical terminology of the Dialog conference as a whole, rather than a particular research area. Furthermore, manipulation of the hyperparameters can have an impact on the specificity of the keyphrases generated, but highly specific terms may distract from the accuracy of the results. More sophisticated techniques can be used in various steps of the pipeline in order to generate higher quality keyphrases. The task of automatic keyphrase extraction remains an active area of research and improvements in these approaches can be useful for aspiring researchers to explore the research areas and technical terminology of a research community.

[16] https://github.com/BMKEG/lapdftext

References

1. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society for Information Science,* vol. 41, no. 6, pp. 391-407, Sep. 1990.
2. T. Hofmann, "Probabilistic Latent Semantic Indexing," in *SIGIR '99*, Berkley, 1999.
3. K. S. Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation,* vol. 28, no. 1, pp. 11-21, 1972.
4. F. J. Damerau, "Generating and evaluating domain-oriented multi-word terms from texts," *Information Processing & Management,* vol. 29, no. 4, pp. 433-447, 1993.
5. J. Jardine and S. Teufel, "Topical PageRank: A Model of Scientific Expertise for Bibliographic Search," in *European Chapter of the Association for Computational Linguistics*, Gothenburg, 2014.
6. А. С. Ванюшкин and Л. А. Гращенко, "Методы и алгоритмы извлечения ключевых слов," *Новые информационные технологии в автоматизированных системах,* no. 19, pp. 85-93, 2016.
7. С. О. Шереметьева and П. Г. Осминин, "Методы и модели автоматического извлечения ключевых слов," *Bulletin of the South Ural State University. Ser. Linguistics,* vol. 12, no. 1, pp. 76-81, 2015.
8. О. С. Недильченко, "Этапы и методы автоматического извлечения ключевых слов," *Молодой учёный,* vol. 22, no. 156, pp. 60-62, Июнь 2017.
9. R. Mihalcea and P. Tarau, "TextRank: Bringing Order into Texts," in *Conference on Empirical Methods in Natural Language Processing*, Barcelona, 2004.
10. K. Hasan and V. Ng, "Automatic Keyphrase Extraction: A Survey of the State of the Art," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, 2014.
11. F. Boudin, H. Mougard and C. D, "How Document Pre-processing affects Keyphrase Extraction Performance," in *Proceedings of the 2nd Workshop on Noisy User-generated Text*, Osaka, 2016.
12. C. Florescu and C. Caragea, "PositionRank: An Unsupervised Approach to Keyphrase Extraction from Scholarly Documents," in *Meeting of the Association for Computational Linguistics*, Vancouver, 2017.
13. X. Wan and J. Xiao, "Single document keyphrase extraction using neighborhood knowledge," in *Cational conference on Artificial intelligence*, Illinois, 2008.
14. N. Teneva and W. Cheng, "Salience Rank: Efficient Keyphrase Extraction with Topic Modeling," in *Association for Computational Linguistics*, Vancouver, 2017.
15. T. Tomokiyo and M. Hurst, "A Language Model Approach to Keyphrase Extraction," in *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment*, 2003.
16. A. Bakarov, A. Kutuzov and I. Nikishina, "Russian Computational Linguistics: Topical Structure in 2007-2017 Conference Papers," in *Dialogue-2018*, Moscow, 2018.

