Bulgarian Language Technology for Digital Humanities: a focus on the Culture of Giving for Education[1]

Kiril Simov
IICT-BAS
Sofia, Bulgaria
kivs@bultreebank.org

Petya Osenova
IICT-BAS
Sofia, Bulgaria
petya@bultreebank.org

Abstract

The paper presents the main language technology components that are necessary for supporting the investigations within the digital humanities with a focus on the culture of giving for education. This domain is socially significant and covers various historical periods. It also takes into consideration the social position of the givers, their gender and the type of the giving act (last posthumous will or financial support in one's lifetime). The survey describes the adaptation of the NLP tools to the task as well as the various ways for improving the targeted extraction from the specially designed corpus of texts related to giving. The main challenge was the language variety caused by the big time span of the texts (80-100 years). We provided two initial instruments for targeted information extraction: statistics with ranked word occurrences and content analysis. Even in this preliminary stage the provided technology proved out to be very useful for our colleagues in sociology, cultural and educational studies.

1 Introduction

Language technology can help in the extraction of useful and focused content from domain texts. We have already worked on a number of such tasks related to Digital Humanities. For example, in the eLearning area (enriching learning objects content or positioning the learner against a predefined level of expected knowledge) – see in Monachesi et al., 2006; in iconography (describing the icons with the help of an ontology for a better comparison and typology) – see in Staykova et al., 2011, etc. Such projects are reflected in the creation of our language resources and technologies during the years – see Zhikov et al., 2013, Savkov et al., 2012, and Simov et al., 2004.

In this paper we focus on the culture of giving for education. Our work was part of the national project (2015-2017) entitled *Culture of giving in the sphere of education: social, institutional and personality dimensions*, coordinated by the Institute for the Study of Societies and Knowledge at Bulgarian Academy of Sciences. Our sociology colleagues adopted two main approaches in their survey: a) application of software developed especially for the content analysis of historical documents; b) application of the theory of planned behavior to the study of philanthropy. Our work was part of the former approach but with influence on the latter.

The collected corpus comprises texts with a time span of 80-100 years. The task was to extract relevant information with the help of statistics and content analysis for displaying the tendencies in the area of giving from the perspective of the language/phrasing/terminology, the social and economic

Kiril Simov and Petya Osenova 2019. Bulgarian Language Technology for Digital Humanities: a focus on the Culture of Giving for Education. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 196–204.

context. Thus, the initial steps include: the creation of a specialized corpus, the creation of a web-based concordance tool and presenting useful statistics and content analysis over the corpus, and adaptation of the existing basic NLP tools.

Our work aims towards the ideas presented in Fokkens et al., 2018. The similarities are as follows: we similarly aim at receiving structured data as output from the NLP processing; we encode metadata characteristics that are common for the corpus (birth date, death date, place of birth, gender, names, etc.); we provide help for getting information on various thematic questions, such as the terminology change through the periods, the target groups preferences, the social behaviour of the givers, etc. The differences are as follows: we do not work with digitized biography dictionaries, but with a specialized corpus of givers' wills that include biographical information; we have not progressed yet to cover also prosopographical information, i.e. to measure characteristics of well-defined groups. However, we envisage this task as our future direction. Last, but not least, the NLP pipeline that was used here does not incorporate any Wordnet concepts or semantic roles despite the fact that we have a word sense disambiguation module for processing newspaper data. These modules will be added later for such specific tasks as the one reported in the paper.

The structure of the paper is as follows: section 2 describes the corpus and its processing; section 3 focuses on the statistics and content analysis; section 4 outlines our efforts in linking the named entities in the corpus – people, locations, organizations; section 5 concludes the paper.

2 Corpus and processing

The specialized corpus of giving for education (abbreviated as CoDar) consists of separate documents from the period after the liberation of Bulgaria (from 1878 onward) until the middle of XX century. Since the aim of the sociologists is to investigate the incentives behind the decision to support education as well as the attitude of the donors together with the most significant causes, the resource includes last will documents, various acts of giving – letters, notarized acts of giving, constitutive documents of charity funds and foundations.

The texts have been gathered from various libraries and then – scanned and digitized. They were represented in an XML format. The following types of information were added: metadata, structural and linguistic ones. The *metadata* provides information about: the title of the document and its type (last will, document of giving, etc.), the place and the time of the document emergence; the gender and the social status of the donor/donors. The *structural information* provides the text, divided into paragraphs and sentences. The *linguistic information* provides parts-of-speech, morphosyntactic characteristics and dependency syntactic analysis.

