sciendo

BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **21**, No 3
Sofia • 2021 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2021-0027

Text Analytics in Bulgarian: An Overview and Future Directions

Gloria Hristova

Department of Statistics and Econometrics, Faculty of Economics and Business Administration, Sofia University "St. Kliment Ohridski", 1113 Sofia, Bulgaria
E-mail: g.hristova@feba.uni-sofia.bg

***Abstract*:** *Text analytics is becoming an integral part of modern business and economic research and analysis. However, the extent to which its application is possible and accessible varies for different languages. The main goal of this paper is to outline fundamental research on text analytics applied on data in Bulgarian. A review of key research articles in two main directions is provided – development of language resources for Bulgarian and experimenting with Bulgarian text data in practical applications. By summarizing the results of a large literature review, we draw conclusions about the degree of development of the field, the availability of language resources for the Bulgarian language and the extent to which text analytics has been applied in practical problems. Future directions for research are outlined. To the best of the author's knowledge, this is the first study providing a comprehensive overview of progress in the field of text analytics in Bulgarian.*

***Keywords*:** *text analytics, natural language processing, Bulgarian text data, practical applications of text analytics, language resources development.*

1. Introduction

Data analytics has become fundamental for modern business companies. Firms implement analytical tools and solutions to better understand both internal processes and the external business environment – customers, competitors, potential partners, and many other aspects of the vast economic world. Meanwhile, with the opportunities offered by the Internet, the widespread use of smart devices, the growth of social networks, the rise of the Internet of Things (IoT) and the remarkable development of the technological world in general, the amount of data we have is multiplying with each passing day. Globally, much of this data is unstructured. This means that we cannot find it in the well-known and easy to process relational databases. A significant part of this unstructured data comes from the digital world and takes mainly the form of images, videos, and text (journals, books, documents, emails, and any other possible type of human communication). This, in turn, gives rise to scientific progress in important research areas, part of the data science and Artificial Intelligence (AI) fields of study. Machine learning provides us with various ways to learn, process, model, and discover patterns in both structured and

unstructured data, thus making it easier to solve various business and economic problems.

Text analytics (or text mining) is a vast interdisciplinary field that broadly defines many analytical techniques used for intelligent analysis of textual data [45]. It combines methods from the fields of computational linguistics, statistics, artificial intelligence, and information science. Text mining is known since the 20-th century, but it was not until the beginning of the 21-st century that it became more widely used and gained much greater publicity. The reason for this is the development of technology and computing power, allowing much more complex mathematical calculations, and the growing vast amounts of textual information on the Internet. According to an industry study [2], 71% of enterprises claim that "unstructured data was growing "somewhat faster" or "much faster" than other business data". Another industry research [12] reveals that 80% of data will be unstructured by 2025 (e.g., business documents, video, audio, IoT, streaming and geo data). Meanwhile, The Journal of Economic Literature publishes leading research in the field revealing how text can be viewed as "data" for important business applications and economic research [19]. Artificial intelligence is already part of the business and economic landscape and the applications of text analytics as listed in [19] are numerous: stock prices prediction based on news and public sentiments, analysis of the impact of central banks communication on financial markets, measuring economic policy uncertainty, measuring the political orientation of media content, topic analysis in research, politics, law, etc.

Nowadays the importance of text analytics is indisputable. However, the extent to which its application is possible and accessible varies for different languages. An inherent characteristic of textual data is that the tools for its processing and analysis are "language dependent". The latter means that for many types of analytical tasks we need to pay special attention to the characteristics of the text data language under analysis. An integral part of the text analytics field is Natural Language Processing (NLP) which concerns the low-level language processing and understanding tasks [45]. Many NLP tools are required in various practical areas such as sentiment analysis, topic modelling, text clustering, etc. However, the availability of NLP tools and similar language resources is one of the factors determining the existing division between the so called "low-resource languages" and "high-resource languages". Of the latter, not surprisingly in the first place is the English language for which there are countless corpora, NLP tools, specialized software systems for text analysis, etc.

The main goal of this paper is to shed light on the development of text analytics in Bulgaria by reviewing key research in the field. Our main aim is to review fundamental research in two main directions. First, a review of studies aimed at the development of language resources for the Bulgarian language is carried out. Then, we focus on studies aimed at experimenting with Bulgarian text data in some practical business and economic applications. The study contributions include a summary and discussion of the results of a large literature review in the text analytics field for a low-resource language like Bulgarian. Drawn are valuable conclusions about the degree of development of the field, the availability and applicability of language resources for the Bulgarian language and the extent to which text analysis has been

applied on Bulgarian text data in practical business and economic problems. The study is comprehensive without being, or claiming to be, exhaustive. Finally, we discuss possibilities for future development. To the best of the author's knowledge, this is the first attempt to outline key research in the field of text analytics for Bulgarian in the two directions described above.

The rest of the paper is organized as follows. Section 2 provides the overall research methodology applied in the critical review of studies. Section 3 presents major results from the literature review as a structured summary. Section 4 provides an analysis of results and discussion of past and future trends, followed by a conclusion.

2. Research methodology

The field of text analytics deals with a number of technologies for processing and analysis of semi-structured and unstructured text data. As defined by M i n e r et al. [45] all these technologies are combined by the need to "convert text to numbers" so that powerful algorithms can be applied to large databases. In the introduction of the current paper, we mentioned that text analytics is an interdisciplinary field and sometimes there are very different concepts that someone may have in mind when referring to it. M i n e r et al. [45] bring clarity to this vast scientific field by outlining seven main practical areas of text analysis. According to the authors a typical task in the field will involve techniques from several of these areas, as they overlap to a large extent. These seven areas are: search and information retrieval, document clustering, document classification, web mining, information extraction, natural language processing, and concept extraction.

As mentioned earlier, the current review is carried out in two main directions. First, research in the practical area of natural language processing is in focus – we aim to outline the main language resources which are available for Bulgarian. The scope of this task includes various instruments or whole systems developed to ease textual data processing and analysis. Examples of such NLP tools commonly used in text analytics projects are different types of parsers, POS tagging and Named Entity Recognition (NER) tools, stemmers etc. We also pay special attention to the development of text corpora in Bulgarian. A corpus is a large collection of structured documents (texts) usually annotated with syntactic and/or morphological information. This type of linguistic resources is crucial in computational linguistics since many NLP tools can be developed based on the knowledge provided from such annotated corpora. The latter can be also used for carrying out various statistical analysis and experiments in the field. The availability of language resources for Bulgarian is a key prerequisite for progress in the practical application of text analytics on real-world problems since such resources ease the analyses and provide opportunities for experimentation.

The second central point in the current research is the practical application of text analytics to solve various economic or business problems. We aim at outlining key research articles focused on the statistical analysis and experimentation with text data in Bulgarian for solving real-world problems. Since the scope of such a task is

very broad, we apply the definition provided by M i n e r et al. [45] and focus mainly on studies in three practice areas of text analytics which, of course, have many intersections – document clustering, document classification, and information extraction. As its name suggests, document clustering deals with the application of cluster (topic) analysis on textual data, while document classification is about the grouping and categorizing documents, using methods, based on labelled training data [45]. The main task in information extraction is to automatically identify and extract events, facts, and relationships from text – it is about extracting structured information from unstructured data.

