Bulgarian-English Parallel Corpus for the Purposes of Creating Statistical Translation Model of the Verb Forms. General Conception, Structure, Resources and Annotation.

Todor Lazarov
Department of Computational Linguistics,
IBL-BAS
todorlazarov91@abv.bg

Abstract

This paper describes the process of creating a Bulgarian-English parallel corpus for the purposes of constructing a statistical translation model for verb forms in both languages. We briefly introduce the scientific problem behind the corpus, its main purpose, general conception, linguistic resources and annotation conception. In more details we describe the collection of language data for the purposes of creating the corpus, the preparatory processing of the gathered data, the annotation rules based on the characteristics of the gathered data and the chosen software. We discuss the current work on the training model and the future work on this linguistic resource and the aims of the scientific project.

1. Introduction and brief background on the subject

The current work on the Bulgarian- English parallel corpus for the purposes of constructing a statistical translation model for verb forms in both languages continues previous works on this subject. As it has been previously stated, translating verb forms is very difficult even for human translation – even though the verb systems of both English and Bulgarian share numerous common characteristics, they differ in the manner in which they express the relations between events and points on the temporal axis, the action denoted by the verb and the information about these events. Nevertheless, as we speak about the opportunities of machine translation, both languages are resource rich, which makes theoretical and practical researches about different aspects of them reliable and the gathered data – practical for the purposes of natural language processing and machine translation.

1.1. The difficulties of translating the verb forms from Bulgarian to English

In numerous previous papers on this subject it has been pointed that the main difficulties in the process of translating the verb forms from Bulgarian to English derive from the grammatical characteristics of these languages. The temporal systems of both languages share a common feature – they consist of different grammatical categories within the hyper-category of tense. Without discussing the grammatical peculiarities of Bulgarian and English, we will outline some of the main differences that contribute to qualitative and quantitative dissimilarities. The most tangible difference is the different number of tenses in the discussed languages: while the English tense system consists of 16 structurally dependable morphological tenses, the Bulgarian system has 9 morphological tenses and different lexical categories that can alter the meaning of the tense forms. Both Bulgarian and English have a category that expresses a completed action in relation to a referential point – the perfect tenses. An obvious difference is the presence of continuous tenses in English, which can express an action that is uncompleted related to the referential point, as opposed to Bulgarian where such tenses do not exist. Another tangible difference is that the Bulgarian verbs have lexical aspect, which is part of the semantics of the lexical unit and expresses the action as finished or unfinished related to the action`s

Keywords: verb form, corpus, annotation structure, statistical machine translation, translation model

own completion (Kucarov, 2007:551). Although many linguists would not hesitate to give a positive answer to the question about whether there is aspect in English and would point to the Progressive as example of an aspectual meaning, there are other linguists who would reject the idea of aspect in English at all. More on the differences between English and Bulgarian regarding the category of aspect can be found in Kabakciev (2000). Also word order in English is a decisive factor in distinguishing meaning when we have the same situation, the same participants, but only different position of the elements of the sentence which influences the meaning. Rendering this meaning in Bulgarian is not a problem; we choose the lexical verb and very often in Bulgarian we have to specify the type of action by adding affixes to the verb (Ivanova, 1968, Nedelcheva, 2012). These are only the main tangible differences of both languages' grammatical systems, but the main point is that the Bulgarian language has the possibility to grammaticalize different linguistic data through greater number of grammatical categories in around 2000 verb forms. The greater number of possible grammatical categories, therefore possible grammaticalized meaning, in Bulgarian contributes to high levels of ambiguity during translation, due to the fact that in English the possible grammatical categories are less and the grammaticalized information from Bulgarian as source language needs to be reduced or unevenly distributed between different grammatical categories in English as target language.