The NLP modules for Bulgarian that have been adapted to the specificities of the corpus are as follows: a tokenizer, a morphological analyzer, a Named Entities recognition and linking module, a lemmatizer and a parser (Savkov et al. 2012). Our state-of-the-art morphosyntactic tagging reaches 97.98 % accuracy (Georgiev et al. 2012). The lemmatization module that depends on the tagging has 95 % accuracy (Savkov et al. 2012), and the dependency parser - 91 % (Simova et al. 2014). Our tagset is rich - it comprises 680 tags[2].

When applied to the specific data, the main problems in the tokenization were related to the proper handling of the abbreviations, especially of titles, named entities, temporal expressions, etc. The morphological analyzer, which is a combination of a morphological dictionary and statistical components, had as its main challenges: rare or archaic words and different orthographical codifications. The lemmatizer depends on the results from the morphological analyzer. We rely on the Bulgarian inflectional lexicon containing near 110 000 lemmas (Popov et al. 1998 and Popov et al. 2003) which is used to map each word form to the lemma given that the grammatical features are predicted correctly. This mapping is almost 100 % accurate. Thus, the main difficulty was the assignment of the word form of a rare word to its lemma. The parser also depends on the previous

[2] http://bultreebank.org/wp-content/uploads/2017/04/BTB-TR03.pdf

steps. Apart from that, the parser had problems with the syntactically different codifications in the contemporary Bulgarian and the texts related to the past times. At this first stage, the parser was used mainly as a source of extracting key phrases. Thus the errors were not viewed as crucial.

The morphological tagging step showed the following situation for the three historical periods (see next section for details on the periodization). The qualitative analysis displayed similar problems for all periods: recognition of dates, names, abbreviations; usage of old case forms, old orthographic forms (including malapropisms, such as '*мие*' (wash) instead of '*ми е*' (to me); '*са*' (are) instead of '*се*' (refl)). In Period 2 in addition come the problematic cases with a wrong gender and a wrong part-of-speech. In Period 3 there appear the wrongly analysed name parts. The quantitative analysis in Table 1 showed the following error rates:

Period	**Error Rate**
Period 1	3.3 %
Period 2	7 %
Period 3	12 %

Table 1: Error rates of morphosyntactic tagging.

At first sight it might seem strange that the error rate of the oldest period is lower than the error rate of the most recent period. It should be vice versa, since the language of the third period is closer to nowadays language. Here, however, the role of the corpus sources has to be taken into account. Many of the documents representing the first and the second period came in a normalized way because they were published in books before the start of the project, while many of the texts from the third period were in their authentic form – they were scanned from archive documents in several large cities in Bulgaria.

3 Statistics and Content Analysis

Two of the most effective ways for observing the behaviour of various words, collocations and phrases are: (a) the statistics over the keywords in some domain and (b) their context-expanding concordances.

On the basis of a frequency analysis over the specialized corpus of giving (CoDar) in comparison to the Bulgarian National Reference Corpus[3] (being a general corpus), frequency lists have been produced for three historical periods: *before 1919* (the Bulgarian Renaissance and the end of the First World War) – 49698 word forms; *between 1919 and 1930* (the period of crisis after the First World War) – 46031 word forms, and *after 1930 to early 1940s* (the years of stability, the Second World War and the first years after 09.09.1944) – 66373 word forms. The content analysis of the corpus showed that a) during all the three mentioned periods the acts of giving aimed at raising the education among the Bulgarian population and it targeted mainly students with modest financial possibilities; b) the texts content reflects the influence of the historical development in Bulgaria during the three periods on the campaigns of giving; c) the orthographic and grammatical style follows the norms that held in the respective period.

For extracting the key words from the corpus we used the program *AntConc* (see Anthony, 2014). We compared the lemmatized version of CoDar corpus with a frequency list constructed on the basis of the lemmatized version of the Bulgarian National Reference Corpus. In this way a list of keywords for each subcorpus was created. The visualization (in form of a word cloud) has been done through the

[3] http://webclark.org/

following web service: https://www.jasondavies.com/wordcloud/. The 'cloud' representations of words for the three periods are outlined on Figures 1, 2 and 3 below.

From the visualizations it can be seen that in the period before 1919 the important keywords are: '*фонд*' (fund), '*училище*' (school), '*ефория*' (board), '*завещание*' (will), '*завет*' (legacy), '*книжа*' (securities), '*лев*' (lev – Bulgarian currency), '*пари*' (money), '*капитал*' (capital); '*София*' (Sofia), '*Букурещ*' (Bucharest). In the period between 1919 and 1930 the following keywords appear: '*банка*' (bank), '*имот*' (estate), '*сума*' (sum), '*съвет*' (council), '*учител*' (teacher); '*Габрово*' (Gabrovo). After 1930 the following words show high frequency: '*гимназия*' (high school), '*година*' (year), '*народен*' (national), '*беден*' (poor). In general, however, the keywords in the three periods are semantically coherent. They belong to the domains of education and finances.