17. Dialogue, "Proceedings," [Online]. Available: http://www.dialog-21.ru/en/digest/. [Accessed 21 July 2020].
18. F. Boudin, Y. Gallina and A. Aizawa, "Keyphrase Generation for Scientific Document Retrieval," in *Annual Meeting of the Association for Computational Linguistics*, Online, 2020.
19. I. Nikishina, A. Bakarov and K. A, "RusNLP: Semantic search engine for Russian NLP conference papers," in *Proceedings of AIST-2018*, Moscow, 2018.
20. X. Zhu, C. Lyu, D. Ji, H. Liao and F. Li, "Deep neural model with self-training for scientific keyphrase extraction," *PLOS ONE*, vol. 15, no. 5, p. e0232547, 2020.
21. R. Meng, S. Zhao, S. Han, D. He, P. Brusilovsky and Y. Chi, "Deep Keyphrase Generation," in *Annual Meeting of the Association for Computational Linguistics*, Vancouver, 2017.
22. S. Bird, R. Dale, B. J. Dorr, B. Gibson, M. T. Joseph, M.-Y. Kan, D. Lee, B. Powley, D. R. Radev and Y. F. Tan, "The ACL Anthology Reference Corpus: A Reference Dataset for Bibliographic Research in Computational Linguistics," in *International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, 2008.
23. A. Kutuzov and E. Kuzmenko, "To Lemmatize or Not to Lemmatize: How Word Normalisation Affects ELMo Performance in Word Sense Disambiguation," in *Proceedings of the First NLPL Workshop on Deep Learning for Natural Language Processing*, Turku, 2019.
24. J. Kamps, C. Monz, M. Rijke and S. B, "Language-dependent and Language-independent Approaches to Cross-Lingual Text Retrieval," in *Comparative Evaluation of Multilingual Information Access Systems*, Berlin, 2014.
25. A. A. Sorokin, "Improving Neural Morphological Tagging using Language Models," in *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conderence "Dialogue 2018"*, Moscow, 2018.
26. W. Comer, "Measured Words: Quantifying Vocabulary Exposure in Beginning Russian," *Slavic & East European Journal*, vol. 63, no. 1, pp. 92-114, 2019.
27. L. Page, S. Brin, R. Motwani and T. Winograd, "The PageRank Citation Ranking: Bringing Order to the Web," *Stanford InfoLab*, Nov. 1999.
28. R. Peter, S. Gopakumar, D. G and S. KP, "Evaluation of SVD and NMF methods for Latent Semantic Analysis," *International Journal of Recent Trends in Engineering*, vol. 1, no. 3, pp. 308-310, 2009.
29. D. O'Callaghan, D. Greene, J. Carthy and P. Cunningham, "An analysis of the coherence of descriptors in topic modeling," *Expert Systems with Applications*, vol. 42, no. 13, pp. 5645-5657, 2015.
30. D. M. Blei, A. Y. Ng and M. I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 3 Mar. 2003.
31. M. Fares, A. Kutuzov, S. Oepen and E. Velldal, "Word vectors, reuse, and replicability: Towards a community repository of large-text resources," in *Proceedings of the 21st Nordic Conference of Computational Linguistics*, Gothenburg, 2017.
32. F. Boudin, "pke: an open source python-based keyphrase extraction toolkit," in *International Conference on Computational Linguistics: System Demonstrations*, Osaka, 2016.

33. A. Kutuzov and E. Kuzmenko, "WebVectores: a toolkit for building web interfaces for vector semantic models," in *International Conference on Analysis of Images, Social Networks and Texts*, Yekaterinburg, 2016.

Appendix

Yr	#	%	Top 5 Keyphrases (λ = 0.6)	Top 10 Keyphrases (λ = 0)
2000	1	32.7	связь, компонент, точка зрение, связывать, отмечать	направление, компьютер, целое, восприятие, информационный, числовой, словосочетание, наблюдение, интернет, требование
	2	7.9	общение, действие, средний, указательный, идиома	идиома, указательный, собеседник, партнер, подросток, добрый, французский язык, референт, выбрать, семья
	3	26.4	правило, понятие, возможность, основа, структура	понимание, словоформ, синтез, понятие, ответ, закон, алфавит, лингвистика, существенно, операция
	4	18.8	узел, граница, символ, синтаксический, сокращение	граница, символ, узел, сокращение, статья, образовывать, естественный язык, запись, словарный статья, переводить
	5	14.2	значение, категория, валентность, семантический, предмет	валентность, наблюдатель, берег, заполнять, квантор, локализация, семантический узел, запрет, нсв, рефлексив
2001	1	18.2	документ, алгоритм, метод, поиск, этап	документ, алгоритм, кластер, ударение, специализированный, получение, пакет, распознавание, фонетический, электронный библиотека
	2	9.9	вербальный, устный, невербальный, спонтанный, селькупский	вербальный, устный, невербальный, спонтанный, селькупский, дискурс, воздействие, коммуникация, реклама, ладно
	3	8.5	тезаурус, идиома, таксон, страх, эмоция	идиома, таксон, страх, эмоция, метафорический, политический, церковь, семья, удивление, периферия таксоный
	4	20.7	значение, обозначать, разный, тип, исключение	анекдот, счетный, нес, обозначать, числительный, член, сходство, совотреть, итак, валентность
	5	42.6	понятие, знание, описание, структура, использование	понятие, стратегия, предметный область, онтология, знание, сегмент, моделирование, гипотеза, лексика, ресурс
2002	1	37.3	документ, задача, информация, позволять, основа	документ, тезаурус, письмо, список, программа, ошибка, объем, транскрипция, флексия, сообщение
	2	17.2	понятие, знание, вербальный, графический, онтология	вербальный, онтология, графический, изображение, когнитивный, наука, ячейка, метафора, реальность, мпо
	3	11.1	общение, жанр, лингвистический, стиль, научный	общение, стиль, стратегия, тысяча, электронный, социальный, точность, совет, конфликт, ядро
	4	12.5	поведение, коммуникация, сознание, радость, собеседник	радость, собеседник, эмоция, мотив, общество, личность, разговорный, разговор, сравнить, номинация
	5	22	значение, синтаксический, действие, описывать, интернет	различие, разница, дискурс, предикат, событие, формирование, предметный область, объяснять, расхождение, глагольный
2003	1	11.6	документ, информация, запрос, поиск, пользователь	документ, запрос, тезаурус, поисковый, интернет, формирование, страница, ресурс, строка, классификация
	2	15.4	значение, событие, семантика, предикат, различать	семантика, различать, смысловой, зависимый, вершина, варьировать, употреблять, концепция, устройство, идиома

