The Political Speech Corpus of Bulgarian

Petya Osenova and Kiril Simov

Linguistic Modelling Department, IICT-BAS
Acad. G.Bonchev 25A, 1113 Sofia, Bulgaria
{petya, kivs}@bultreebank.org

Abstract

The paper introduces the Political Speech Corpus of Bulgarian. First, its current state has been discussed with respect to its size, coverage, genre specification and related online services. Then, the focus goes to the annotation details. On the one hand, the layers of linguistic annotation are presented. On the other hand, the compatibility with CLARIN technical Infrastructure is explained. Also, some user-based scenarios are mentioned to demonstrate the corpus services and applicability.

Keywords: Political Speech Corpus, Speaker Annotation, Topic Annotation, Sentiment Annotation

1. Introduction

Political speech has always been in the center of the media discourse. It is known as one of the most manipulative and figurative types of speech. At the same time, it inevitably influences people's language and attitude. All these facts have made the analysis of political speech a priority for the NLP community. The following existing political speech corpora can be mentioned, among others: CORPS, which has been tagged with audience reactions (Guerini et. al, 2008), German Political Speeches Corpus and Visualization (Barbaresi, 2012), Congressional Speech Data (Thomas et. al, 2006), etc.

The language that the politicians use as well as politicians' attitude to a specific topic can be observed with the help of large specialized corpora, which would include: Parliament speech, Interviews, Election speech, Apels, etc.

Within the framework of a national project we focus on modeling the linguistic knowledge and providing related adequate services for the better applicability of the corpus. Thus, we aim at providing a corpus for research in various areas, such as sociology, politology, linguistics, etc. The basic services would be:

1. Observations of words and expressions in a context (concordance), and
2. Retrieval of information on attitude or opinion in the domain of political speech.

Some possible scenarios for relevant research are the following: survey on the political speech based on Politicians' speeches in a synchronous timespan, but in different discourses (parliament, interview, apel, etc.), survey on the political language strategies based on Politicians' speeches in a diachronous timespan (when in office and when in opposition). The usual approaches for achieving the above-mentioned scenarios are as follows:

- Sentiment analysis (positive, negative or neutral attitude towards a fact or person)
- Opinion analysis (a view, attitude, or appraisal on an object from an opinion holder)

Our approach includes considerations on positive/negative attitude and opinion, expressed in the corpus.

The paper is structured as follows: Section 2 outlines the current version of the Political Speech Corpus; Section 3 discusses the parameters of corpus annotation. Section 4 focuses on the types of services that operate over the corpus. Section 5 concludes the paper.

2. The Corpus of Political Speech

The corpus is collected from sources that provide political speech in predominantly textual form. Such sources might include: Parliament debates transcriptions, public institution websites with news sections where the important speeches are provided in transcribed form, also transcriptions of interviews on the web pages of media companies. When such transcriptions are missing, we are transcribing by ourselves. Since this is an expensive and time-consuming process, we rely on relatively simple transcription guidelines. In the process of selecting the corpus content we take into account the typology of communication - direct (parliament debates; interviews; political websites) vs. indirect (articles on politicians).

The focus at this stage has been put on the direct type of communication. Thus, the current processed subcorpora are as follows: Parliament Control Speech, Interviews with politicians as well as Pre-election debates and speeches. The Parliament Control speech includes predominantly data from Parliament Control Sessions from years 2006-2012. The Interviews and Pre-election appearances come from popular public broadcasts on the national TVs and radio.

The data is formatted into a standardized XML encoding and TEI specification, then a classification is performed. Since there are also occurrences of words in a non-conventionalized form, after the NLP processing these words are detected and handled.

The corpus is publicly available at the following link: political.webclark.org, and is part of WebCLaRK.

WebCLaRK (www.webclark.org) is a portal hosting different language services for Bulgarian. The services are grouped into two categories: (1) services providing access to language resources and (2) services providing access to language technologies. In a long term perspective they will include at least the following components:

Language Resources Services: Concordance over plain text; Concordance over annotated text.