1.2. Overview of the proposed solutions and the current work on the problem

Nevertheless, as it has been pointed out before, the characteristics of grammaticalized information in Bulgarian and English verb forms share numerous similarities. While structurally the verb forms in Bulgarian and English can be studied as a specific type of grammatical collocations (Sinapova and Dochev, 1999), other studies (Vassileva, 2003) on Bulgarian and English temporal systems prove in a convincing way that the two languages are different in many respects. However, many linguists note parallels and similarities in the tense system, the categories of aspect and the temporal variations. That is why we have similar grammatical meaning in most of the verb forms that can be formally described and analysed. The current work on analysing and describing in what manner the grammatical information is transferred during translation can be divided in two main approaches, each based on two fundamental methods of machine translation.

1.1.1. Rule-based machine translation of the verb forms from Bulgarian to English

Previous researches on the matter have proposed that the similarities between Bulgarian and English are strong enough for constructing transfer-based rules for the grammatical categories and give a reliable linguistic explanation of how the grammatical information is transferred during translation. Different possible rule-based systems have been described for the purposes of constructing reliable transfer-based rules especially for Bulgarian-English translation. A common feature of the rule-based systems is that they consist of several structural layers that aim at deep formal linguistic comprehension of the language data (Iliev, 2014). Although the rule-based method in machine translation is reliable, as it depends on language models, which are constructed by people (and represent exterior linguistic competence), it is still the human perception of the linguistic phenomena. It is needless to point out that for the rule-based method we need large and accurate grammars and dictionaries, which must take into account all possible language variations. The possibilities of the rule-based method can offer an insightful comparativistic view of the linguistic processes that occur during translation, but they have limited application with regard to describing the complex process of translating the grammaticalized information of the verb forms.

1.1.2. Statistical machine translation and statistical translation models

Incorporating linguistic knowledge into statistical models is an everlasting topic in natural language processing. The last two decades of development in the field of NLP are considered to be the second flourishment of applying statistical methods in the field after the 1980's. Recently a number of machine translation efforts have focused on grammatical formalisms for performing source language analysis, transfer rule application and target language generation. It is worth mentioning several works, such as (Bond et. al, 2005) exploiting DELPH-IN1 infrastructure for developing HPSG grammars; (Riezler and Maxwell, 2006) using LFG grammar; working on a hybrid architecture consisting of an LFG grammar, an HPSG grammar, partial parsing; and using the Functional

Generative Description framework for language analysis on analytical and grammatical level. All the approaches rely on the advances in the development of deep grammar natural language parsing. The approaches share similar architecture and techniques to overcome the drawbacks of the deep processing in comparison to statistical shallow methods. Manually created word aligned bi- or multilingual corpora have proven to be useful resources in variety of tasks, e.g. for the development of automatic alignment tools, but also for lexicon extraction, word sense disambiguation, machine translation, annotation transfer, etc. However, one of the limitations of statistical machine translation is that it only translates words within the context of a few words before and after the translated word. For small sentences, it works pretty well. For longer ones, the translation quality can vary from very good to, in some cases, borderline nonsensical. It is almost always possible to see it has been machine-generated. Nowadays the statistical methods are incorporated into the neural machine translation approaches. Neural Machine Translation (NMT) is an end-to-end learning approach for automated translation, with the potential to overcome many of the weaknesses of conventional phrase-based translation systems. Unfortunately, NMT systems are known to be computationally expensive both in training and in translation inference. Also, most NMT systems have difficulty with rare words. These issues have hindered NMT's use in practical deployments and services, where both accuracy and speed are essential. In the late 2000s, a new machine learning technology called deep learning or deep neural networks, one that tries to mimic how the human brain works (at least partially), became a viable option for many hard to crack computer science problems thanks to advances both on the research side (how to build, train and run these large neural networks) and on the imputer side with the arrival of the extremely large scale computing power of the cloud. For the purposes of this article we will restrict ourselves from further discussion on the characteristics of the different MT systems and circumscribe a general description of statistical translation modelling. The statistical translation models consist of two general components:

- Language models: The goal of statistical language modelling is to build a statistical language model that can estimate the distribution of natural language as accurate as possible. A statistical language model is a probability distribution ***P(s)*** over strings ***s*** that attempts to reflect how frequently a string ***s*** occurs as a sentence. Having a reliable language model is the first step towards building a statistical translation model.
- Translation models: The goal of statistical translation modelling is to represent the probability of a string in a target language to be the translation of a string in the source language. A string of a given language (e) is translated according to the probability distribution ***p(e|f)*** that a string ***e*** in the target language is the translation of a string ***f*** in the source language.

