Svelta Koeva

Natural Language Processing in Bulgaria (from BLARK to competitive language technologies)

1. Introduction

The meaning of the terms *Natural Language Processing* and *Computational Linguistics* can be interpreted in different ways. Linguistics, in contrast to the other sciences, began to use formal methods for description much later. If by "computational" we mean the application of formal methods for the description of linguistic data and the improvement of the accuracy and speed of analysis with the aid of specialized computer programmes, then modern linguistics is computational linguistics in the same way as modern physics, for example, might be called computational physics. Computational linguistics to the extent that we understand it has a wider meaning. In addition to the formal (to be understood as complete and consistent) description of natural language this concept also refers to Natural Language Processing. This means the development, on the one hand, of effective theoretical models and language technologies, while on the other hand – computational applications and systems to enhance the quality and effectiveness of communication at various levels – spelling and grammar checking; machine translation; categorisation and summarisation of documents, searching and extraction of information, transformation of written text into speech and vice versa; and much else. This understanding is synchronous with the definition "Computational linguistics (CL) is a discipline between linguistics and computer science which is concerned with the computational aspects of the human language faculty" (Uszkoreit 2000).

In this paper a brief overview of the history of Natural Language Processing in Bulgaria is presented, as well as a short survey over the basic language resources and some innovative research achievements.

2. The beginnings of Natural Language Processing in Bulgaria

The beginnings of Natural Language Processing in Bulgaria are connected with the Machine Translation in the Mathematical linguistics group led by Prof. Alexander Ludskanov in early 1970. The group began work at the Institute of Mathematics of the Bulgarian Academy of Sciences and developed a research programme devoted to the problems of Russian-Bulgarian machine translation as well as quantitative and statistical studies of Bulgarian language. The Institute of Mathematics and Informatics at the moment includes a Department on mathematical linguistics as well.

At the end of the 1980's a new section was formed – the Laboratory for linguistic modelling – which brought together leading researchers (logicians, mathematicians, linguists) from a range of Bulgarian research institution of the Bulgarian Academy of Sciences and the University of Sofia. Over a short period of time the laboratory won financing for a number of research projects from European institutions: LaTeSLav[1] (1991-1994) – aimed

[1] http://www.coli.uni-saarland.de/projects/lateslav1.html.

at developing a prototype of a grammar checker; BILEDITA[2] (1996-1998) – for the development of bi-lingual electronic dictionaries; GLOSSER[3] (1996-1998) – aimed at supporting foreign language training and others. In 1994 a number of researchers from the laboratory led by Prof. Yordan Penchev established a new unit at the Institute for Bulgarian (Bulgarian Academy of Sciences). In 2003 it was renamed as the Department for Computational Linguistics.

Since 1995 there has been a significant increase in the number of projects supported by European funds and nationally-financed projects, supported mainly by the Fund for Academic Research of the Ministry of Education, Youth and Science. The Multext-East[4] (1995-1997) extension of the previous Multext and EAGLES EU projects provided the Bulgarian language resources in a standardized format with standard mark-up and annotation, and these resources were later expanded and upgraded in the TELRI[5] I and II (Trans European Language Resources Infrastructure 1995-1998/1999-2001) and Concede[6] (Consortium for Central European Dictionary Encoding 1998-2000) projects.

In parallel with this, language resources are being developed at the University of Sofia (for example speech corpora), Plovdiv University (for example, electronic dictionaries), the New Bulgarian University (translation memory resources), South-West University (parallel corpora) and others.

A number of years ago five Bulgarian academic institutions founded a consortium to create and develop an integrated national academic infrastructure for language resources. Bulgarian institutions are also involved in the CLARIN[7] project. Other ongoing projects include those comprised by META-NET,[8] EUROPEANA[9] and ATLAS[10] aimed at developing the basic technologies and standards necessary to make knowledge on the Internet more widely available in the future.

In addition to many other smaller-scale funded projects, the above-mentioned projects have led to the development of competences in the field of Language Technology as well as a basic technological infrastructure of language tools and resources for Bulgarian. As a consequence over the past decade a number of important electronic language resources (dictionaries, corpora, lexical data bases) as well as programmes for their processing (spell checking, information extraction, word sense disambiguation, machine translation, etc.) have been developed.