The scope of the current survey is not limited to reviewing only the most recent research (for example, in the last five years) – instead, we are focused on including any key research in the field of text analytics in Bulgarian. Since the number of contributions is large, an exhaustive review of all related work cannot be given here. However, setting such a scope enables us to follow and document key trends and developments in the field. When reviewing research focused on the development of NLP tools for Bulgarian we are mainly interested in the applicability of such instruments and whether they are accessible by a web interface, software program or implemented in established software for statistical programming as R or Python. When reviewing works towards the development of text corpora, we are mainly interested in their type (monolingual/bilingual/multilingual), size, text data domain, and level of annotation (if such exists). When reviewing research on the practical application of text analytics we are mainly focused on: the business/economic problems being addressed; text data domain; practice area of text analytics; availability of language resources as datasets, models or programming code for experiments provided as a result from the study.

3. Text analytics applications

The current section provides a review of key research articles in the two main areas described in the previous section. First, we provide a summary of research aimed at development of text corpora in Bulgarian since such language resources could be the basis for the development of NLP tools. Next, we outline the existing tools and systems for common NLP tasks such as parsing, stemming, part-of-speech tagging, etc. Finally, we summarize key research aimed at the practical application of text analytics on Bulgarian textual data for solving real-world problems (i.e., research with more business and economic focus).

3.1. Text corpora

Reviewing the available language resources for Bulgarian, we should first pay special attention to the development of text corpora. As mentioned earlier, these large collections of structured documents, annotated on various linguistic levels can serve as a basis in the development and validation of natural language processing tools. In the development of language resources intended for the Bulgarian language, undoubtedly very great progress in the field has been achieved by the BulTreeBank group. This group of researchers works on various projects in the field of

computational linguistics and is part of the Department of Artificial Intelligence and Language Technologies at the Bulgarian Academy of Sciences (BAS). The main goal of the BulTreeBank group is to create language resources for Bulgarian. One of their very successful projects is the BulTreeBank project (which also gives the name of the group of researchers) which aims at the development of the Bulgarian National Reference corpus (the BulTreeBank corpus). The group has several other projects in the field of NLP. However, the scope of the current paper does not allow a detailed review of each – within this section we will focus only on the development of text corpora in Bulgarian, and in the next section we will mention some of the text processing tools for Bulgarian developed by the BulTreeBank group and other researchers.

Simov, Popova and Osenova [55] present the BulTreeBank project, which aims to create an annotated with detailed syntactic information (in the form of syntactic trees) TreeBank in Bulgarian. At the time the project starts, there is no such corpus in Bulgarian which may be used to enable the development of linguistics theories and the application of NLP techniques. The main goal is, if it is not possible to create a corpus containing all the syntactic structures found in the Bulgarian language, then at least to contain the most common ones, covering the structure of simple sentences in the Bulgarian language. Henceforth, such set can serve as a basis for the development of larger corpora and other linguistic resources. In [53] the main components of the monolingual BulTreeBank corpus are presented – basic set of sentences ("gold standard") analysed manually, Treebank and a corpus of texts in Bulgarian, covering a wide range of domains with a size of 100 million words. The corpus contains not only simple, but also complex sentences [55]. In [56] is described how the main functionalities of the CLaRK system for corpora development are exploited in the BulTreeBank project. The latter is an XML based software for corpora development first introduced a little earlier – in 2001 [54]. Following is a number of publications related to different stages and tasks in the development of the BulTreeBank corpus and a detailed list of all can be found on the BulTreeBank group's official website. Currently, the corpus contains texts in the domains of fiction, newspapers, legal documents, government texts and other. The corpus is being constantly updated with new texts and is annotated at various linguistic levels.

Koeva, Mihov and Tinchev [35] present the architecture and logic behind the development of the Bulgarian Wordnet (BulNet) – a lexical database in Bulgarian, which enables the analysis of semantic relations between words. The semantic relations in Bulgarian Wordnet are based mainly on Princeton WordNet and EuroWordNet. BulNet was created within the European project BalkaNet, and as of 01.01.2020 it consists of more than 85,000 synsets. Koeva, Blagoeva and Kolkovska [33] present the monolingual Bulgarian National Corpus (BulNC). In 2010, the large-scale representative corpus comprised of about 320 million words including text data from various thematic domains has been presented. The corpus contains four main components (sub-corpora) – Bulgarian Brown Corpus, Structured Corpus of Bulgarian electronic documents (2001-2009), Structured Corpus of Bulgarian Printed Editions (1945-2009) and transcripts of conversations. The corpus is morpho-syntactically annotated. The authors emphasize the great practical

applications of the corpus in the development of tools for spelling and grammar, machine translation, classification, and extraction of topics from text. In 2012 the Bulgarian National Corpus has about 469.5 million words. The Bulgarian X-language Parallel corpus is integrated [37] in BulNC. The Bulgarian X-language corpus contains texts in the domains of administration, fiction, science, and media. It contains texts in 33 languages, among which the largest parallel corpus is the Bulgarian-English one.

Aside from monolingual corpora, bilingual and multilingual corpora also play a very important role in the field of linguistics and NLP. Parallel corpora can be used in the development of machine translation systems, NLP tools, semantic analysis, comparative analysis, and other similar tasks. In the study of D i m i t r o v a et al. [17] a Bulgarian corpus created within the MULTEXT-East project is proposed. The latter is a continuation of the MULTEXT project, which incorporates the development of several linguistic resources for languages in Western Europe. The resources developed during the MULTEXT-East project include a parallel corpus, comparative corpus, and lexicons. As a result of the project, an annotated multilingual corpus for seven languages (six from Central and Eastern Europe + English) has been created. The multilingual corpus is freely available for research purposes.

In 2018 the first multilingual argumentative corpus for the Balkan languages (Turkish, Greek, Albanian, Croatian, Serbian, Macedonian, Bulgarian, Romanian) and Arabic has been introduced [58]. Argument extraction has great practical importance, as it could facilitate decision-making and content analysis on social networks, making it a powerful tool in the social and political sciences. The development of tools for argument extraction is most often carried out by applying supervised machine learning on annotated data. However, such data is available mainly in English. To achieve their goal, the authors apply tools for argument annotation in English texts, and then project the information on the other languages of interest by making use of bilingual parallel corpora. News articles are used during development, and the corpus is freely available for research purposes. K o e v a, O b r e s h k o v and Y a l a m o v [36] have developed a corpus of legislative documents in Bulgarian. Its development is a part of a larger project to create a multilingual corpus of legislative documents in seven European countries (the MARCELL project). The corpus consists of a little over 25,000 documents in several categories and its main data source is the Bulgarian State Gazette.

The first wordnet in Bulgarian (BulNet) was presented earlier in the current section [35]. O s e n o v a and S i m o v [48] also propose a lexical database enabling the study of semantic relations between words – BulTreeBank WordNet (BTBWN) for Bulgarian. Authors point out that the existing wordnets for Bulgarian are not publicly available. Since such linguistic resources are very valuable in a number of NLP tasks, the authors focus their research in this area. In [48] the architecture and steps behind the development of BTBWN are presented. The BulTreeBank corpus is utilized in the development of BTBWN. The latter can be freely used and accessed via the Python library “wn”. Since the initial development of BTBWN, there have been research efforts put into its enrichment by integration with the Bulgarian

Wikipedia – [57, 30]. Such integrations would play a significant role in improving important NLP tasks such as word sense disambiguation and relation extraction.

It is not possible to include a detailed review of all the existing literature about corpora development in the current paper. Therefore, without claiming to be exhaustive, here we mention other important research [14, 34, 63, 15, 16, 26].