Fig. 1: Keywords for the period *before 1919*.

Fig. 2: Keywords for the period *between 1919 and 1930*.

Fig. 3: Keywords for the period *after 1930*.

In Table 2 below we give the first ten most frequent words from the lists with ranked keywords for the three periods.

While in the first period the concept of *will* dominates, in the second and third one this is the concept of *fund*. As third in the ranking list during the first period comes the Bulgarian word '*училище*' (school), while in the next two periods it is the Bulgarian word '*сума*' (sum). In the first and second periods the fourth place has been taken by the same concept *board*, but with two lexicalizations. In the third period the word is *the secondary school*. In case of the board concept the terminology change can be traced.

Ranking of keywords for the three periods					
Before 1919		*Between 1919 and 1930*		*After 1930*	
Word	**Rank**	**Word**	**Rank**	**Word**	**Rank**
завещание (will)	7.87	*фонд* (fund)	6.42	*фонд* (fund)	7.12
фонд (fund)	4.22	*завещание* (will)	5.73	*завещание* (will)	5.85
училище (school)	3.42	*сума* (sum)	3.67	*сума* (sum)	3.69
ефория (board of trustees)	2.71	*настоятелство* (board of trustees)	3.40	*гимназия* (secondary school)	2.78
имот (property)	2.23	*беден* (poor)	3.11	*беден* (poor)	2.60
сума (sum)	2.19	*училище* (school)	2.43	*просвещение* (education)	2.54
лихва (interest)	2.14	*завещавам* (leave one's will)	2.40	*лихва* (interest)	2.51
МНП (Ministry of national education)	2.14	*лихва* (interest)	2.35	*гимназията* (the secondary school)	2.07
душеприказчици (confessors)	1.93	*гимназия* (secondary school)	1.69	*дарение* (donation)	2.06
завещавам (leave one's will)	1.76	*дарение* (donation)	1.66	*завещавам* (leave one's will)	1.80

Table 2: The first 10 words with the highest rank, presented per period.

The lexicalization of the concept *board* '*ефория*' (ephoria) changes as follows: in the first period it takes fourth position, in the second period – 280th position and in the third period – 602nd position. On the contrary, the other lexicalization '*настоятелство*' comes as 426th in the first period, as 4th in the second period and as 14th – in the third period. In all the periods within the first 10 words there appears the verb '*завещавам*' (leave by will). Thus, as a whole, mainly the terminology has changed, not the content.

Of special interest are the keywords that appear in one or two periods, but not in all three. Such an example is the adjective *беден* (poor). It the second and third periods it takes fifth position of frequency, but it does not appear within the first 10 most frequent words in the first period. Thus, an assumption might be made that the giving after 1919 was oriented exclusively to the poor students. In the period before 1919 also the word '*имот*' (property) has been used frequently. This means that giving through estates was a charity form that was not so popular in the periods after 1919. The data might give hints on the distribution of roles within society. For example, before 1919 the roles of the members of the boards of trustees as well as the executors were more popular, while in the next two periods the role of the legator became the most frequent one.

The concordancing service has been customized on the base of the *webclark.org* concordancer. Several use cases have been tested, such as: finding information about female donors or executors of wills (we got 20 results); finding information about the beneficiaries of the donors (we got around 56 results); finding cases on what the support has been given for (we got 70 results where the preferences concern the schools and then – some specific persons).

Concerning the first use case, the names of the donors can be derived through the keyword *donor* or through the metadata search, or both. It is interesting to get the information about the female donors. Besides their names, there is information about their native towns (Gabrovo, Plovdiv, Tarnovo, Gorna Oryahovitsa, etc.), their birth and death dates in case they are known. If not known, this position is marked as *unknown*. Additionally, the type of giving act is mentioned: last will, charity, letter of interest, etc. It turns out that the charity documents are used more frequently in comparison to last will ones and letters of interest.

Concerning the second use case, the documents include information about the circumstances under which the beneficiary might be deprived from its grant. Such circumstances refer to cases when the student does not behave or gets low grades. Some pre-conditions for the grants might be declared in advance, such as the students to return in Bulgaria after their studies or to work in an appointed area (in education or in church, etc.).