3. Language resources

Electronic language resources (as well as methods for describing language data) for Natural Language Processing are radically different from traditional methods of working in

[2] http://www.cis.uni-muenchen.de/projects/BILEDITA/leaflet_cover.html.
[3] http://www.let.rug.nl/glosser/.
[4] http://nl.ijs.si/ME/.
[5] http://telri.nytud.hu/.
[6] http://www.itri.brighton.ac.uk/projects/concede/.
[7] http://www.clarin.eu/external/.
[8] http://www.meta-net.eu/meta/about.
[9] http://www.europeana.eu/portal/.
[10] http://kms.atlasproject.eu/index.

linguistics. In order that it can be used in a wide range of computational applications, data within the electronic language resources has to be as complete and consistent as possible and the properties and relations between the units of which it is composed must be explicitly encoded.

The term 'language resources' refers to a large variety of electronic data which includes both written and spoken language forms. Depending on their structure, language resources can generally be divided into corpora, dictionaries (including terminological data bases, thesauri and ontologies), lexical-semantic networks, grammars and language models. The term 'language resources' also refers to large variety of language processing tools (tokenizers, taggers, lemmatizers, parsers and so on). The BLARK (Basic Language Resources Kit) concept was defined in a joint initiative between ELSNET (European Network of Excellence in Language and Speech) and ELRA (European Language Resources Association). BLARK is defined as the minimal set of resources that is necessary to do any precompetitive research and education at all (Krauwer 2003). BLARK includes many different resources, such as (mono- and multilingual) written and spoken language corpora, mono- and bilingual dictionaries, terminology collections and grammars, taggers, morphological analysers, parsers, speech analysers and recognisers, etc. ELDA[11] (Evaluations and Language resources Distribution Agency) elaborated a report defining a (minimal) set of Language resources to be made available for as many languages as possible.

3.1 Corpora

The following definition might be proposed as a compilation of the numerous and varied definitions of corpus: "A corpus is a large collection of language samples presented in such a manner as to allow for computational processing and selected on the basis of certain (linguistic) criteria, in order to represent an adequate language model" (Koeva 2010b, 9).

It could be said that some of the most extensively developed language resources in Bulgaria or for the Bulgarian language are corpora. There is a wide range of data for monolingual corpora and archives which reflect various periods in the development of the Bulgarian language, mainly connected with its current status (for example: Bulgarian National Corpus, BgSpeech[12] collection, BulTreeBank Text Archive, Corpus of Old Slavic Texts from the XIth Century[13] and others).

The Bulgarian National Corpus (Koeva et al. 2009) undoubtedly occupies central place amongst them. The Bulgarian National Corpus project began development at the Institute for Bulgarian of the Bulgarian Academy of Science at the beginning of 2009. The project is aimed at compiling and annotating a very large general corpus representative of the synchronous state of the Bulgarian language. The Bulgarian National Corpus reflects the conditions of the Bulgarian language from the middle of the XXth century (specifically from 1945 – the year of the last orthographical reform in Bulgaria) to the

[11] http://www.blark.org/.

[12] http://www.bgspeech.net/index_en.html.

[13] http://www.hf.ntnu.no/SofiaTrondheimCorpus/.

present day. At this present moment about 10% of the total number of texts are documents published between 1945 and 1989, and 90% are documents published between 1990 and 2011.

At the present moment the Bulgarian National Corpus has more than 420 million words and includes more than 11,000 samples. It is envisaged in the very near future that the volume of the Corpus will exceed 500 million words (1 billion words is an achievable aim).

Every document is accompanied with metadata in XML format containing information relating to: the author (authors) of the text, translator (translators) of the text (in the case of translated works), the year of first publication of the text, number of words in the text, genre category of the text, style and thematic area, text source, data of addition, additional commentaries, etc. The unified description of texts facilitates their processing and grouping in relevant subcorpora on the basis of various criteria (for example, author, date of creation, genre category, etc.). The corpus was automatically processed for sentence borders, part of speech tags, lemma and grammatical features of words, word senses (according to data from the Bulgarian wordnet). Recently shallow parsing is performed by means of detecting of phrase structure and assigning phrase boundaries, labels and heads.

The Bulgarian National Corpus is a language resource of national importance and provides a wide range of possibilities for theoretical and practical applications in a number of areas. Since mid 2009 the Bulgarian National Corpus has been publicly accessible on the Internet.[14]

The annotated corpus contains additional "interpretative and predominantly linguistic information" (EAGLES 1996). Separate levels of linguistic annotation can be defined (Leach 1997, 8-15), for example: morphological, morpho-syntactical, syntactical, semantic and discourse (EAGLES 1996), and annotated corpora are usually associated with more than one level of annotation. A number of Bulgarian annotated corpora should also be mentioned: for parts of speech (POS), word senses and dependency structure.