3.2. Natural language processing tools and systems

In this section we focus on identifying key research aimed at the development of NLP tools and systems for processing text data in Bulgarian. The availability of such tools is a key prerequisite for making progress in the practical application of text analytics in Bulgarian in real-world business and economic problems.

In the previous section we started the review by presenting key research of the BulTreeBank group. The researchers pointed out as a main problem the almost complete lack of Bulgarian language processing tools to use during the development of the BulTreeBank corpus. As a result, they had to create most of the additional tools they need. S i m o v et al. [52] describe the infrastructure of a collection of language processing tools intended for Bulgarian. These resources include a tokenizer, morpho-syntactic analyser, morpho-syntactic disambiguation tool, sentence splitting, chunking, and named-entity recognition grammars. The research has great practical and theoretical contributions. First, it marks the start of linguistic resources development for a less spoken language, for which such resources almost did not exist before – this is of great practical importance for researchers in the field of natural language processing who experiment with data in Bulgarian. The work also draws attention to the strategies that can be employed in the development of such processing tools when it comes to less spoken languages as Bulgarian. A list of all linguistic resources intended for the Bulgarian language and developed by the BulTreeBank group can be found on the official website of the group. Among them are tokenizer and sentence splitter, morpho-syntactic tagger, lemmatization tool, dependency parser, POS taggers for Bulgarian, stopword list, and others. Some of these resources are available to use together or as separate applications.

T a n e v and M i t k o v [60] propose the LINGUA system for natural language processing in Bulgarian. The paper reveals the architecture of the LINGUA system – it contains modules for tokenization, POS tagging, segmentation, and others. The authors do not mention whether the system has been implemented and made accessible. T a n e v [59] proposes “Socrates” – a prototype of question-answering system for the Bulgarian language. The prototype can answer to three types of questions – definitions (“What is/Who is…?”), locations (“Where is…?”) and temporal questions (“When is…?”). When returning a response, the system uses encyclopaedic resources and the Google search engine. The input text data is processed by the LINGUA system.

N a k o v [46] develops BulStem – an inflectional stemmer for Bulgarian. The author’s approach combines machine learning methods and utilization of large morphological dictionary in Bulgarian. The tool is freely available and since 2003 has been used in many studies involving text analysis in Bulgarian. It is implemented in Perl, Java, and Python. M a r i n o v and N i v r e [39] present the first dependency

parser for Bulgarian. The tool is developed and tested using the BulTreeBank corpus. The authors do not mention if the tool is made available to the wider research community. S a v o y [51] also presents a stemming tool for Bulgarian. The author compares his approach and results with those presented in [46]. It should be noted that Savoy's stemming algorithm is intended only for removal of inflectional suffixes (feminine/plural forms and definite article).

G e o r g i e v et al. [20] present their work focused on the task of named entity recognition in Bulgarian. Their approach is tested on news data and entities fall into the following categories – persons, organizations, locations and miscellaneous. The tool is developed and validated using annotated data from the BulTreeBank corpus. The authors use a Conditional Random Fields (CRF) model and compare its performance with different types of features (orthographic, morpho-syntactic variables, etc.). The study is the first to apply a statistical approach for named entity recognition in Bulgarian – at that point of time most of the other research applied rule-based methods. The authors continue their research by publishing results of their experiments with the OpenNLP toolkit [21]. Language resources in OpenNLP are developed for English but can be adapted for other languages if training data is available. For this purpose, the authors use the BulTreeBank corpus and focus on five basic tasks in NLP – sentence detection, tokenization, POS tagging, chunking, and parsing. Results in this study are treated as a baseline for the Bulgarian language.

S a v k o v et al. [50] present the Linguistic Processing Pipeline for Bulgarian – BTB-LPP. The system has the following modules for text processing – tokenization, lemmatization, POS tagging, dependency parsing. Similar to other studies, experiments are carried out on the BulTreeBank corpus. BTB-LPP is implemented and can be accessed through the CLaRK system. G h a y o o m i, S i m o v and O s e n o v a [25] adapt two constituency parsers (Stanford and Berkley parser) for the Bulgarian language. Authors rely on a statistical approach, rather than a rule-based one. Again, the BulTreeBank corpus, which provides a large amount of annotated data, is used for training and validation.

G e o r g i e v et al. [22] present a series of experiments focused on POS tagging in Bulgarian. What distinguishes their work from similar studies is the fact that they use a much larger list of morpho-syntactic tags based on the BulTreeBank corpus. Authors achieve accuracy comparable to that accomplished for English. A r k h i p o v et al. [1] present their work focused on named entity recognition employing the most popular technologies in the field of natural language processing nowadays – Transformer models. Authors focus on four Slavic languages (Russian, Bulgarian, Czech, and Polish) and propose a NER method based on maybe the most widely adapted Transformer model – BERT [13]. As a basis for solving the NER task for several languages, a multilingual BERT model on data extracted mainly from Wikipedia is used. Authors draw valuable conclusions on the performance of BERT models trained for specific tasks and languages. The programming code for experiments is published, as well as the NER model with highest performance.

P o p o v, O s e n o v a and S i m o v [49] present a project for implementation of NLP tools for Bulgarian in the spaCy library, which is a free open-source library for NLP in Python and a leader in the employment of industry-strength solutions in the

field. The current survey reveals that this is the first study explicitly aimed at the implementation of Bulgarian language tools in the Python programming language. The current review brings evidence that such tools for Bulgarian exist, but they "live" on different platforms and are not integrated. A little earlier in the current paper, we presented a similar toolkit for NLP in Bulgarian, which has been implemented in the CLaRK system [50]. The BulTreeBank corpus and BTBWN are used as main sources of training data for each of the processing modules in the proposed toolkit in [49]. The latter includes the following tasks for text processing in Bulgarian – tokenization, lemmatization, POS tagging, dependency parsing, NER, word sense disambiguation. The presented work is in its initial stage. In the same year Marinova et al. [40] present a study aimed at the discovery of new types of entities for NER in Bulgarian. Authors rely on deep learning methods for the detection of inconsistencies in initial annotation and for discovery of new types of objects in text. Main contribution of the research is the enrichment of NLP resources for named entity recognition in Bulgarian. Data with the added two new entity categories ("Events" and "Products") is used by Popov, Osenova and Simov [49] during the implementation of NER tool for Bulgarian in spaCy.

3.3. Practical applications of text analytics

In this section we focus on identifying key research applying text analytics for solving practical business and economic problems. Table 1 summarizes the reviewed articles and provides information for some of their key characteristics – practice area of text analytics, text data domain, provided language resources (publicly available).

Zhikov et al. [68] are the first to propose a method for keyword extraction on a dataset of articles published in a Bulgarian media website. The authors do not explicitly define the task as "topic modelling", but their goal is to extract certain parts of text to describe the main themes of news articles. The business problem being addressed is related to the support of the website Svejo.net in which each published article must contain keywords describing its content briefly. The authors approach the task in two ways – as a supervised and as an unsupervised machine learning problem. The experiments continue in [69] – the authors propose a system based on NLP techniques and aimed at overall optimization of the website. The system has two components – keyword extraction and article categorization into predefined general categories in order to speed up the addition of new articles to the platform. Tanev and Steinberger [61] conduct another study in the news articles domain. The authors create a methodology for event extraction from news headlines in Bulgarian and Czech. Authors claim that this is among the first attempts to perform such task for both the languages. A tool for event extraction is useful for structuring huge amounts of data in the form of news, outlining more interesting articles and article summarization.