	3	32.7	правило, описание, задача, структура, класс	класс, синтез, множество, технология, признак, разрабатывать, сигнал, набор, преобразование, предметный область
	4	32.9	понимание, знание, подобный, цель, фраза	понимание, общение, сообщение, грамматика, коммуникация, деятельность, внимание, взаимодействие, суффикс, коммуникативный акт
	5	7.4	анекдот, испытуемый, значение, сонет, собака	анекдот, испытуемый, сонет, собака, пушкин, кошка, петя, актант, кличка, подзначение
2004	1	7.6	событие, стратегия, письменный, дизъюнкция, реакция	стратегия, событие, письменный, реакция, устный, генератор, чередование, нарратив, ориентация, когнитивный
	2	22.9	значение, вовремя, резко, отрицание, предикат	вовремя, резко, толкование, приходить, сельскупский, сходство, оставлять, трансформация, период, неопределенный
	3	39.4	информация, задача, использование, база, позволять	база, пользователь, метод, электронный, запрос, поиск, разрабатывать, алгоритм, автоматический, построение
	4	23.1	класс, падеж, сравнение, значение, собственный	падеж, средний, поведение, референциальный, референт, килограмм, валентность, тонна, зависимый, именной
	5	7	спам, окончание, мнута ч, письмо, фонетический	спам, мнута ч, письмо, фонетический, рекламный, сайт, просодический, звуковой, аллофон, фонема
2005	1	36	разный, семантический, синтаксический, выражать, оказываться	синтаксический, фактор, различие, средний, толкование, указание, характерный, объяснять, ожидание, целое
	2	8.5	предлог, предикат, анекдот, валентность, все-таки	предикат, анекдот, все-таки, посессор, наблюдатель, референт, придаточный, императив, внешний посессор, продолжительность
	3	10	метафора, коррупция, поведение, рациональный, обоюдный	метафора, коррупция, поведение, рациональный, обоюдный, адресат, норма, конфликт, эмоциональный, социальный
	4	23.2	документ, запрос, поиск, информация, пользователь	документ, запрос, поиск, распознавание, аннотация, технология, поисковый, онтология, интерфейс, компания
	5	22.3	правило, структура, алгоритм, знак, фонема	фонема, звук, эксперт, синтез, предметный область, ядро, синсет, вычисление, испытуемый, строка
2006	1	18.7	действие, понимание, глубина, наблюдатель, дискурс	действие, понимание, наблюдатель, дискурс, отрицательный, точка зрение, французский, стратегия, слой, общение
	2	51.3	информация, понятие, структура, основа, связь	понятие, метод, документ, информация, количество, алгоритм, знание, онтология, подход, поиск
	3	5.9	аллофон, твердый, гласный, польский, фонема	аллофон, твердый, фонема, мягкий, ячейка, сочинять, позиционный, гласный, польский, формат
	4	7.7	правдоподобие, фраза, естественный, синтезировать, событие	правдоподобие, синтезировать, событие, восприятие, сигнал, акцент, интонация, изображение, река, рема
	5	16.4	добрый, размер, предлог, толщина, употреблять	добрый, толщина, судьба, высота, тело, ширина, единица измерение, косвенный, пословица, ладонь
2007	1	6.4	множественный число, недавний, запятая, давний, художественный	множественный число, недавний, запятая, вчерашний, круглый, периферия, омонимия, жестовый язык, пунктуация, магазин