Bulgarian POS and sense annotated corpora are excerpts from the Bulgarian Brown corpus.[15] In the Bulgarian POS-annotated Corpus (+150,000 words) each word form is annotated by hand with the relevant part of speech and grammatical features, with which it is used in the context, selected from a majority of possibilities from the large Grammar dictionary of Bulgarian (Koeva et al. 2006). In the Sense-annotated corpus (+100,000 words) each lexical unit is linked manually with the most appropriate synonym set from the Bulgarian wordnet (BulNet) (Koeva 2010b). Unlike the bulk of sense-annotated corpora where only (sets of) content words are annotated, in the Bulgarian Sense-annotated corpus[16] each lexical unit has been assigned a sense.

The Dependency part of BulTreeBank represents the syntactic information (based on HPSG) encoded in BulTreeBank. It consists of two sets of sentences: grammar derived examples (1,500) and corpus-derived ones (10,000 sentences) (Osenova/Simov 2004).

[14] http://search.dcl.bas.bg.
[15] http://dcl.bas.bg/Corpus/home_en.html.
[16] http://dcl.bas.bg/semcor/en/.

It gives examples from sentences from Bulgarian grammar textbooks, newspapers, literature and other sources of texts. The main function of the three resources is to serve as training and test corpora in the development of basic programmes for automatic annotation at a morpho-syntactical level (tagger), semantic level (word sense disambiguation tool) and syntactical level (parser) with sufficient accuracy and coverage.

Corpora might contain texts from one language only or more than one language. These are accordingly monolingual and multilingual corpora. Multilingual corpora can be divided into translated (consisting of originals and translated equivalents), parallel corpora (consisting of originals and translated equivalents, sentence (and word) aligned – for example the multi-lingual corpus of documents from the European Parliament JRC-ACQUIS[17]) and comparable corpora (collection of thematically similar texts in one or more languages) – for example news translation on Hristo Botev Bulgarian National Radio.

The Bulgarian-X language parallel corpora already compiled or under development are mainly focused on other Slavic, Balkan and West European Languages. One of the aims of the short-term European SEE-ERA NET project *Building Language Resources and Translation Models for Machine Translation Focused on South Slavic and Balkan Languages* (Tufiş et al. 2009) was to develop parallel corpora for Bulgarian, Greek, Romanian and Slovene plus Czech, English, French and German excerpts from Acquis Communautaire (called SEE-ERA.net Administrative Corpus – SEnAC) and for Jules Verne's novel *Around the world in 80 days* translated into French, German, Spanish, Portuguese, Italian, Romanian, Russian, Serbian, Croatian, Bulgarian, Macedonian, Polish, Slovenian, Hungarian and Greek (called SEE-ERA.net Literary Corpus – SEnLC). The SEnAC resulted in 60,389 translation units, each containing one sentence translated in the 8 languages. The SEnLC total number of segments is 4,409 and the average number of words per language is about 60,000. The selected texts are tokenised, tagged, lemmatised and aligned at the sentence level for both corpora subparts and at the word level for the SEnAC.

In the scope of the project Multext-East the versions of Orwell's novel *Nineteen Eighty-Four* in six languages (Bulgarian, Czech, Estonian, Hungarian, Romanian, and Slovene) were tagged for part-of-speech and aligned to English (Dimitrova et al. 1998). Another project resulted in the development of a bilingual collection of cultural texts in Greek and Bulgarian (Ghouli et al. 2009). The corpus amounts to 700,000 tokens in total (350,000 tokens per language): the literature sub-corpus is about 550,000 tokens, the folklore and legend sub-corpus is about 150,000 tokens.

There are other projects aimed at compiling and processing parallel corpora (targeting Bulgarian as well) – i.e. the RuN Corpus (Grønn/Marijanovic 2010), a parallel corpus consisting (mostly) of Norwegian and Russian texts, extended recently with parallel texts in other European languages including Bulgarian); the Bulgarian-Polish-Lithuanian Corpus (Dimitrova et al. 2009); the ParaSol (Waldenfels 2006), known as the Regensburg Parallel Corpus – a parallel aligned corpus of translated and original belletristic texts in Slavic (Bulgarian among them) and some other languages, etc.