Language Technologies Services: Morphological analysis; Lemmatization; Syntactic analysis.

The concordance tool of WebCLaRK is an integration of Lucene full text search system and the CLaRK system. Lucene is used for indexing the corpus, and the CLaRK system is used for the actual concordance application. In order to reduce the size of the resulting concordance documents, the number of examples is limited to 3000, which we hope to be enough for most of the envisaged tasks. The query language provides wildcard symbols. Thus, the users are able to specify complex templates. The result can be downloaded as an HTML document or as an XML document for further observations and/or processing. The system is being further developed to allow also querying over the additional corpus annotations like *source* - parliament debates, interviews; *speakers* - selected by their party or other characteristics. The system is the first one which provides free web services within CLARIN-BG.

At the moment the corpus comprises 15 million tokens, and it has been enriched systematically.

3. Corpus Annotation

The corpus provides two types of information: extralinguistic (contextual, situation-based) and linguistic (encoded into the words, phrases and sentences). Thus, our annotation process has been split into two tasks: 1. Annotation of topics and speakers, and 2. Annotation of linguistic units. The latter task has been further subdivided into two subtasks: a) annotating the corpus with an NLP pipe for Bulgarian and b) annotating the sentiment and opinion expressions.

We adopted two standards for annotation of corpora documents - TEI guidelines for document and paragraph level of annotation and the Text Corpus Format (TCF) which supports a standalone annotation in XML (Hinrichs et al., 2010). TCF provides an inventory for grouping annotations by types and separates them in different groups. Thus, TCF is very easy to be extended with new groups for new types of annotations. TCF is specially developed for interchange of results from language tools within the framework of D-SPIN Project - German part of CLARIN infrastructure. We use TCF as a format for the Bulgarian language pipeline and for that reason the annotations that depend on the results from the pipeline are encoded as extensions of TCF. However, for the manual annotation task, we implemented transformations from an add-in type of annotation for the different groups of annotation in TCF. The two annotation standards are interrelated via XPointer link from TCF documents as well as TEI documents.

3.1. Annotation schema

Our annotation schema comprises several layers of annotation: topics, speaker utterances and linguistic units (morphosyntactically processed data plus sentiment and opinion statements).

Topics and Speaker utterances

The topic annotation is performed manually, since in parliament debates as well as during an interview, topics can change quite often. Thus, performing detailed topic annotation over the whole corpus is an extensive and non-trivial task. For that reason the topic annotation is performed on two levels: *document level* and *sentence level*. A document in the corpus can represent a text on a single topic which usually follows from the structure of the debate or the interview. We do not have an initial list of topics, but ask the annotators to formulate the topic as a list of key words, separated by a semicolon. At the next stage of corpus annotation, we plan to process the created topics by organizing them in a hierarchy. The sentence level topic annotation at the moment is performed only when the sentence is annotated as a sentiment or opinion statement.

The speaker annotation is performed semi-automatically, with minimal post-checking, since there is enough explicit information in the data. In the first step the speakers are marked within the paragraphs, they have appeared in. Then, some propagation is done of the speaker to the other related paragraphs. The speakers can be the chair of the Parliament, the journalist, the interviewed politician, etc. However, at this stage the speakers have been identified only by their names. Their role: being a member of the Parliament, a member of the Government or a party leader/activist has been stored in their meta-profile. In the meta-profile there are details on: the social role of the politician, time history of his political positions, party membership, etc. In TEI annotation we rely on the standard XML element for speaker annotation. In TCF the annotation of the speaker is done by an attribute (@sp) added to the sentence element. Additionally, we put element <ns2:speakers> with children elements <ns2:speaker> with the same attribute and the same value of the attribute, and a textual description of the speaker, if any.

Linguistic units

Concerning the NLP processing, the following modules have been executed: a tokenizer of Bulgarian, which segmented the text into tokens and sentences. The next steps were: a morphosyntactic tagger, a lemmatizer and a dependency parser. All of these steps are combined into an NLP pipe for Bulgarian (Savkov et al., 2012). The pipeline has been developed within EuroMatrixPlus project. In the result, the morphosyntactic annotations, the lemmas as well as the clause boundaries are kept. Our assumption is that the smallest linguistic unit which expresses sentiment or opinion, is the clause. The paragraphs from the original text are also kept together with the related sentences. The annotation is done in XML within the CLaRK system.