The studies of Mihaylov, Georgiev and Nakov [42] and Mihaylov et al. [43] reflect a crucial problem in the political, social, and economic life in Bulgaria and in Eastern Europe in general, namely the delusion of society through paid and false opinions on the Internet by supporters of various political parties. The studies concern the political turmoil in Bulgaria during Oresharski's government – at

that moment it became clear that the Socialist Party was paying the so-called "internet trolls" to express opinions in their support on the Internet. The authors approach the task as a classification problem. In the first study, they assume that a person who is claimed to be a troll by at least 5 people in a discussion forum, is indeed a troll. In the second study, they include information about officially exposed trolls. While in these two studies the analysis is on user level (troll user vs. non-troll user), the authors continue their research in [41] this time experimenting on a "comment level". The study carried out in 2019 aims to create a methodology for the detection of troll versus non-troll comments.

K a p u k a r a n o v and N a k o v [31] are the first to develop a system for fine-grained sentiment analysis of movie reviews in Bulgarian. Usually, studies on sentiment analysis consider the problem as a classification task with three categories – positive, negative, neutral. However, in their fine-grained approach [31] have a target variable with 11 categories describing the reviewers' sentiment from positive to negative. The authors approach the task in three different ways – as classification, regression, or ordinal regression problem. As a result from the study, the authors provide a publicly available dataset of movie reviews in Bulgarian and a sentiment lexicon for the Bulgarian language (based on movie reviews data). M i h a y l o v a et al. [44] are the first to address the problem of community question answering in Bulgarian – study's main aim is to create a method for relevant answers ranking. Data used in experiments is in the form of annotated question-answer pairs generated in the largest Bulgarian online forum (BGMamma). A classification model is applied to assess the relevance of an answer. The authors' approach includes translating the data from Bulgarian into English, as the methodology applied is originally created for text data in English. In addition, domain adaptation is used because the training data sample is small.

H a r d a l o v, K o y c h e v and N a k o v [27] make the first attempt to distinguish between credible and fake news in Bulgarian. They are focused on distinguishing between serious news and news designed to sound humorous but fake. The authors develop a rich set of linguistic, semantic, sentiment and other text features. Since the study is the first to address this important social and economic problem, the authors create a dataset with credible and fake articles from several media websites in Bulgaria. Furthermore, the programming code for experiments is also publicly available. G e o r g i e v a-T r i f o n o v a, S t e f a n o v a and K a l c h e v [24] perform text analysis on customer feedback data. One of the main contributions of their work is the development of publicly available dataset of user reviews in Bulgarian. The dataset is manually annotated with the following categories – compliments, complaints, mixed, suggestions. The authors perform document classification and propose a new approach for text data representation as a combination of the vector space model and pointwise mutual information.

In 2018 the extremely interesting and important problem of fake news detection is addressed again [32]. Similar to opinion trolls, such articles aim to manipulate public opinion in order to achieve political and economic goals and interests. In this category also fall the so-called "click-baits", which are shocking news or content that provokes people to open it. K a r a d z h o v et al. [32] are the first to address the

problem of fake news detection as a task to distinguish between serious news and news that are designed to make the reader believe they are real (as opposed to humorous ones in [27]). Several language resources are created as part of the study (listed in Table 1). The authors use a method for document classification which combines a rich set of hand-crafted features and a deep learning technique for text representation.

Table 1. Studies applying text analytics on practical business and economic problems

No	Research article	Year	Practice area of text analytics	Text data domain	Language resources provided (publicly available)
1	[68]	2012	Document clustering (keyword extraction)	News data (articles in a website)	No
2	[69]	2012	Document clustering (keyword extraction). Document classification.	News data (articles in a website)	No
3	[61]	2013	Information extraction	News data (news titles)	No
4	[42]	2015	Document classification	Publications and user comments in media forum (Dnevnik.bg)	No
5	[43]	2015	Document classification	Publications and user comments in media forum (Dnevnik.bg)	No
6	[31]	2015	Document classification	Movie reviews	Yes. 1. Dataset with movie reviews. 2. Sentiment lexicon for the Bulgarian language
7	[44]	2016	Document classification	Community question answering data (BGMamma)	No
8	[27]	2016	Document classification	News data (articles)	Yes. 1. Dataset with credible and fake news articles in Bulgarian. 2. Programming code for the experiments
9	[24]	2018	Document classification	User reviews for online stores	Yes. 1. Dataset with customer feedback data
10	[32]	2018	Document classification	News data (articles)	Yes. 1. Domain-specific word embeddings and topic models; 2. Fact-checking lexicon; 3. Four lexicons useful for modelling the difference in language use between fake and factual news articles
11	[41]	2019	Document classification	Publications and user comments in media forum (Dnevnik.bg)	No
12	[18]	2019	Document classification	News data (articles)	Yes. 1. Dataset with toxic news articles. 2. Programming code for the experiments
13	[66]	2019	Document classification	Interviews	No
14	[29]	2021	Document clustering (topic modelling)	Chat data (customer support chats)	No

Dinkov, Koychev and Nakov [18] propose a methodology for detection of toxic content in news articles. They create a dataset with news articles in nine categories (eight different toxic categories plus one non-toxic). The toxic categories

include fake news, sensations, hate speech, conspiracies, anti-democratic, pro-authoritarian, defamation, and delusion. This is the first work that captures such a broad scope of toxic news categories in Bulgarian. Despite the data being originally in Bulgarian, it is translated to English by using Google Translate API. However, experiments are carried out and compared for both languages. The authors use a rich dataset of features some of which created by models like BERT, Facebook XML model, Google's Universal Sentence Encoder (or USE), ELMo. As a result of the study, the dataset and programming code for experiments are made publicly available.

V e l i c h k o v, K o y c h e v and B o y t c h e v a [66] present their experiments aimed at the prediction of sport events outcome. The authors' approach consists of the application of NLP techniques on interviews with players made prior to the sport event. The authors test several well-known classical approaches for text classification and extend the experiments by utilizing modern technologies for NLP like the BERT model. Results indicate that BERT models achieve better performance in the task of sport events outcome prediction than models trained on structured data only (for example, player rank, age, etc.). H r i s t o v a [29] presents the first study in which chat data in Bulgarian is being analysed. The aim of the study is to extract the main topics of client interest expressed in customer support chats generated in the contact centre of a large bank. The study provides an overall methodology for text processing and topic modelling of customer support chats in Bulgarian. The analysis of real-world chat data in the banking domain reveals interesting insights into customer behaviour in respect to the beginning of the COVID-19 crisis. The extracted topics are evaluated in terms of their complexity. The latter is used to yield recommendations for effective chatbot development based on the assumption that a good approach for developing a chatbot system is to begin with simple and easier to handle topics and gradually include more complex ones.

3.4. Other practical applications of text analytics – the biomedical domain

Reviewing research focused on the practical applications of text analysis in Bulgarian, we should mention the progress made in the biomedical domain. There are a number of research articles analysing medical text data in Bulgarian. Most of the work in this domain is carried out by researchers at BAS. Since biomedical text mining is a large domain having its own specific characteristics, we briefly review key research articles in a separate section.