[17] http://langtech.jrc.it/JRC-Acquis.html.

Two basic approaches are implemented in the compilation of the Bulgarian-X language corpora: 1) extracting them from well known multilingual databases of parallel texts available on the Internet, i.e. Acquis Communautaire (Steinberger et al. 2006), and 2) compiling new collections of parallel documents. In the scope of the combination of the two approaches special efforts have been made towards the development of Bulgarian-English-X language parallel corpus. It consists of Bulgarian English parallel fiction texts (34,553,474 words in Bulgarian), European union law documents in 23 languages (30,082,860 words in Bulgarian) and news items in 9 Balkan languages and English (7,056,104 words in Bulgarian). The corpus is aligned at the sentence level, the Bulgarian texts are tagged and lemmatized.

The conclusion that can be drawn from this brief and not complete overview of parallel corpora available, where Bulgarian is one of the languages in focus, is that those corpora are not very extensive; they represent generally administrative or literally texts and they are built from the available texts on the Internet, rather than on a planned strategy for developing a balanced and representative parallel corpus.

3.1 Dictionaries and lexical-semantic networks

Dictionaries are other basic components in Natural Language Processing. Computational dictionaries are different from electronic dictionaries in which words are normally presented as lists of basic forms. The term 'computational' is used to mean a dictionary the format which allows for more complex processing – for example the generation of all word forms relating to a given lemma or the link of a lemma and the relevant grammatical features with a specific word form. The format, structure and content of computational dictionaries are designed to serve the various applications of the Natural Language Processing.

Large morphological dictionaries developed by a number of centres (Institute for Bulgarian, University of Plovdiv, Language Modelling Laboratory) have existed for a long time (Koeva 1998; Totkov et al. 1988; Paskaleva 1997). They allow for the automatic analysis and synthesis of word forms and thus provide the ability to construct a paradigm (all possible forms) of a given word, the recognition of a given form as a part of a paradigm and to ascribe the grammatical features. Some of them are used for the development of spell checkers. For example, applications that have been developed at an academic level for spell checking and hyphenating both for Windows and MacOS, for example ItaEst[18] and MacEst.[19] However, such non-commercial applications despite providing high level functionality for correctness and convenience, cannot be expected to develop quickly on the market. A series of commercial products called Slovnik Plus (spell checker, hyphenator, translation dictionary from and into English, electronic synonym dictionary for Bulgarian) and Slovnik Expert (grammar checker) are offered by Sirma Media.[20] Kirila Korekt 10, a product offering full compatibility with Windows 7 and MS Office (spell checker and hyphenator, grammar checker, stylistic appropriate-

[18] http://www.bacl.org/itaestbg.html.
[19] http://dcl.bas.bg/MacEst.html.
[20] http://www.sirma.com/?Sirma_Media.

ness recommendations, synonym dictionary with added antonyms and search and replace functions based on all forms of a given word) is distributed by BMG Ltd.[21]

Wordnet and FrameNet undoubtedly occupy an important place amongst lexical resources which have been very important for the creation of more complex applications in the area of Natural Language Processing. Wordnet and FrameNet have been successfully used in intelligent information search and information retrieval from documents in different languages, text categorisation and text summarisation, word sense disambiguation, machine translation, as well as in many other Natural Language Processing tasks.

The Bulgarian wordnet (Koeva 2010a) is a lexical-semantic network which nodes are synonym sets (so-called synsets) which contain words or multiword expressions (called literals), while arcs contain semantic, morpho-semantic, derivational and extra-linguistic relations between objects placed within the nodes (Fellbaum 1998). The meaning of the lexical nodes in wordnet is expressed by means of the relations to the other nodes in the network, on the one hand and through the properties of the nodes itself (implicitly through the synonym relation between the literals in the synonym set and explicitly through the interpretative meaning and examples of meaning), on the other. Wordnet is one of the most complete and consistent lexical resources (in comparison the literals in the Bulgarian wordnet are much greater in number than the word list in a standard spelling dictionary), at the same time the synonym sets from different languages are connected by means of inter-language equivalence relations, which are used as a basis for the development of the wordnet multilingual lexical-semantic network, the so called global wordnet. Wordnet combines the qualities of the existing language resources. It contains definitions and examples, like ordinary dictionaries, but also organises synonym sets into a conceptual network by means of the semantic relations which exist between them. At the moment the Bulgarian data base contains more than 33,000 synonym sets. The Bulgarian wordnet is approximately one quarter the size of the English wordnet and is one of the biggest in Europe. The European organisation ELDA disseminates the Bulgarian wordnet.