	2	22.9	сочетание, семантика, адресат, указывать, употреблять	адресат, семантика, предлог, предикат, толкование, анекдот, выглядеть, указательный, тезис, изображение
	3	47.7	задача, использование, структура, понятие, основа	алгоритм, задача, категория, пользователь, сайт, маркер, разрабатывать, учитывать, база, автоматический
	4	16.6	признак, эксперт, класс, клиент, кавычка	эксперт, клиент, цвет, дискурс, поле, коммуникация, устный, прием, воздействие, вещь
	5	6.4	диктор, сегмент, испытуемый, индикатор, сигнал	диктор, испытуемый, индикатор, сигнал, аллофон, фонетический, слог, интонационный, тон, гласный
2008	1	8.3	пауза, падение, акцентный схема, подъем, акцент	пауза, падение, акцентный схема, подъем, акцент, предикация, тон, дискурс, рассказ, реплика
	2	11.9	агент, фраза, состояние, клауза, адресат	агент, клауза, протокол, придаточный, хозяин, относительный, платонов, анекдот, реакция, оборот
	3	11	устный, диалектный, взаимодействие, разновидность, современный	устный, диалектный, разновидность, взаимодействие, жанр, персонаж, смена, наблюдение, языковой личность, диалект
	4	48	задача, информация, словосочетание, метод, структура	задача, словосочетание, метод, понятие, содержать, схема, список, зависимый, пользователь, онтология
	5	20.8	поведение, эмоциональный, вещь, семантический, семантика	эмоциональный, вещь, тело, толкование, апресян, творительный, локативный, акт, колокол, нос
2009	1	8.9	пауза, граница, положение, рассказ, ссылка	пауза, граница, положение, рассказ, ссылка, дискурс, вершина, просодический, страница, говорящий
	2	22.9	разметка, концепт, омограф, создание, информация	концепт, омограф, автоматический, разметка, сеть, граф, программа, узел, информационный, морфологический
	3	15.7	поиск, распределение, алгоритм, круглый, ключевой	распределение, алгоритм, ключевой, круглый, частота, параллельный, опорный, выпуклый, коллекция, оператор
	4	35.3	семантический, глагольный, называть, сочетание, указывать	сочетание, обращение, нос, нулевой, толкование, предмет, идиома, теплый, сочетаться, нуль
	5	17.2	единица, заменять, запись, действие, тело	заменять, действие, тело, хотеться, испытуемый, шрифт, собирать, отрицательный, блог, художественный
2010	1	24.9	задача, подход, связь, понятие, метод	алгоритм, референциальный, документ, процедура, предметный область, референциальный выбор, анализатор, тезаурус, извлечение, точность
	2	7.9	тело, соматический, анекдот, грудь, щека	тело, соматический, анекдот, грудь, щека, коннотация, сердце, семиотический концептуализация тело, орган, любовь
	3	18.8	создание, база, программа, запись, информация	знак препинание, распознавание, напр, синтез, блок, чтение, обучение, рсс, программный, заголовок
	4	26.4	метафора, энциклопедический, семантический, описание, структура	метафора, энциклопедический, разговорный, метафорический, адресат, звук, зона, характеризоваться, штраф, белорусский
	5	21.9	испытуемый, идиома, отвечать, средний, экспериментальный	идиома, испытуемый, отвечать, дискурс, событие, скорее, дискурсивный, собеседник, отражение, отрицание