Combining these two components a statistical translation model attempts to calculate the most likely translation of a string ***ê*** of the source language:

$$\hat{e} = argmax_e P(f \vee e) P(e)$$

In this way the probability distribution ***p(e|f)*** is calculated by combining the probabilities of the translation model for the two languages and the language model of the target language. A major benefit of this approach is that it allows the use a language model. This can be very useful in improving the fluency or grammaticality of the translation model’s output.

As they calculate the statistical translation probabilities, statistical translation models directly depend on the quantity and quality of the available linguistic resources. The main principle of this approach is "more data is better data", thus a statistical model of certain language evaluates the probability of certain string of words to appear not by their grammatical correctness, but by the frequency of their usage in the available resources. That is why the first step towards statistical translation modelling is to have sufficient and dependable linguistic corpora with enough language data to ensure that the constructed models based on these resources are reliable and scientifically effective.

2. Resources for the creation of the Bulgarian – English parallel corpus for the purposes of constructing statistical translation model for verb forms

2.1. Requirements and selection of the suitable resources

The step that precedes the creation of the corpus itself is the collection and evaluation of reliable language resources that are suitable for the purposes of the corpus. As it has been stated before (Lazarov, 2016) there are several existing reliable corpora that can provide linguistic data for the purposes of our project. The fundamental requirements of our corpus determine the major characteristics that the available resources must have:

- the language resources must represent parallel Bulgarian-English sentence-aligned texts;
- in its meta-information it must be stated which language is the original language and which is the translation of the original;
- a verified layer of PoS-tags would be beneficial, but not necessary for our needs. The different types of language corpora contain different metadata. Some corpora do not contain information about the morphological characteristics of words, yet they are a valuable resource for monitoring and describing linguistic phenomena and their verification.

Having defined these requirements for the linguistic data that will be included in the corpus, we restrict our choice to the Bulgarian-English Sentence- and Clause-Aligned Corpus (BulEnAC). BulEnAC was created as a training and evaluation data set for automatic clause alignment in the task of exploring the effect of clause reordering on the performance of SMT. The BulEnAC is an excerpt from the Bulgarian-English Parallel Corpus – a part of the Bulgarian National Corpus (BulNC) of approximately 280.8 million tokens and 8.2 million sentences for Bulgarian and 283.1 million tokens and 8.9 million sentences for English. The Bulgarian-English Parallel Corpus has been processed at several levels: tokenization, sentence splitting, lemmatization. The BulEnAC consists of 366,865 tokens altogether. The Bulgarian texts comprise 176,397 tokens in 14,667 sentences, with average sentence length 12.02 words. The English part totals at 190,468 tokens and 15,718 sentences (12.11 words per sentence). The number of clauses in a sentence averages 1.67 for Bulgarian compared with 1.85 clauses per sentence for English. (Koeva et al, 2012).

Another resource that can provide reliable parallel language data is the Bilingual Library [`http://www.bglibrary.net/`], which although it does not provide PoS-tags or meta-information about the texts, includes a sufficient volume of Bulgarian-English parallel texts, which can be included in our corpus.