Tokens are annotated as <ns2:token> elements with included @ID, @start and @end attributes. The morphosyntactic annotation is done by elements <ns2:tag> with attributes @ID for the tag and @tokID for the corresponding token.

The sentences and clause boundaries are marked by sentence elements: <ns2:sentence> containing references to corresponding tokens. The sentences can be discontinuous in cases of interruptions between several speakers and in cases of parenthetical expressions. Some sentences can be incomplete. These segments are also referred to as sentences although in many cases they are clauses.

The processing is envisaged to support several tasks: 1. For detecting unconventional usages of words; 2. For better recognition of sentiment/opinion linguistic holders, since most of them appear at the level of phrases and clauses; 3. More precise searches by the users.

Sentiment/opinion statements

There is a lot of work done on sentiment/opinion analysis in NLP. One popular way of handling this task is the usage of SentiWordNet (Esuli et. al, 2006) or other sentiment-based lexicon. The role of WordNet in multilingual context when modeling subjective language is considered, for example, in (Maks and Vossen, 2010). The authors discuss the problem of subjectivity ambiguity with respect to the content of synsets.

On methodological level, our vision on sentiment/opinion analysis is based on (Pang and Lee, 2008), since they take into account the role of all the levels of linguistic analyses, and also provide domain-specific considerations. We view the text as consisting of subjective and objective statements. Any of the speakers of a given text can express statements of both kinds. However, the annotators explicitly annotate only the subjective statements. Each subjective statement consists of consecutive range of sentences of the same speaker. Each subjective statement is annotated as element <ns2:statement> with three attributes: @type with possible values subjective (sentiment and opinion) and objective (at the moment not annotated); @attitude with possible value positive, negative and neutral (by default); @topic – topic id. The attribute 'attitude' holds for sentiment. In our schema, opinion refers to modality, i.e. whether the speaker is certain about something, whether he doubts it or does not know. The description of each topic is represented in an additional element <ns2:topic> with the same attribute and a textual content. The annotators are free to add their own topics, but the topic description is kept short by imposing constraints on the length of the description.

Similarly to most of the approaches, we started with a sentiment-based lexicon of Bulgarian. This lexicon was derived from an Explanatory dictionary of Bulgarian, where the meanings of the words have been marked with connotative or register markers, such as: ironically, scornfully, disapprovingly, rudely, figuratively, colloquial, slang, etc. The domain markers were ignored (medical, geological, grammatical etc.). The meanings were combined with respect to their label. Thus, the candidates for negative attitude were gathered together and checked. From 545 candidates around 300 have been confirmed. The rest labels, which are not so transparent (such as colloquial, figurative, etc.) have also been grouped for checking – it has more than 6000 lexical entries. However, this group contains a lot of ambiguities, and thus – was harder for checking. It turned out that 2350 entries have a negative marking, while only 611 lexical entries – a positive one. The following should be noted: Bulgarian explanatory dictionaries happen to mark explicitly mostly the negative meanings in comparison to the positive ones; also, this preliminary seed should be augmented with corpus-based meanings; last, but not least, the domain specific features should be taken into account.

The first probes on part of the corpus showed that speech of parliament control is much more sentiment-oriented in comparison with interviews. Also, the negative expressions are four times more than the positive ones.

We compared the negative entries from the lexicon to the annotation in the corpus. The initial experiments suggest that sentiment expressions are more on phrasal level than on the word one. In our future experiments, when we provide more annotations, this situation might change slightly.

4. Corpus Services and Usage

The services behind the corpus consist of several modules. They provide various possibilities to the user, such as: better aggregated statistics, manipulation over the processed version of the corpus, concordancing and extraction. At the moment, the online services provide concordancing and basic statistics over results from the raw version of the corpus. However, our aim is to gradually develop the web portal, adding annotated versions of corpora as well as processing services.