B o y t c h e v a [3] presents a prototype of the EVTIMA system, which performs information extraction from patient records (written in free text format) and recognizes various fragments of the text, such as a description of patient's disease. The EVTIMA system aims to cover a range of tasks for automated text analysis, including – structuring patient records, easing the process of side-to-side comparison between them and segmentation of documents. In 2011 the author continues research on diagnosis recognition in patient records, which are in free text format [4]. As not all records include the ICD-10 (International Classification of Diseases revision 10) disease code, this leads to loss of valuable information. This problem requires the usage of NLP techniques to first locate the description of the diagnosis in a given

document and then classify it according to ICD-10 by using machine learning methods. The latter will lead to the extraction of more complete information from medical records. A study with similar main goals, but conducted using different methods, is also presented by G e o r g i e v et al. [23] – they perform disease name recognition in Bulgarian discharge summaries (epicrises). The study identifies several factors that make the task of disease name recognition quite complex.

N i k o l o v a et al. [47] present a business intelligence tool (BITool) for natural language processing and information extraction from outpatient records. The tool extracts and analyses important information about patients' treatment (for example, drug names, dosages, duration of treatment, etc.). The goal of the research project is to accelerate the development of the Register of diabetes patients in Bulgaria by making use of natural language processing techniques and business intelligence. The authors demonstrate the usage of the BITool for a specific problem – identification of patients who have diabetes but have not been diagnosed with this disease. The classification approach used in the study is a combination of rule-based methods and machine learning.

B o y t c h e v a et al. [7] present a study whose main goal is to discover relations between the treatment of a disease and other disorders a patient is suffering from (specifically, how treatment can affect these disorders). The authors present the design of a system applying text analytics techniques on outpatient records. A mixture of methods for frequent pattern mining and frequent sequence mining is applied on data to discover important dependencies and correlations concerning serious diseases. Another important topic in the field of healthcare is tackled in [10], namely reducing the risk factors that lead to the development of serious diseases such as Chronic Obstructive Pulmonary Disease (COPD) and diabetes. Text analytics can be applied on outpatient records to extract direct information on patients' medical condition. This type of data can serve as a source of information about the presence of certain risk factors that could lead to the development of such diseases. The authors make an attempt to extract such risk factors affecting patients by applying association rules to outpatient records. The method described can be used to create a system for early warning and detection of patients who are at risk of developing a serious disease.

A similar approach is used by the same authors in another study [9] in which they apply association rules for information extraction from outpatient records and discussions in healthcare forums – thus, a combination of formal and informal documents in the biomedical domain is being analysed. The methods applied allow to compare the informal expressions of the medical terminology with the formal ones and to study the differences between professional and colloquial medical language. Z h a o [67] presents an innovative approach based on deep learning and natural language processing, which aims at information extraction and normalization of electronic health records. The combination of techniques manages to extract information about the biomarker of patients diagnosed with cancer.

In most of the reviewed studies, experiments have been performed on data from the Bulgarian Diabetic Register, which was created with the help of a collection of outpatient records for over 500,000 patients with diabetes treated in the period

2010-2016 [62]. Since it is not possible to include a detailed review of all the existing literature on biomedical text mining in Bulgarian, without claiming to be exhaustive here, we mention other important research [8, 6, 65, 5].

4. Discussion

Progress in the text analytics field for a specific language, in the first place, largely depends on the availability of data for this language. This is one of the factors that define English as a "high-resource language". Although there are different definitions for a "high-resource" and a "low-resource" language [11], the one that is most often utilized considers whether a given language has large, annotated monolingual/bilingual/multilingual corpora, and whether various linguistic resources and technologies for working with the given language have been developed (the latter largely depends on the former). Of course, the extent to which a language falls into one category or another depends on many factors, the most important of which are rather economic, social, and cultural. Among these factors are how many people speak the language and to what extent they use modern information technologies and the Internet. The latter determines the availability of text data in different domains, as in most studies the main source of data is the Internet (or digitally stored text data, in general). From this point of view, compared to other languages, it could be stated that the Bulgarian language belongs to the category of low-resource languages.

From all that has been said so far, it is clear that for a language like Bulgarian there are barriers other than simply investing more in research and development. In fact, the current study reveals that such research capacity exists, and many efforts have been made by various specialists in the field to develop a range of language resources for Bulgarian. If we glance at the global development of text analytics and more specifically – NLP, we can state that it goes through three main stages. Despite being somewhat shifted in time, the development of NLP for the Bulgarian language follows the same path. Initially, mostly rule-based approaches were used. Such methods despite being easier to understand are time-consuming and prone to human errors and do not generalize well. After this period, the application of statistical approaches to NLP tasks began. Machine learning methods are far more reliable than rule-based approaches since statistical inference is used to interpret and detect patterns in textual data. However, there is one important prerequisite – the availability of data. Such algorithms "learn" by utilizing annotated training data. As mentioned earlier, data in a particular language suitable for a specific NLP task is not always available, while usually manual annotation is expensive in time and resources. The third stage of NLP development occurred in the recent years with focus on transfer learning and utilization of Transformer models [64]. Even before the introduction of Transformer models, there was a tendency to focus more on the application of deep learning methods to text analytics tasks.

Quite naturally the progress in text analytics for Bulgarian follows the same three stages described above. The present review reveals that most commonly applied are methodologies based on machine learning approaches different from deep learning. So, one direction for future research is the utilization of deep learning

techniques in both NLP tools development and applications with more practical focus. However, it should be noted that the last depends to a very large extent on the amount of available data – traditional neural network-based methods require a lot of data. In this regard, the advent of transfer learning may be a workaround solution for such low-resource languages as Bulgarian. Transfer learning allows the utilization of models trained on huge amounts of data which can be fine-tuned for specific tasks and languages. Such fine-tuning can be performed with significantly smaller amount of data. Without doubt another direction for future development is the study of transfer learning and how it can be utilized for solving text analytics problems in Bulgarian. Studies already focused on this hot topic are [18, 28, 66, 65].

The current review was divided into three logical parts on purpose. First, we made a review of corpora development since such language resources enable the effective development of NLP tools for text data processing. Our review reveals that so far, such resources for Bulgarian are few. Among them the main are – two monolingual corpora annotated on different linguistic levels (BulTreeBank corpus and Bulgarian National Corpus), two wordnets (BulNet, BTBWN) and a few bilingual/multilingual parallel corpora. The Bulgarian National Corpus is so far the largest in size corpus in Bulgarian, judging by the officially provided information. The BulTreeBank corpus is characterized by high quality annotation at various linguistic levels and many key research articles use it in the development of different NLP tools for tasks like POS tagging, named entity recognition, dependency, and constituency parsing etc. The domain of texts included in the BulTreeBank corpus, and the Bulgarian National Corpus is diverse. As of February 2021, there are 16 language resources published on LRE Map (Language Resources and Evaluation) for the Bulgarian language, while for English there are 961 resources, for German – 216, for Spanish – 130 [38]. Languages with similar number of resources as for Bulgarian are Romanian, Estonian, Slovenian, Egyptian Arabic and other.