2.2. Assessment and relevance evaluation of the selected resources

We have to point out that both of the described resources do not meet the preset requirements for them. Although BulEnAC represents a reliable parallel PoS-tagged Bulgarian – English corpus with a sufficient volume, it does not contain information about the source and target language for each of the consisting sub-corpora. For the purposes of our project such information must be subjectively attached to each set of sentences, based on extra linguistic characteristics such as origin of the text, author, its source, etc. Contrasting to that, the Bilingual Library offers parallel texts with information about their source language, author, target language and translator, but it does not contain any linguistic information about the included texts or any alignment. Each of the resources' advantages and disadvantages were taken into account when the structure of the Bulgarian-English parallel corpus for the purposes of creating statistical translation model for verb forms was constructed. The corpus is constructed of small pieces of both resources, which were evaluated and selected after reviewing not only the linguistic information they can provide, but also the meta-linguistic. The approved resources span from particularly selected single sentences to entire coherent texts from different sources – such as news, short narratives, drama pieces and other literary works. The meta-information of the corpus includes data about the source (file name or URL) of each sub-partition of it, the source and the target language, the date of collection and the date of incorporation, the fact that a sub-partition is part of the BulEnAC provides information about whether it had a PoS-tag layer or not. The PoS-tag layer of

BulEnAC is used for correction and confirmation after both of the annotation layers of our corpus have been implemented.

3. Annotation of the corpus

3.1. The first annotation layer - principles and selected tools

During the phase of collection and evaluation of the appropriate resources for the corpus the problem about its annotation structure arose. The used linguistic material did not have equally distributed quality and quantity of annotation layers therefore the general annotation structure was constructed.

The linguistic data in the corpus has two layers of annotation. The first layer is the PoS-tags layer and it consists of PoS-tags of the words. For both languages the tool TreeTagger is used. The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid (1995) in the TC project at the Institute for Computational Linguistics of the University of Stuttgart. Because the TreeTagger is adaptable to other languages if a lexicon and a manually tagged training corpus are available, it was chosen to be trained on the training data, obtained from the corpus after its initial manual second layer annotation.

The gathered linguistic data was first divided in small working files and, where needed, aligned sentence by sentence. Each aligned pair of sentences in Bulgarian and English receives a unique identifying number in order to be recognizable in the subsequent work. After the initial process of dividing and aligning the data, the TreeTagger is used to annotate both languages. For the tagsets used by the TreeTagger see: Santorini (1991) for English and Simov et al. (2004) for Bulgarian. After the process of annotation the data is manually checked and corrections are applied where needed. The annotated working files are separated for each language and meta-information is added to them. Each file receives an ID number in and it is saved as three column tsv (tab-separated values) file. The first column of each line of the file contains the word/token, the second column represents the lemma of the word and the third column is the prescribed PoS-tag. A blank line represents the sentence boundary.

3.2. The second annotation level - structure and tagset

For the second layer of annotation the tool WebAnno (Yimam et al, 2014) is used. WebAnno is a general purpose web-based annotation tool for a wide range of linguistic annotations including various layers of morphological, syntactical, and semantic annotations. Additionally, custom annotation layers can be defined, allowing WebAnno to be used also for non-linguistic annotation tasks. Different modes of annotation are supported, including a correction mode to review externally pre-annotated data, and an automation mode in which WebAnno learns and offers annotation suggestions. WebAnno accepts several file formats, but for the purposes of our project the CONLL file format was chosen. WebAnno uses a revised version of the CoNLL-X format. Annotations are encoded in plain text files (UTF-8, using only the LF character as line break, including an LF character at the end of file) with three types of lines: Word lines containing the annotation of a word/token in 6 fields separated by single tab characters; Blank lines marking sentence boundaries; and Comment lines. Sentences consist of one or more word lines, and word lines contain the following fields:

- ID number of the sentence – the prescribed unique number of the sentence
- ID number of the word/token – the length of the word/token marked by the initial character and the ending character.
- The word/token
- PoS tag – the first annotation layer
- The verbal tag – the second annotation layer
- Numerical relation between the two annotation layers – which elements of the first annotation layer are included in the second annotation layer.

For examples of the file format see Appendix A.

The second annotation layer is done manually through the WebAnno tool. The tagset of this layer consist of smaller number of possible tags than the first annotation layer. They can be staged over the first

layer. The WebAnno tool treats these entities as chunks, which means that a single tag can be prescribed to more than one entity from the first layer. The tagset of the second annotation layer is presented in Table 1.