2011	1	26.6	впечатление, идиома, пословица, употреблять, семантический	идиома, пословица, впечатление, предикат, реагировать, энантиосемия, толкование, студент, отрицание, семантический поле
	2	11	падеж, анекдот, ожидание, персонаж, допускать	падеж, анекдот, ожидание, вспоминать, помогать, хлеб, антецедент, тело, независимый, повторять
	3	12.3	синтез, диалектный, фонетический, цитирование, устный	цитирование, эпизод, эмоциональный состояние, слог, граница, чужой, цитация, знак препинание, синтагм, движение
	4	27.1	ошибка, опечатка, вариант, тональность, исправление	опечатка, тональность, ошибка, исправление, документ, вершина, синтаксический анализатор, частотность, подход, распознавание
	5	23.1	языковой, помощь, различие, полный, функция	различие, общение, программа, история, указание, факт, доля, грамматика, научный, статус
2012	1	17.6	загадка, запрос, вариант, идиома, современный	загадка, идиома, страница, узус, обувь, формат, вхождение, сайт, сегмент, компьютерный
	2	11	препарат, образец, метафора, лекарство, средство	образец, метафора, препарат, рема, параметрический, коммуникативный членение, солнце, информант, запах, составляющая
	3	27.2	правило, разметка, класс, фрагмент, классификация	разбор, алгоритм, классификатор, эксперт, классификация, процедура, токен, вершина, дерево, китайский язык
	4	39.2	момент, предполагать, совпадение, компонент, событие	событие, совпадение, выражать, предикат, предполагать, обозначать, английский, коммуникативный, невозможный, предшествовать
	5	5.1	реплика, ага, произношение, спонтанный, мужчина	ага, произношение, спонтанный, междометие, диктор, вызов, ожидание, произнесение, отрывок, дискурсивный единица
2013	1	17	век, движение, пушкин, анализатор, пупок	пушкин, пупок, анализатор, жанр, щека, исторический, диалект, волос, номинализация, болеть
	2	40.3	признак, описание, лексический, тело, характеристика	тело, характеристика, фрейм, параметр, поле, предлог, выражаться, толкование, средний, отличать
	3	26.4	анекдот, фраза, адресат, действие, выражать	фраза, ленин, ирония, отрицание, сравнить, событие, угроза, двое, дискурсивный, читать
	4	4.1	запрос, ошибка, исправление, опечатка, тестовый	опечатка, асессор, метрика, исправлять, исправление, ранжирование, оценка качество, тестовый, классификатор, сценарий
	5	12.2	частотный, грамматический, метод, словоформа, распознавание	украинский, вариативность, текстовый, модуль, тональность, сентимент, словоформ, вручную, распознавание, двуязычный
2014	1	15.5	конфигурация, собственно щепоть, точь, перо, палец	собственно щепоть, точь, перо, движение, ладонь, щепоть, плоскость, конфигурация, жестикуляция, кольцо
	2	37.2	цель, действие, любой, семантика, связь	ответ, вопросительный, метр, акт, кандидат, средство, иллокутивный, указание, применение, кластер
	3	12.1	интонация, реализация, сослагательный наклонение, отрицание, предлог	интонация, сослагательный наклонение, предлог, реализаций, валентность, приведять, индикатив, синтез, выражаться, петя
	4	19.8	концепт, интонационный, испытуемый, коллекция, ударный	интонационный, коллекция, диалектный, рисунок, диктор, алгоритм, испытуемый, концепт, порождение, начальный

	5	15.4	фильм, модификация, самоисправлений, встречаться, фрагмент	модификация, самоисправлений, чего-ловек, самоисправление, тизер, коррекция, сиквел, устный, изоморфизм, вхождение
2015	1	20.6	база, частотность, информация, окончание, частота	окончание, частотность, гласный, словоформа, запись, гикрь, алгоритм, интернет, регион, полезный
	2	9.1	сочетание, бывать, мастер, аплодисменты, гугл	аплодисменты, гугл, перебиваться, мастер, кузнечный, фразеологический, корпусной лингвистика, мера ассоциация, коллокация, устойчивый словосочетание
	3	31.6	размер, семантика, семантический, свойство, плод	плод, размер, толкование, называться, ожидание, предложный, предикат, приходить, лингвоспецифичный, дательный
	4	24	признак, движение, граница, социальный, информант	движение, фонетический, строка, сигнал, диктор, формальный, опыт, код, распознавание, единичный
	5	14.8	собеседник, персонаж, функция, акцент, коммуникативный	собеседник, персонаж, акцент, крайний мера, знакомиться, квантор, рема, эмоция, говорящий, иллюстративный
2016	1	32.3	диктор, признак, событие, семантика, интерпретация	нсв, устный, связь, испытуемый, строка, последовательность, характерный, естественный, просодический, святой
	2	2.9	диктор, идентификация, аудитор, пауза, дыхательный	диктор, аудитор, дыхательный, дыхание, озвучивать, озвучивание, звуковой, знакомый, темпоральный пауза, дыхательный пауза
	3	21.5	признак, частота, класс, использование, выборка	частота, выборка, документ, запрос, коллекция, обработка, лексикон, построение, вектор, обучение
	4	10.6	семантика, интерпретация, мужчина, свойство, момент	мужчина, коннектор, иван, река, дистрибутивный, этаж, дейктический, подарок, актант, справа
	5	32.6	событие, разный, фрагмент, понятие, оценка	событие, фрагмент, понятие, отрицание, исчезать, пара, ответ, акцент, маркер, начинать
2017	1	4.5	файл, предикатив, идиолект, элемент, пространственный	предикатив, идиолект, сигнал, словник, яма, идиолекть, предикативый, файл, произношение, повествовательный
	2	25.1	частота, событие, тип, именной, устный	событие, наблюдать, просодический, движение, частота, идиома, положение, интонационный, исследовать, стратегия
	3	23.1	пример, отрицание, семантический, якобы, семантика	отрицание, якобы, коннектор, пропозиция, французский, валентность, языковой единица, эквивалент, выходить, итальянский
	4	15.9	признак, связь, правило, набор, метод	реализация, обучение, цепочка, процедура, сказуемое, формула, согласный, образование, граф, токен
	5	31.4	различный, использование, проводить, показывать, количество	фактор, грамматический, участие, основание, количественный, классификация, выбор, остальной, доля, сложность
2018	1	37.9	тип, указывать, фраза, процесс, структура	абы, длительный, как-нибудь, граница, пауза, иллокутивный, императив, относительный, указывать, устный
	2	29.8	отрицание, возможный, предикат, семантический, компонент	отрицание, кончаться, нкрь, состояние, ход, длиться, ответ, приступать, толкование, по-видимому
	3	2.7	расти, приходить, адресат, верить, сообщать	расти, приходить, верить, сообщать, школа, поверить, смерть, метатекстовый, избирать, уходить