The second focus of the current survey is progress made in the development of NLP tools for processing text data in Bulgarian. The current review reveals that existing language resources cover a range of natural language processing tasks including tokenization, stemming and lemmatization, POS tagging, dependency parsing, NER, word sense disambiguation, constituency parsing. However, not all of these resources are available and easy to use and implement in real industry applications. The work of Popov, Osenova and Simov [49] marks the beginning of research efforts put into the integration of NLP tools for Bulgarian with software like Python. Nowadays, NLP professionals and data scientists in large companies work mostly with software suitable for text processing like Python and popular frameworks for NLP and big data which usually are integrated with Python. Currently, language resources for Bulgarian, if implemented at all, are “living” separately on different platforms/specialized software programs. What can be said for sure is that future developments in the field depend on and should include integration with software such as Python, R or NLP frameworks like spaCy, NLTK, TextBlob, etc. R and Python are used by a wide range of academic and industry professionals. Integration of language resources with such software would help to assess their applicability in industry. Naturally, if such integrations exist, this will

ease practitioners and lead to more experiments and research with a practical focus, which is precisely the third point of the current review. Furthermore, the analysis of NLP tools development for Bulgarian reveals that the application of statistical approaches in experiments starts relatively later if compared to NLP development in general. The latter could be explained by the lack of annotated data at that point of time. Finally, progress in the development of methodologies and tools for processing text data in Bulgarian requires the availability of benchmark models for different tasks. Knowledge about state-of-the art performance will naturally lead to further developments and constant improvement of methodologies and algorithms designed for Bulgarian.

The current survey reveals that main research efforts in the text analytics field for Bulgarian start in the beginning of the 21-st century. Having this in mind, it is quite normal that studies focused mainly on practical text mining applications appear even later – around 2010. Furthermore, Table 1 shows that more than half of the reviewed articles are published in the last 4 years (from 2016 to 2020). In some studies, the research methodology and available linguistic resources simply do not allow application on text data in Bulgarian (or do allow it but at the price of worse results). In such cases, text data is translated from Bulgarian to English – [18, 44]. The current review brings evidence that there are not many research articles focused on the applications of text analytics on practical problems. However, there is a lot of research and progress in the field of biomedical NLP for the Bulgarian language. Mainly traditional machine learning methods are utilized in this domain, so it would be valuable to put more research efforts into experimentation with deep learning methods. Apart from biomedical NLP, there is a clear scientific interest and progress in the detection of toxic and misleading behaviour in community forums and fake news detection. The latter are important economic and social problems.

The current review reveals that almost half the studies focused on practical business/economic problems analyse news data. More should be made in the direction of human behaviour analysis and social networks communication analysis. For example, BGMamma is an online communication platform and a source of historical data about huge volumes of discussions on various topics. This platform provides an interesting opportunity for experimentation in the field of community question answering and answer selection [44], sentiment analysis and topic modelling. Valuable insights into social and economic behaviour could be made by experimenting with data from Twitter and Facebook, for example. More efforts should be put in the direction of sentiment analysis which is extremely valuable in the analysis of political and social events and has applications in financial data analysis. Finally, the present review reveals that most of the research focus mainly on supervised machine learning methods, but efforts should be put also into the utilization of unsupervised techniques such as topic modelling, for example. The larrer also has great practical applications as demonstrated by H r i s t o v a [29]. The future of text analytics in Bulgarian should gradually shift to application of more advanced and resource–rich NLP tasks such as question-answering, summarization, reading comprehension, relation extraction, semantic textual similarity, conversational AI. All these tasks are hot topics in the text analytics field.

5. Conclusion

To the best of the author's knowledge, this is the first study outlining key research and progress in the field of text analytics for Bulgarian. The main aim is to provide an overview and more complete picture of the past and future research trends in the field. Surveying the existing literature and drawing conclusions regarding the progress and future directions in the text analytics field for a low-resource language like Bulgarian is considered as a major contribution of the presented work. The current study achieved its main objectives by providing a comprehensive review of key research articles in two main directions – development of language resources for Bulgarian (corpora and NLP tools for text data processing) and experimenting with Bulgarian text data in practical applications of text analytics. First, a summary of key research articles was provided and then a discussion outlining past, current, and future developments in the field was carried out.

Artificial intelligence is all around us, and it has a strong impact on modern economics and the business world. Text analytics allows machines to process, understand, and draw conclusions on huge amounts of text, which would otherwise be impossible for a human to perform. The field has opened completely new horizons to explore important economic, political, and social problems, as well as to address various business issues and support decision-making. The current study revealed several such problems addressed with the help of text analytics in Bulgarian – website optimization, detection of manipulation trolls deceiving public opinion on the Internet, sentiment analysis of user reviews, fake news detection, customer behaviour analysis and other. In the future, such applications for Bulgarian without doubt will continue to develop and become an integral part of modern business and economic studies.

References

1. Arkhipov, M., M. Trofimova, Y. Kuratov, A. Sorokin. Tuning Multilingual Transformers for Named Entity Recognition on Slavic Languages. – In: Proc. of 7th Workshop on Balto-Slavic Natural Language Processing (BSNLP'19), August 2019, pp. 89-93.
2. 451 Research. Addressing the Role of Unstructured Data with Object Storage. 2018. **https://whitepapers.theregister.com/paper/view/7081/451-research-addressing-the-changing-role-of-unstructured-data-with-object-storage?td=s-uu**
3. Boytcheva, S. Assignment of ICD-10 Codes to Diagnoses in Hospital Patient Records in Bulgarian. – In: Proc. of International Workshop "Extraction of Structured Information from Texts in the BioMedical Domain" (ESIT-BioMed'10), Associated to the 18th Int. Conference on Conceptual Structures (ICCS'10), Kuching, Sarawak, Malaysia, July 2010, pp. 56-66.
4. Boytcheva, S. Automatic Matching of ICD-10 Codes to Diagnoses in Discharge Letters. – In: Proc. of 2nd Workshop on Biomedical Natural Language Processing, September 2011, pp. 11-18.
5. Boytcheva, S. Structured Information Extraction from Medical Texts in Bulgarian. – Cybernetics and Information Technologies, Vol. **12**, 2012, No 4, pp. 52-65.
6. Boytcheva, S., G. Angelova, Z. Angelov, D. Tcharaktchiev. Text Mining and Big Data Analytics for Retrospective Analysis of Clinical Texts from Outpatient Care. – Cybernetics and Information Technologies, Vol. **15**, 2015, No 4, pp. 58-77.