Concerning the third case, the frequent situation is when the main support goes to the primary and secondary schools but the sum interest goes for the living expenses of poor and/or blind children who are hard working and capable in their studies.

Since in the giving act also the place of giving is important, Table 2 shows a list of ranked places that appear within the top 150 selected keywords.

Before 1919		*Between 1919 and 1930*		*After 1930*	
Name	**Rank**	**Name**	**Rank**	**Name**	**Rank**
Букурещ (Bucharest)	1.64	*Чепеларе* (Chepelare)	0.84	*Копривщица* (Koprivshtitsa)	1.45
Търново (Tarnovo)	1.06	*Габрово* (Gabrovo)	0.76	*София* (Sofia)	0.89
Габрово (Gabrovo)	0.89	*София* (Sofia)	0.71	*Шумен* (Shumen)	0.64
Свищов (Svishtov)	0.61	*Неврокоп* (Nevrokop)	0.48	*Търново* (Tarnovo)	0.55
Галац (Galats)	0.49	*Търново* (Tarnovo)	0.46	*Габрово* (Gabrovo)	0.34
Одеса (Odessa)	0.44	*Прилеп* (Prilep)	0.39	*Севлиево* (Sevlievo)	0.31
Браила (Braila)	0.39	*Севлиево* (Sevlievo)	0.30	*Пазарджик* (Pazardzik)	0.30

София (Sofia)	0.38			*Етрополе* (Etropole)	0.21
Карлово (Karlovo)	0.29				

Table 3: The words in first 9 positions (where applicable) with the highest rank, presented per period.

It can be observed that in the first period there appear more cities outside Bulgaria, such as Bucharest, Galats and Braila in Romania and Odessa in Russia. In the second period there appear also cities from South-West Bulgaria and Vardar Macedonia like Nevrokop and Prilep. In the third period the focus is only on cities that are within the boundaries of nowadays Republic of Bulgaria. Another interesting thread is the role of today's capital Sofia. In the first period its rank is 0.38. In the second period it grows to 0.71, while in the third one it is already 0.89. Also, after 1930s the most important towns happen to be the smaller province ones such as Koprivshtitsa and Etropole.

4 Named Entity Annotation and Linked Open Data

All the Named Entities have been annotated in XML format with respect to their categories: Person, Location, Organization, Date, Amount. In this way the actual charity documents have been connected to the biographies of the donors. Thus, we established a connection between the events within donors' biographies and the overall acts of giving. This information will be used in at least two ways: (1) the creation of Linked Open Data datasets interconnected with the existing datasets like DBpedia, GeoNames, etc; and (2) support for better understanding of the culture of giving, motivation for donation, etc.

For each document we manually explicated all the mentions of persons . The metadata includes: the names of the persons who donated the sum, the date of the issue of the document, the place of issue. For each person we recorded events in which they participated: birth – place, date, parents; education, working periods, marriage, etc. Most of the places mentioned in the documents were associated with one or more events of these types. Having this factual information explicitly in the text, we could find relationships between the institution or the place of education and the beneficiary of the giving document. For example, when somebody was born is some place A, but studied in place B and finally worked in place C, through the recording of this biographical information it would become clear why the donation was performed in favor of the school in place A or place B.

Additionally, the data has been encoded as RDF statements in such a way that: (1) if there is an appropriate DBpedia URI for the instance, then we use it; (2) if there is no appropriate DBpedia URI, then we create one for the corresponding entity attempting to resemble DBpedia ones. For the corresponding new instances we selected appropriate ontology classes like *dbp:Person, dbp:Politician, dbp:Village*, *dbp:Location*. If it is a location, but not represented in DBpedia, we searched for an appropriate GeoNames instance and if found, we established an *owl:sameAs statement*.

For the moment our Linked Open Data dataset is relatively small, but we consider it important with respect to the representation of people that played a crucial donor role in the Bulgarian society without having been recorded in the big datasets. The sets of documents are: 89 documents in the first period, 111 documents in the second period and 185 documents in the third period. They contain information for 461 people (some documents are related to more than one person - usually couples). For each of them we have records of information for their names, place of birth, date of birth, university, place of work, place of donation, amount of donation, currency, date and place of death. The statements are a little more that 3200. As mentioned above, we envisage to combine out approach with the micro biographies of Fokkens et al. 2018. This will ensure interoperability with other biographical datasets.

It will be interesting to compare such datasets on European level to check how many of the donors lived in different European countries and what their donating coverage was.