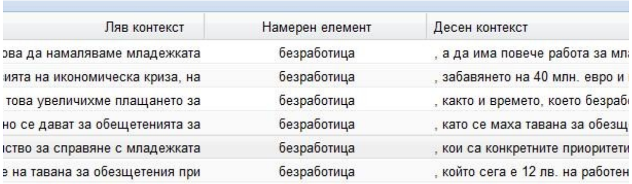
An example of a survey in the service system is the following: some statistics has been made on the frequency of the word occurrences in two subtypes of the political corpora (parliament speech (over 2 mln. tokens) and interviews (about 500 000 tokens)) in comparison with a subpart of a general media corpus (about 70 000 000 tokens) – from the Bulgarian National Reference Corpus. For example, the following Table 1 presents such a survey on the probability of the occurrence of modals (must, want, can) in present tense, all persons and two numbers. It can be seen that the modals *мога* 'can-I' and *можем* 'can-we' are predominant in the interviews in contrast to the parliament speech and the general corpus, while *трябва* 'must' is comparable in the general corpus and the interviews, while not so probable in the parliament speeches.

	INTERVIEW	**%**	**PARLIAMENT**	**%**	**GENERAL CORPUS**	**%**
трябва / must-3p-sg	808	0,1870	2287	0,1008	132428	0,1892
може / can-3p-sg	680	0,1574	1986	0,0875	152472	0,2178
искам / want-1p-sg	245	0,0567	1049	0,0462	13176	0,0188
могат / can-3p-pl	168	0,0389	675	0,0297	49222	0,0703
мога / can-1p-sg	158	0,0366	436	0,0192	18332	0,0262
можем / can-1p-pl	110	0,0255	374	0,0165	13384	0,0191
иска / want-3p-sg	54	0,0125	160	0,0071	19365	0,0277
искат / want-3p-pl	51	0,0118	95	0,0042	11856	0,0169
искаме / want-1p-pl	50	0,0116	157	0,0069	4868	0,0070
искате / want-2p-pl	25	0,0058	157	0,0069	3377	0,0048
можете / can-2p-pl	19	0,0044	191	0,0084	5396	0,0077
можеш / can-2p-sg	9	0,0021	10	0,0004	7301	0,0104
искаш / want-2p-sg	4	0,0009	5	0,0002	4696	0,0067

Table 1: Distribution of modal verbs in the three corpora

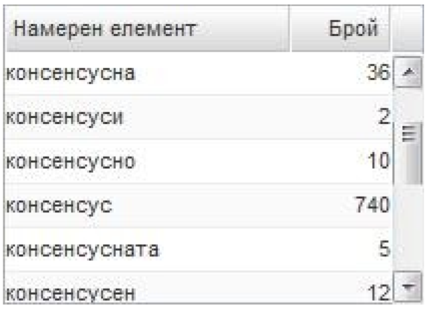
Another statistics shows that (after removing the stop words) the most frequent words in the interviews are: *България* (Bulgaria), *година* (year), *в момента* (at the moment), *въпрос* (question), *хора* (people); in the parliament: *господин* (Mr), *благодаря* (Thank you), *министър* (Minister), *председател* (Chair), *България* (Bulgaria), and in the general media corpus - *България* (Bulgaria), *година* (year), *време* (time), *каза* (said), *София* (Sofia), *вчера* (yesterday).

The concordance web service provides the usage of a word or a phrase in its context. The screenshot below shows the word *безработица* 'unemployment', found in 48 contexts in the mixed corpus (parliament speech + interviews). The found examples are 48. All politicians recognize that there is unemployment and that it has been rising. Also, the topic of unemployment co-occurs with the topics of the retirement reform, health system, re-qualification, etc.