The Bulgarian FrameNet represents general semantic and language-specific lexical-semantic and syntactic combinatory properties of Bulgarian lexical units (the pairing of a word (either a single word or a multi-word expression) and word sense). The Bulgarian FrameNet database (Koeva 2010c) so far contains unique descriptions of over 3,000 Bulgarian lexical units, approx., one tenth of them aligned with appropriate semantic frames (Ruppenhofer et al. 2006). A lexical entry in Bulgarian FrameNet consists of a lexical unit, a semantic frame from the English FrameNet expressing abstract semantic structure, a grammatical class, defining the inflexional paradigm, a valency frame describing (some of) the syntactic and lexical-semantic combinatory restrictions (an optional component) and (semantically and syntactically) annotated examples.

The unique character of the Bulgarian FrameNet is determined by the fact that it defines classes of lexical units in relation to: their place in a given semantic frame at an inter-language level, their productivity in the formation of diathesis, semantic and syntactic alternations, the expression of general morpho-syntactic characteristics and the description of (combinations of) obligatory and permissible contexts.

[21] http://www.bmg.bg/LiveContent/English.aspx.

With regard to resources such as lexicons, wordnets and framenets in Bulgaria substantial resources have been developed in recent years, although their enlargement and cross-validation are subject to further work.

4. Basic language processing tools

The automatic pre-processing and annotation of texts is a necessary precondition for the majority of Natural Language Processing systems. The identification of word and sentence boundaries in the majority of cases includes the removal of ambiguity in the use of punctuation, i.e. when a given symbol designates the end of a sentence and when not. The tokenization is the process of identifying words, phrases, symbols, or other meaningful elements in a text called tokens (the simplest definition of a token is a sequence of symbols between blanks). Many of the interesting problems in the area of computational linguistics, as well as many of the most important applications for the natural language processing require an automatic system for correct association of words with suitable grammatical categories and their values – a tagger. In the most general terms, tagging (the analysing of words according to parts of speech and the relevant values of their grammatical categories) includes the inputting of ambiguous grammatical information and disambiguation. Usually taggers are associated with tokenizers and sentence splitters. Again, Bulgarian taggers developed by a number of centres (Institute for Bulgarian, University of Plovdiv, Language Modelling Laboratory) have existed for a long time (Koeva 2008; Doychinova/Mihov 2004; Chanev/Krushkov 2006).

Lemmatisation is closely connected with the tagging of parts of speech and consists of ascribing a lemma, i.e. the basic form of inflectional words, to each word in the text after the performance of a morpho-syntactical analysis, as well as the relevant grammatical characteristics which characterise the used form of the word.

In order for a parallel corpus to be useful, it needs to be processed with sentence and word alignment – the process of connecting pairs of words, phrases, terms or sentences in texts from different languages which are translated equivalents. Although there are manually aligned Bulgarian parallel corpora, automatic alignment of parallel corpora is used due to the large volumes of texts (Tufiş et al. 2009).

Recently a word sense disambiguation tool was developed for Bulgarian. The principal application of Bulgarian Sense-annotated corpus is in training and evaluation of a multi-component word sense disambiguation system currently under development. The corpus is used in almost every stage of the system creation and tuning. Currently, it uses 4 independent “weak” classifiers (two knowledge-based and two implementing Hidden Markov Models) and fifth weak classifier assesses the confidence for a particular sense according to its frequency in Sense-annotated corpus. The current version outperforms the calculated random sense baseline by 24 points with an overall precision of ~65% (vs ~40% for random sense).

5. Main areas of applications of language resources

The main areas in which language resources and technologies are applicable are searching and extracting information, categorisation and summarisation, automatic question answering and machine translation as well as speech synthesis and recognition.