7. B o y t c h e v a, S., G. A n g e l o v a, Z. A n g e l o v, D. T c h a r a k t c h i e v. Mining Clinical Events to Reveal Patterns and Sequences. – In: Innovative Approaches and Solutions in Advanced Intelligent Systems, Springer, Cham, 2016, pp. 95-111.
8. B o y t c h e v a, S., G. A n g e l o v a, Z. A n g e l o v, D. T c h a r a k t c h i e v. Mining Comorbidity Patterns Using Retrospective Analysis of Big Collection of Outpatient Records. – Health Information Science and Systems, Vol. **5**, 2017, No 1, pp. 1-9.
9. B o y t c h e v a, S., I. N i k o l o v a, G. A n g e l o v a. Mining Association Rules from Clinical Narratives. – In: Proc. of International Conference Recent Advances in Natural Language Processing, RANLP 2017, September 2017, pp. 130-138.
10. B o y t c h e v a, S., I. N i k o l o v a, G. A n g e l o v a, Z. A n g e l o v. Identification of Risk Factors in Clinical Texts through Association Rules. – In: Proc. of Biomedical NLP Workshop Associated with RANLP 2017, September 2017, pp. 64-72.
11. C i e r i, C., M. M a x w e l l, S. S t r a s s e l, J. T r a c e y. Selection Criteria for Low Resource Language Programs. – In: Proc. of 10th International Conference on Language Resources and Evaluation (LREC'16), May 2016, pp. 4543-4549.
12. Solutions Review. 80 Percent of Your Data Will Be Unstructured in Five Years. 2019. **https://solutionsreview.com/data-management/80-percent-of-your-data-will-be-unstructured-in-five-years/**
13. D e v l i n, J., M. W. C h a n g, K. L e e, K. T o u t a n o v a. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805, 2018.
14. D i m i t r o v a, L., R. P a v l o v, K. S i m o v. The Bulgarian Dictionary in Multilingual Lexical Data Bases. – Cybernetics and Information Technologies, Vol. **2**, 2002, No 2, pp. 33-42.
15. D i m i t r o v a, L., R. G a r a b í k. Bulgarian-Slovak Parallel Corpus. – In: Proc. of 6th International Conference NLP, Multilinguality SLOVKO'2011, Modra, Slovakia, 2011, pp. 44-50.
16. D i m i t r o v a, L., V. K o s e s k a-T o s z e w a. Bulgarian-Polish Language Resources (Current State and Future Development). – International Journal Cognitive Studies/Études Cognitives, Vol. **13**, SOW, Warsaw, 2013.
17. D i m i t r o v a, L., R. P a v l o v, K. S i m o v, L. S i n a p o v a. Bulgarian MULTEXT-East Corpus– Structure and Content. – Cybernetics and Information Technologies, Vol. **5**, 2005, No 1, pp. 67-73.
18. D i n k o v, Y., I. K o y c h e v, P. N a k o v. Detecting Toxicity in News Articles: Application to Bulgarian. arXiv preprint arXiv:1908.09785, 2019.
19. G e n t z k o w, M., B. K e l l y, M. T a d d y. Text as Data. – Journal of Economic Literature, Vol. **57**, 2019, No 3, pp. 535-574.
20. G e o r g i e v, G., P. N a k o v, K. G a n c h e v, P. O s e n o v a, K. S i m o v. Feature-Rich Named Entity Recognition for Bulgarian Using Conditional Random Fields. – In: Proc. of International Conference RANLP-2009, September 2009, pp. 113-117.
21. G e o r g i e v, G., P. N a k o v, P. O s e n o v a, K. S i m o v. Cross-Lingual Adaptation as a Baseline: Adapting Maximum Entropy Models to Bulgarian. – In: Proc. of Workshop on Adaptation of Language Resources and Technology to New Domains, September 2009, pp. 35-38.
22. G e o r g i e v, G., V. Z h i k o v, P. O s e n o v a, K. S i m o v, P. N a k o v. Feature-Rich Part-of-Speech Tagging for Morphologically Complex Languages: Application to Bulgarian. arXiv preprint arXiv:1911.11503, 2019.
23. G e o r g i e v, G., V. Z h i k o v, B. P o p o v, P. N a k o v. Building a Named Entity Recognizer in Three Days: Application to Disease Name Recognition in Bulgarian Epicrises. – In: Proc. of 2nd Workshop on Biomedical Natural Language Processing, September 2011, pp. 27-34.
24. G e o r g i e v a-T r i f o n o v a, T., M. S t e f a n o v a, S. K a l c h e v. Customer Feedback Text Analysis for Online Stores Reviews in Bulgarian. – IAENG International Journal of Computer Science, Vol. **45**, 2018, No 4, pp. 560-568.
25. G h a y o o m i, M., K. S i m o v, P. O s e n o v a. Constituency Parsing of Bulgarian: Word-vs Class-Based Parsing. – In: Proc. of 9th International Conference on Language Resources and Evaluation (LREC'14), May 2014, pp. 4056-4060.
26. G i o u l i, V., K. S i m o v, P. O s e n o v a. A Parallel Greek-Bulgarian Corpus: A Digital Resource of the Shared Cultural Heritage. – In: Language Technology for Cultural Heritage, Berlin, Heidelberg, Springer, 2011, pp. 99-112.

27. H a r d a l o v, M., I. K o y c h e v, P. N a k o v. In Search of Credible News. – In: Proc. of International Conference on Artificial Intelligence: Methodology, Systems, and Applications, Springer, Cham, September 2016, pp. 172-180.
28. H a r d a l o v, M., I. K o y c h e v, P. N a k o v. Beyond English-Only Reading Comprehension: Experiments in Zero-Shot Multilingual Transfer for Bulgarian. – arXiv preprint arXiv:1908.01519, 2019.
29. H r i s t o v a, G. Topic Modeling of Chat Data: A Case Study in the Banking Domain. – In: AIP Conference Proceedings, Vol. **2333**, March 2021, No 1, 150014.
30. K a n c h e v a, Z., I. R a d e v. Linguistic vs Encyclopaedic Knowledge. Classification of MWEs from Wikipedia Articles. – Cybernetics and Information Technologies, Vol. **20**, 2020, No 4, pp. 125-140.
31. K a p u k a r a n o v, B., P. N a k o v. Fine-Grained Sentiment Analysis for Movie Reviews in Bulgarian. – In: Proc. of International Conference Recent Advances in Natural Language Processing, September 2015, pp. 266-274.
32. K a r a d z h o v, G., P. G e n c h e v a, P. N a k o v, I. K o y c h e v. We Built a Fake News & Click-Bait Filter: What Happened Next Will Blow Your Mind!. – arXiv preprint arXiv:1803.03786, 2018.
33. K o e v a, S., D. B l a g o e v a, S. K o l k o v s k a. Bulgarian National Corpus Project. – In: Proc. of LREC-2010, Valletta, 2010, pp. 3678-3684.
34. K o e v a, S., S. L e s e v a, I. S t o y a n o v a, E. T a r p o m a n o v a, M. T o d o r o v a. Bulgarian Tagged Corpora. – In: Proc. of 6th International Conference Formal Approaches to South Slavic and Balkan Languages, October 2006, pp. 78-86.
35. K o e v a, S., S. M i h o v, T. T i n c h e v. Bulgarian Wordnet-Structure and Validation. – Romanian Journal of Information Science and Technology, Vol. **7**, 2004, No 1-2, pp. 61-78.
36. K o e v a, S., N. O b r e s h k o v, M. Y a l a m o v. Natural Language Processing Pipeline to Annotate Bulgarian Legislative Documents. – In: Proc. of 12th Language Resources and Evaluation Conference, May 2020, pp. 6988-6994.
37. K o e v a, S., I. S t o y a n o v a, R. D e k o v a, B. R i z o v, A. G e n o v. Bulgarian X-Language Parallel Corpus. – In: Proc. of 8ht International Conference on Language Resources and Evaluation (LREC'12), May 2012, pp. 2480-2486.
38. LRE Map.
https://lremap.elra.info/
39. M a r i n o v, S., J. N i v r e. A Data-Driven Dependency Parser for Bulgarian. – In: Proc. of 4th Workshop on Treebanks and Linguistic Theories (TLT'05), 2005, pp. 89-100.
40. M a r i n o v a, I., L. L a s k o v a, P. O s e n o v a, K. S i m o v, A. P o p o v. Reconstructing Ner Corpora: A Case Study on Bulgarian. – In: Proc. of 12th Language Resources and Evaluation Conference, May 2020, pp. 4647-4652.
41. M i h a y l o v, T., P. N a k o v. Hunting for Troll Comments in News Community Forums. – arXiv preprint arXiv:1911.08113, 2019.
42. M i h a y l o v, T., G. G e o r g i e v, P. N a k o v. Finding Opinion Manipulation Trolls in News Community Forums. – In: Proc. of 19th Conference on Computational Natural Language Learning, July 2015, pp. 310-314.
43. M i h a y l o v, T., I. K o y c h e v, G. G e o r g i e v, P. N a k o v. Exposing Paid Opinion Manipulation Trolls. – In: Proc. of International Conference Recent Advances in Natural Language Processing, September 2015, pp. 443-450.
44. M i h a y l o v a, T., I. K o y c h e v, P. N a k o v, I. N i k o l o v a. Finding Good Answers in Online Forums: Community Question Answering for Bulgarian. – In: Proc. of 2nd International Conference Computational Linguistics in Bulgaria, September 2016, pp. 54-63.
45. M i n e r, G., J. E l d e r, A. F a s t, T. H i l l, R. N i s b e t, D. D e l e n. Practical Text Mining and Statistical Analysis for Non-Structured Text Data Applications. Academic Press, 2012.
46. N a k o v, P. BulStem: Design and Evaluation of Inflectional Stemmer for Bulgarian. – In: Proc. of Workshop on Balkan Language Resources and Tools (1st Balkan Conference in Informatics), Thessaloniki, Greece, November 2003.