Ляв контекст	Намерен елемент	Десен контекст
ова да намаляваме младежката	безработица	, а да има повече работа за мл
ията на икономическа криза, на	безработица	, забавянето на 40 млн. евро и
това увеличихме плащането за	безработица	, както и времето, което безраб
но се дават за обещетенията за	безработица	, като се маха тавана за обезщ
ство за справяне с младежката	безработица	, кои са конкретните приоритети
е на тавана за обезщетения при	безработица	, който сега е 12 лв. на работен

The statistics tool provides the number of occurrences of each found wordform, based on the concordance query. In this way, not only the most frequent words and phrase can be taken into account, but also the most infrequent ones. For example, in Bulgarian the noun *консенсус* 'consensus' is a relatively new one. If we are interested how its grammatical behavior is viewed by the politicians, we can rely on the concordance tool and the statistics. In the screenshot below it can be seen that the most frequent element is the lemma noun in masculine, singular (740 occurrences). Then come the usages for feminine singular adjective (36 occurrences), masculine singular adjective (12 occurrences) and the rarest one is the plural form of the noun (2 occurrences).



Намерен елемент	Брой
консенсусна	36
консенсуси	2
консенсусно	10
консенсус	740
консенсусната	5
консенсусен	12

5. Conclusions

The reported work is in progress. The first phase of annotation has been finished (topic/speaker and linguistic annotation), and the second one has started (sentiment annotation). Such a corpus has many applications. To start with, after the sentiment annotation, the linguists will survey the corpus for sentiment markers using the concordance and statistics services. Then, various systems for sentiment annotation can be trained. Last, but not least – a corpus-based lexicon with sentiment indicators for Bulgarian will be compiled. In this way, the existing seed lexicon will be extended with new meanings and expressions.

In future, we plan to extend the corpus to cover bigger periods of time. Also, we would like to enrich our speaker database with more factual knowledge about politicians, political parties, locations and events.

6. Acknowledgements

This work has been done within the national project "Center for Investigation of Political and Journalistic speech" – DTK 02/42 from year 2009, funded by the Ministry of Education, Youth and Science, Bulgaria. The authors would like to thank their project co-workers Vladislav Milanov, Nadya Mihaylova and Nadya Terzijska for providing the data, and three anonymous reviewers for their critical comments.

References

A. Barbaresi 2012. *German Political Speeches Corpus and Visualization.* Technical report (http://perso.ens-lyon.fr/adrien.barbaresi/corpora/technical-paper_v2.pdf)

Andrea Esuli and Fabrizio Sebastiani. 2006. *SentiWordNet: A Publicly Available Lexical Resource for Opinion Mining***.** In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*,* Genova, IT, 2006, pp. 417-422.

Erhard Hinrichs, Marie Hinrichs, and Thomas Zastrow. 2010. *Weblicht: Web-based lrt services for German*. In Proceedings of the ACL 2010 System Demonstrations, pages 25– 29, Uppsala, Sweden.

Bo Pang and Lillian Lee. 2008. *Opinion mining and sentiment analysis*. Foundations and Trends in Information Retrieval Vol. 2, No 1-2 (2008) 1–135

Isa Maks and Piek Vossen. 2010. *Modeling Attitude, Polarity and Subjectivity in Wordnet*. In: Proceedings of the 5th Global WordNet Conference (GWC2010), Mumbai, India, January 31-February 4, 2010, Ed. P. Bhattacharya. Fellbaum, P. Vossen, Narosa Publishing House (cover by Sam Vossen), ISBN 979-81-8487.

Aleksandar Savkov, Laska Laskova, Stanislava Kancheva, Petya Osenova, Kiril Simov. *Linguistic Analysis Processing Line for Bulgarian*. In the Proceedings of LREC 2012. Istanbul, Turkey.

Guerini M., Strapparava C. & Stock O. "CORPS: A Corpus of Tagged Political Speeches for Persuasive Communication Processing". Journal of Information Technology & Politics, 5(1): 19-32, Routledge, 2008.

Matt Thomas, Bo Pang, and Lillian Lee 2006. Get out the vote: Determining support or opposition from Congressional floor-debate transcripts. The original version of the paper appeared in the Proceedings of EMNLP, 2006, pp. 327–335.