Even big search engines like Google do not use all the options for "intelligent" searching, especially for languages like Bulgarian which have a relatively small number of native speakers and relative small amount on texts exposed on the Internet. Jabse.com is a Bulgarian search engine (Jabse is an acronym of: Just Another Bulgarian Search Engine). Jabse uses its own spider to recognize and correctly index various types of documents (including MS Word, Adobe pdf, MS Power point, Flash swf). It can process Cyrillic domains and possesses its own evaluation system to define the importance of pages and terms contained therein on the basis of a range of criteria, including the number of incoming links. Certain Bulgarian portals have crawlers similar to those used by global search engines designed to index sites included within their categories. These portals provide the most accurate search results since their data bases include not only key words in the text description, but also words from the contents of the entire site and pages contained therein. Dir.bg, one of the first and largest web portals in Bulgaria launched a standalone service – Diri.bg. "Diri" (in Bulgarian "дири") is an old word for "search" ("tarsi" – "търси"). This new service is in direct competition with the existing Jabse and claims to have in the order of 50 million pages within its index. It is still to be seen whether Jabse or Diri.bg will develop sufficiently to become a significant factor in the Bulgarian Internet sphere.

The automatic categorisation of documents (in the Internet and specialised archives) can be performed on the basis of various criteria, for example the specific nature of the text, with the help of key words and phrases, but usually these phrases are not sufficiently reliable in themselves. Language processing can be used in automatic categorisation as a basic classification mechanism by providing semantic interpretation. Recently automatic categorisation of documents is provided in the scope of the Atlas[22] project aiming at the development of a platform combining three separate solutions: i-Publisher, that will provide a powerful web-based instrument for creating, running and managing content-driven web sites; i-Librarian that will allow its users to store, organize and publish their personal works, to locate similar documents in different languages, and to obtain easily the most essential texts from large collections of unfamiliar documents, and EUDocLib – a publicly accessible repository of EU documents.

In contrast to information extraction systems the purpose of which is to provide users with an approximate list of search coincidences, a question-answering system must be able to provide its users with specific information relevant to the question asked, rather than a list of close coincidences. Socrates (Tanev 2004) is an online system for question answering in Bulgarian. It searches for definitions, authors, inventors and discoverers, geography, maps, family links and dates. It also offers online demonstration of the functionality of the question answering system.

[22] http://kms.atlasproject.eu/index.

There are many areas of communication in which machine translation can be successfully used: for example access to multi-lingual data bases, the creation of search systems, extraction of information and translation of documents, foreign language training – both in traditional forms and in new forms of distance or electronic learning, in communications: for the translation of electronic messages or other documents wherein the rapid transfer of information is of vital importance, in working with the contents of documents aimed at the automatic definition of the text theme, localisation of description of products for the needs of national and regional markets through the creation of the necessary documentation, and last but not least, in professional translation through the use of translation memory technologies in systems to assist translators, in order to improve and increase the speed of their work, as well as to automate the basic part of the translation process.

Machine translation is particularly challenging for Bulgarian. The rather flexible word order which when combined with the lack of morphological distinction for nominal cases and subject omission is a real challenge for natural language processing of Bulgarian and especially for machine translation.

One of the good examples is WebTrance by SkyCode[23] – a machine translation (MT) system which automatically translates texts, help files, menus, windows and Internet pages from English, German, French, Spanish, Italian and Turkish into and from Bulgarian. Meaning-based translation, rather than word-for-word translation, is a challenge for many people studying a foreign language. The aim of WebTrance is to provide meaningful translation of texts. Provided good adaptation in terms of user-specific terminology and workflow integration, the use of MT can increase productivity significantly.

Bultra[24] is a translation system which translates from English into Bulgarian. The original English texts can be sourced from various areas of knowledge. The advantages are: the creation of its own proprietary lexical data bases: the ability to work with several lexical bases; the inputting of words and expressions which do not need to be translated; and integrated electronic English-Bulgarian dictionary.

The ongoing project iTranslate4[25] will offer not only full coverage of EU languages, but also will provide for each language pair the best quality available at the time and mediates easy transfer to professional translators. Translation service is already available online (the translation will be available from any to any language, in many cases directly or if needed through English).

There also exist individual products with limited functionality in subfields such as speech synthesis and speech recognition. Ciela – a Bulgarian publisher of legal literature has its own system for Bulgarian speech recognition. The system was developed as an academic project based on a corpus of legal texts containing over 200 million words used to compile a dictionary of 450,000 word forms (Mitankin et. al. 2009). On the Bulgarian market, there are a few Bulgarian text-to-speech systems. One of these is SpeechLab 2.0[26] provid-

[23] http://webtrance.skycode.com/?lang=bg.
[24] http://transdict.com/translators/bultra.html.
[25] http://itranslate4.eu/project/index.html.
[26] http://www.bacl.org/speechlab.html.

ed free-of-charge to computer users with visual disabilities. SpeechLab 2.0 (Andreeva et al. 2005) allows non-sighted computer users to work in the Microsoft Windows 98/ 2000/XP/2003 environment. It has a synthesizing speed of approximately 108 words/sec. The speech synthesizing method used is diaphonic concatenation. The speech synthesizer works in Bulgarian and also provides for the correct pronunciation of English words.