	4	5.9	диалектный, запись, коммуникация, объем, сообщение	диалектный, информантовый, информант, комбинировать, март, социальный, дискурсивный маркер, эпизод, написание, записывать
	5	23.8	синтаксический, разметка, необходимый, оценка, модальный	синтаксический, оценка, разметка, модальный, ресурс, база, дальнейший, подход, необходимый, множество
2019	1	17.1	вопросительный, позиция, косвенный, предположение, согласование	вопросительный, согласование, предположение, подлежащее, именной, сфера действие, подлежать, по-видимому, акцентоноситель, подтверждать
	2	5.5	импликатура, дискурсивный, отменять, значение, валентность	импликатура, отменять, валентность, неопределенность, параллельный, соглашаться, лексический значение, внутренний состояние, погибать, подозрительно
	3	39.3	пример, значение, приходить, факт, встречаться	приходить, факт, условие, препозитивный, происходить, получаться, заголовок, количественный, связь, касаться
	4	28.6	разметка, результат, задача, качество, сущность	символ, обучение, диктор, сущность, чтение, пользователь, ошибка, пауза, разметка, длина
	5	9.5	коррекция, строка, монолог, пересказчик, дискурс	коррекция, монолог, пересказчик, строка, комментатор, груша, рассказчик, локутор, переходить, партия
2020	1	5.2	обратно, состояние, открывать, бить, замерзать	обратно, состояние, замерзать, снег, мужчина, закрытый, рвать, тройка, бить, воздух
	2	16	тоска, запрос, нсв2, префикс, сочетаемость	тоска, префикс, семантический класс, оригинал, скетч, поле, оригинальный, желание, способный, переводной
	3	19.1	частотность, тест, проводить, правильно, ошибка	частотность, тест, правильно, английский язык, влияние, взаимодействие, распределение, пользователь, орфографический, частота
	4	56.5	вводный, тип, рассматривать, возможность, разный	петя, предположение, интерпретация, выражать, отрицание, уезжать, разметка, настоящий, коннектор, предикат
	5	3.2	ресурс, дополнять, качественный, влиять, совместный	ресурс, дополнять, качественный, совместный, совместно, пациент, обрабатывать, извлекать, праздник, скорость
All	1	9.5	дискурсивный, реплика, устный, коммуникация, акт	дискурсивный, реплика, акт, коммуникативный, адресат, пропозиция, акцент, рема, испытуемый, иллокутивный
	2	24	выражать, семантика, отрицание, семантический, событие	выражать, движение, петя, тоска, отрицание, толкование, уезжать, сравнить, эффект, итак
	3	41.9	показывать, число, оценка, разный, вариант	оценка, вариант, показатель, частотный, раздел, частотность, допускать, высоко, гипотеза, коннектор
	4	8.6	запрос, разметка, обучение, документ, поиск	запрос, обучение, документ, интернет, сайт, выдача, электронный, обучать, сеть, поисковый
	5	15.9	задача, алгоритм, метод, связь, признак	тег, аннотация, классификатор, извлечение, тестирование, модуль, интерфейс, эксперт, атрибут, граф