47. N i k o l o v a, I., D. T c h a r a k t c h i e v, S. B o y t c h e v a, Z. A n g e l o v, G. A n g e l o v a. Applying Language Technologies on Healthcare Patient Records for Better Treatment of Bulgarian Diabetic Patients. – In: Artificial Intelligence: Methodology, Systems, and Applications, Springer, Cham, September 2014, pp. 92-103.
48. O s e n o v a, P., K. S i m o v. The Data-Driven Bulgarian WordNet: BTBWN. – In: Cognitive Studies/Études Cognitives, Vol. **18**, 2018.
49. P o p o v, A., P. O s e n o v a, K. S i m o v. Implementing an End-to-End Treebank-Informed Pipeline for Bulgarian. – In: Proc. of 19th Workshop on Treebanks and Linguistic Theories, 2020, pp. 162-167.
50. S a v k o v, A., L. L a s k o v a, S. K a n c h e v a, P. O s e n o v a, K. S i m o v. Linguistic Processing Pipeline for Bulgarian. – In: Proc. of 8ht International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 2012.
51. S a v o y, J. Searching Strategies for the Bulgarian Language. – Information Retrieval, Vol. **10**, 2007, No 6, pp. 509-529.
52. S i m o v, K., P. O s e n o v a, S. K o l k o v s k a, E. B a l a b a n o v a, D. D o i k o f f. A Language Resources Infrastructure for Bulgarian. – In: Proc. of LREC'04, 2004, Lisbon, Portugal, pp. 1685-1688.
53. S i m o v, K., P. O s e n o v a, M. S l a v c h e v a, S. K o l k o v s k a, E. B a l a b a n o v a, D. D o i k o f f, …, M. K o u y l e k o v. Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank. – In: Proc. of LREC 2002, Canary Islands, Spain, May 2002, pp. 1729-1736.
54. S i m o v, K., Z. P e e v, M. K o u y l e k o v, A. S i m o v, M. D i m i t r o v, A. K i r y a k o v. CLaRK-an XML-Based System for Corpora Development. – In: Proc. of Corpus Linguistics 2001 Conference, 2001, pp. 558-560.
55. S i m o v, K., G. P o p o v a, P. O s e n o v a. HPSG-Based Syntactic Treebank of Bulgarian (BulTreeBank). – In: A Rainbow of Corpora: Corpus Linguistics and the Languages of the World, 2002, pp. 135-142.
56. S i m o v, K., A. S i m o v, M. K o u y l e k o v, K. I v a n o v a, I. G r i g o r o v, H. G a n e v. Development of Corpora within the CLaRK System: The BulTreeBank Project Experience. – In: Demonstrations, 2003.
57. S i m o v, K., P. O s e n o v a, L. L a s k o v a, I. R a d e v, Z. K a n c h e v a. Aligning the Bulgarian Btb Wordnet with the Bulgarian Wikipedia. – In: Proc. of 10th Global Wordnet Conference, 2019, pp. 290-297.
58. S l i w a, A., Y. M a, R. L i u, N. B o r a d, S. Z i y a e i, M. G h o b a d i, …, A. A k e r. Multi-Lingual Argumentative Corpora in English, Turkish, Greek, Albanian, Croatian, Serbian, Macedonian, Bulgarian, Romanian and Arabic. – In: Proc. of 11th International Conference on Language Resources and Evaluation (LREC 2018), May 2018.
59. T a n e v, H. Socrates: A Question Answering Prototype for Bulgarian. – In: Recent Advances in Natural Language Processing III, Selected Papers from RANLP 2003, 2004, pp. 377-386.
60. T a n e v, H., R. M i t k o v. Shallow Language Processing Architecture for Bulgarian. – In: Proc. of 19th International Conference on Computational linguistics COLING'02, Vol. **1**, 2002, pp. 1-7.
61. T a n e v, H., J. S t e i n b e r g e r. Semi-Automatic Acquisition of Lexical Resources and Grammars for Event Extraction in Bulgarian and Czech. – In: Proc. of 4th Biennial International Workshop on Balto-Slavic Natural Language Processing, August 2013, pp. 110-118.
62. T c h a r a k t c h i e v, D., S. Z a c h a r i e v a, G. A n g e l o v a, S. B o y t c h e v a, Z. A n g e l o v, P. M a r i n o v a, …, T. T o m o v. Building a Bulgarian National Registry of Patients with Diabetes Mellitus. – Bulgarian Journal of Social Medicine, Vol. **2**, 2015, pp. 19-21 (in Bulgarian).
63. T i e d e m a n n, J. News from OPUS-A Collection of Multilingual Parallel Corpora with Tools and Interfaces. – Recent Advances in Natural Language Processing, Vol. **5**, October 2009, pp. 237-248.
64. V a s w a n i, A., N. S h a z e e r, N. P a r m a r, J. U s z k o r e i t, L. J o n e s, A. N. G o m e z, …, I. P o l o s u k h i n. Attention is All You Need. – arXiv preprint arXiv:1706.03762, 2017.

65. V e l i c h k o v, B., S. G e r g i n o v, P. P a n a y o t o v, S. V a s s i l e v a, G. V e l c h e v, I. K o y c h e v, S. B o y t c h e v a. Automatic ICD-10 Codes Association to Diagnosis: Bulgarian Case. – In: Proc. of 11th International Conference on Computational Systems-Biology and Bioinformatics (CSBio'20), November 2020, pp. 46-53.
66. V e l i c h k o v, B., I. K o y c h e v, S. B o y t c h e v a. Deep Learning Contextual Models for Prediction of Sport Event Outcome from Sportsman's Interviews. – In: Proc. of International Conference on Recent Advances in Natural Language Processing (RANLP'19), September 2019, pp. 1240-1246.
67. Z h a o, B. Clinical Data Extraction and Normalization of Cyrillic Electronic Health Records Via Deep-Learning Natural Language Processing. – JCO Clinical Cancer Informatics, Vol. **3**, 2019, pp. 1-9.
68. Z h i k o v, V., I. N i k o l o v a, L. T o l o ş i, Y. I v a n o v, G. G e o r g i e v. Theme Extraction in Bulgarian: Experiments in Supervised and Unsupervised Settings. – In: Proc. of CLoBL, 2012.
69. Z h i k o v, V., I. N i k o l o v a, L. T o l o ş i, Y. I v a n o v, B. P o p o v, G. G e o r g i e v. Enhancing Social News Media in Bulgarian with Natural Language Processing. – INFOtheca, Vol. **2**, 2012, No 13, pp. 6-18.

Received: 23.03.2021; Second Version: 11.07.2021; Accepted: 23.07.2021