5 Conclusions

The paper presents the specialized corpus in the area of giving for education – CoDar, as well as the basic steps of its processing from both – linguistic and statistical points of view (including word clouds). Since Bulgarian belongs to the morphologically rich languages, the most important step was the morphosyntactic tagging. However, the error analysis showed that not only the historical periods of the language are important but also whether the texts were normalized, or not, and if yes, to which extent.

The ultimate aim of our efforts is to facilitate the extraction of appropriate content that might answer research questions and support objective generalizations within the area of socio-economic humanities. Our future work refers to: cleaning the corpus from errors added during the digitization; standardizing normalization of old and rare words; re-training of the processing modules on the cleaned and normalized data.

The project (as other similar projects before it) has impact on the planning of the developments within the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG. From the perspective of language resources our goal is to integrate the existing ones, to extend them to a reasonable size and to construct new ones. The integration will be done through direct references and annotation according to the appropriate standards, including the LOD technologies. In addition to handling the necessary language resources, our goal will be to construct a knowledge graph of the data and tools represented within CLaDA-BG. The knowledge graph will be populated with entities extracted from existing sources like Bulgarian DBpedia, but also with entities extracted from the Bulgarian National Reference Corpus. It will be further automatically extended with documents from the Bulgarian Web Space, but also with the OCR versions of old documents, newspapers. In addition to entities from textual sources we will provide descriptions of cultural and historical artefacts. In this way the different entities will be contextualized in the sense of their co-occurrence in time and space.

With respect to language technologies behind the standard modules, mentioned here, we will work in the direction of Named Entities Recognition and Identification, Semantic Role Labeling, Event Recognition, and Coreference Resolution. Thus, we will be able not only to process the language structure of new texts, but relate them to the previously processed and represented data.

We believe that the combination of Language Resources, Language Technologies and Semantic Technologies is the only objective way to support successfully the research in the Social Sciences and Humanities.

Acknowledgements

The work reported here is done partially within the Bulgarian National Project: *Culture of giving in the sphere of education: social, institutional and personality dimensions*, Grant DFNI-K02/12. In addition, some of the work has been done within the *Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG*, Grant DO01-164/28.08.2018.

References

Anthony, L. 2014. AntConc (Version 3.4.4w) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Fokkens et al. 2018. Fokkens, A., Ter Braake, S., Ockeloen, N., Vossen, P., Legêne, S., Schreiber, G., De Boer, V. BiographyNet: Extracting Relations Between People and Events. At: arXiv:1801.07073 [cs.CL]

Georgiev G., Zhikov V., Simov K., Osenova P., Nakov P. 2012. Feature-Rich Part-of-speech Tagging for Morphologically Complex Languages: Application to Bulgarian. In: proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon, France. pp 492-502.

Monachesi, P., Lemnitzer, L., Simov, K. Language Technology for eLearning. In: *Innovative Approaches for Learning and Knowledge Sharing. EC-TEL 2006*. Lecture Notes in Computer Science, vol 4227. Springer, Berlin, Heidelberg, 2006, p. 667-672.

Popov, D., Simov, K. and Vidinska, S. 1998. A Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language. (in Bulgarian) Atlantis KL, Sofia, Bulgaria. 927 pages.

Popov D., Simov, K., Vidinska, S. and Osenova, P. 2003. A Spelling Dictionary of Bulgarian Language. (in Bulgarian), Nauka i Izkustvo, Sofia, Bulgaria. 808 pages.

Savkov, A., Laskova, L., Kancheva, S., Osenova, P., Simov, K. *Linguistic Analysis Processing Line for Bulgarian*. Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC 2012), ELRA, 2012, 2959-2964

Simov, K., Osenova, P., Kolkovska, P., Balabanova, E., Doikoff, D. *A Language Resources Infrastructure for Bulgarian*. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), ELRA, 2004, 1685-1688.

Simova, I., Vasilev, D., Popov, A., Simov, K., Osenova P. 2014. Joint Ensemble Model for POS Tagging and Dependency Parsing. In: Proceedings of the First Joint Workshop on Statistical Parsing of Morphologically Rich Languages and Syntactic Analysis of Non-Canonical Languages. Dublin, Ireland. pp 15–25.

Staykova, K., Simov, K., Agre, G., Osenova, P. Language Technology Support for Semantic Annotation of Iconographic Descriptions. In: *Proceedings of the Workshop on Language Technologies for Digital Humanities and Cultural Heritage*, *RANLP 2011*, 2011, p. 51-56.

Zhikov, V., Georgiev, G., Simov, K., Osenova, P. *Combining POS Tagging, Dependency Parsing and Coreferential Resolution for Bulgarian*. Proceedings of RANLP, 2013, 755-762.