6. Conclusions

Due to the volume restrictions of this submission it is not possible to list and compare in any detail the qualities of the existing language resources, technologies and software available for Bulgarian.

To sum up, the results indicate that Bulgarian stands reasonably well with respect to the most basic language technology tools and resources, such as tokenizers, POS taggers, morphological analyzers, reference corpora. However, such a study leads to the following general conclusions: a small number of research centres and companies are involved in the creation of language resources and programmes for their use, but they lack sufficient coordination between them. This has led to the parallel creation of language resources and programmes of one and the same type, such as morphological dictionaries and taggers. But even this fact can be viewed positively as there can be no absolute duplication, i.e. there are variations in the completeness, quality and application. However, there needs to be reliable documentation, accessible results from validation tests, in such a way that future users will be able to choose resources or programmes depending on the specific needs of their developments. The results would be even better if there were capabilities for the standardisation and convertibility of the resources, as well as the link between commercial products and research developments.

From this it is clear that more effort needs to be directed towards the development of resources for Bulgarian as well as into research, innovation, and development. It is also to be hoped that Bulgaria's participation in CESAR[27] and META-NET will make it possible to develop, standardise and make available several important Language resources and thus contribute to the growth of Bulgarian language technology.

7. References

Andreeva, M./Marinov, I./Mihov, S. (2005): SpeechLab 2.0 – A high-quality text-to-speech system for Bulgarian: In: *Proceedings of the RANLP 2005, Borovets, September 2005*. Borovets, 52-58.

Chanev, A./Krushkov, H. (2006): A simple part-of-speech-tagger for Bulgarian. In: *Research and Applied Conference in Mathematics, Informatics and Comuter Science*. Veliko Tarnovo, 195-198.

Dimitrova, L./Ide, N./Petkevic, V./Erjavec, T./Kaalep, H.J./Tufiş, D. (1998): Multext-East: parallel and comparable corpora and lexicons for six Central and Eastern European languages. In: Boitet, C./Whitelock, P. (eds.): *Proceedings of the Joint 17th International Conference on Computational Linguistics, Montréal, Canada, August 1998*. Montréal: Université de Montréal, 315-319.

[27] http://ec.europa.eu/information_society/apps/projects/factsheet/index.cfm?project_ref=271022.

Dimitrova, L./Koseska, V./Roszko, D./Roszko, R. (2009): Bulgarian-Polish-Lithuanian Corpus – Current Development. In: *Proceedings of the RANLP 2009*. *Borovets, Bulgaria, 17 September 2009*. Borovets.

Ghouli, V./Simov, K./Glaros, N./Osenova, P. (2009): A web-enabled and speech-enhanced parallel corpus of Greek-Bulgarian cultural texts. In: *EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*. 35-42.

Grønn, A./Marijanovic, I. (2010): Russian in contrast. In: *Oslo Studies in Language* 2, 1, 1-24.

Doychinova, V./Mihov, S. (2004): High performance part-of-speech tagging of Bulgarian. In: *Proceedings of Eleventh International Conference on Artificial Intelligence: Methodology, Systems, Applications* (*AIMSA-2004*). (= LNAI 3192). 246-255.

EAGLES (1996) = *EAGLES: Recommendations for the morphosyntactic annotation of corpora* (1996). (= *EAGLES Document EAG-TCWG-MAC/R*). Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

Fellbaum, C. (ed.) (1998): *Wordnet: An electronic lexical database*. Cambridge, MA: MIT Press.

Koeva, S. (1998): Gramatichen rechnik na balgarskiya ezik. Opisanie na koncepciyata za organizaciyata na lingvistichnite danni. In: *Bulgarian Language*.5, 49-58.

Koeva, S./Leseva, S./Stoyanova, I./Tarpomanova, E./Todorova, M. (2006): Bulgarian tagged corpora. In: *Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, 18-20 October 2006, Sofia, Bulgaria*. 78-86.

Koeva, S. (2007): Multi-word term extraction for Bulgarian, ACL 2007. In: *Proceedings of the Conference on Balto-Slavic NLP*. 59-66.

Koeva, S. (2010a): Bulgarian Wordnet – current state, applications and prospects. In: Miltenova, A.L. (ed.): *Bălgaro-amerikanski dialozi* (*Bulgarian-American Dialogues*). Sofia: Prof. Marin Drinov Academic Publishing House, 120-132.

Koeva, S. (2010b): Balgarskiyat semantichno anotiran korpus – teoretichni postanovki. In: Koeva, S. (ed.): *Balgarskiyat semantichno anotiran korpus*. Sofia: Institute for Bulgarian Language, 7-42.

Koeva, S. (2010c): Lexicon and grammar in Bulgarian FrameNet. In: Calzolari, N./Choukri, K./Maegaard, B./Mariani, J./Odjik, J./Piperidis, S./Rosner, M./Tapias, D. (eds.): *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (*LREC '10*), *Valletta*. Valletta: European Language Resources Association (ELRA), 3678-3684.

Koeva, S./Blagoeva, D./Kolkovska, S. (2010): Bulgarian National Corpus project: In: Calzolari, N./Choukri, K./Maegaard, B./Mariani, J./Odjik, J./Piperidis, S./Rosner, M./Tapias, D. (eds.): *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (*LREC '10*), *Valletta*. Valletta: European Language Resources Association (ELRA), 3678-3684.

Krauwer, S. (2003): The Basic Language Resource Kit (BLARK) as the first milestone for the Language Resources Roadmap. In: *Proceedings of SPECOM 2003*. Moscow, 8-15. www.elsnet.org/dox/krauwer-specom2003.pdf.

Leech, G. (1997): Introducing corpus annotation. In: Garside, R./Leech, G./McEnery, A.M. (eds.): *Corpus annotation: linguistic information from computer text corpora*. London: Longman.

Mitankin, P./Mihov, S./Tinchev, T. (2009): Large vocabulary continuous speech recognition for Bulgarian. In: *Proceedings of the RANLP 2009*. *Borovets, Bulgaria, 17 September 2009*. Borovets, 246-250.

Osenova, P./Simov, K. (2004): *BTB-TR05: BulTreeBank Stylebook*. (= BulTreeBank Project Technical Report No. 05).

Paskaleva, E. (1997): Bulgarian language resources and tools in joint European initiatives. In: Marcinkevičienė, R./Volz, N. (eds.): *Proceedings of the Second European Seminar of TELRI*. Kaunas: Institut für Deutsche Sprache/VDU, 99-109.

Ruppenhofer, J./Ellsworth, M./Petruck, M.R.L./Johnson, C.R. (2006): *FrameNet II: extended theory and practice.* Berkeley: Unpublished manuscript. http://framenet.icsi.berkeley.edu/book/book.html.

Steinberger, R./Pouliquen, B./Widiger, A./Ignat, C./Erjavec, T./Tufiş, D. (2006): The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation* (*LREC '06*)*, Genoa.* Genoa, 2142-2147.

Tanev, H.T. (2004): Socrates – a question answering prototype for Bulgarian. In: Nicolov, N./Bontcheva, K./Angelova, G./Mitkov, R. (ed.): *Recent advances in Natural Language Processing: selected papers from RANLP 2003*. Vol. 3. Amsterdam: John Benjamins, 377-385.

Totkov, G./Krushkov, H./Krushkova, M. (1988): Formalization of the Bulgarian Language and development of a linguistic processor (morphology). In: *Travaux scientifiques* 26, 3,1, 988 – Mathematique, 301-310.

Tufiş, D./Koeva, S./Erjavec, T./Gavrilidou, M./Krstev, C. (2009): ID 10503 Building language resources and translation models for machine translation focused on south Slavic and Balkan languages. In: Machačová, J./Rohsmann, K. (eds.): *Scientific results of the SEE-ERA.NET Pilot Joint Call*. Vienna: Centre for Social Innovation (ZSI), 37-48.

Uszkoreit, H. (2000): *What is Computational Linguistics?* www.coli.uni-saarland.de/~hansu/what_is_cl.html.

Waldenfels, R. (2006): Compiling a parallel corpus of slavic languages. Text strategies, tools and the question of lemmatization in alignment. In: Brehmer, B./Zdanova, V./Zimny, R. (eds.): *Beiträge der Europäischen Slavistischen Linguistik* (*POLYSLAV*) *9*. München: Sagner, 123-138.