Bulgarian		English	
Vaor	Verbal form in Aorist	Vprs	Verbal form in Present Simple
Vfutexact	Verbal form in futurum exactum	Vps	Verbal form in Past Simple
Vfutexpraet	Verbal form in futurum exactum praeteriti	Vfs	Verbal form in Future Simple
Vfutpraet	Verbal form in futurum praeteriti	Vprp	Verbal form in Present Perfect
Vfutur	Verbal form in Futurum	Vpp	Verbal form in Past Perfect
Vimperf	Verbal form in Imperfect	Vfp	Verbal form in Future Perfect
Vperfect	Verbal form in Perfect	Vprc	Verbal form in Present Continuous
Vplusqperf	Verbal form in plusquamperfect	Vpc	Verbal form in Past Continuous
Vpraesens	Verbal form in Praesens	Vfc	Verbal form in Future Continuous
		Vprpc	Verbal form in Present Perfect Continuous
		Vppc	Verbal form in Past Perfect Continuous
		Vfpc	Verbal form in Future Perfect Continuous
		Vfsp	Verbal form in Future Simple in the Past
		Vfpp	Verbal form in Future Perfect in the Past
		Vfcp	Verbal form in Future Continuous in the Past
		Vfpcp	Verbal form in Future Perfect Continuous in the Past

Table 1: Tagset of the second annotation layer

The targeted volume of the training data is 1,000 aligned sentences with the two layers of annotation. Since this layer of annotation is manually done by a single person, the pre-defined annotation conventions with an extended and elaborate tag-set will be made available and published later on after this stage of the annotation process is finished.

4. Current work and evaluation of the working process

The working process on the corpus can be divided in three major stages: collection and evaluation of the linguistic material; annotation of training data; and correction and evaluation of automatically annotated data. The current working flow is concentrated on the second stage. The manual annotation of the targeted volume of the training data appears to be the most time consuming stage of the working process and the most problematic. The assessment of the encountered problems and issues during the first two stages of our current work can be divided as follows:

- Pros:

1. The choice of both tools – the TreeTagger and WebAnno brought most of the positives to the working process. The fact that the TreeTagger provides already established set of annotation conventions provided the opportunity to reuse large sets of already annotated data and thus reduce the technical time needed for annotation and manual correction of the gathered linguistic data. Another contribution of the TreeTagger is that it can be trained on user predefined tagsets which allows alternating the used tagsets at any given point of the working process and creating new unique ones entirely for the purposes of this project.

The other main tool of the project - the WebAnno annotation tool also provides numerous opportunities for working with the collected data and versatile functions that meet the initial needs. One of the most beneficial features of the tool is that it offers import and export of the datasets in more than 12 different working file formats that are suitable for different purposes. The tool also provides different annotation levels which can be independent or logically bounded. In the case of the current work on the corpus the greatest advantage of the tool is its ability to perform a predictive annotation. The predictive annotation of the tool prescribes tags on the subsequent language data with certainty based on the previous occurrences of the tag. It can be tuned to be context-sensitive or subordinate constructed. This feature of the WebAnno annotation tool provides the opportunity to annotate the identical tokens in massive sets of data more efficiently and with fewer errors. It is also applicable for the process of manual correction since it offers the possibility to calculate inconsistencies between the automatically assigned tags.

2. The choice of language material also contributes to the efficiency of the working process and the achieved results. The fundamental requirement for the training data is to be representative. This means that the gathered data must demonstrate all of the studied language phenomena in a variety of contexts. The inner structure and the meta-information of BulEnAC provide the opportunity to select language data based on its targeted qualities – e.g. language pragmatics, source and target language, source of the text, year of publishing, etc. This feature of BulEnAC contributed to the greater variety of language material that is included in the training data.

- Cons.

1. The main problem faced during the manual work on preparing the training data is the insufficient variety of verb forms in both Bulgarian and English. Although the initially selected language resources were able to provide various language materials, they were not able to ensure the grammatical variety of the linguistic data. Previous studies (Lazarov, 2017) have shown that the distribution of tense forms in Bulgarian, as source language, is uneven and reliable statistical data can be obtained through large and representative corpora. The distribution of tense forms according to Lazarov (2017) is provided in Table 2. This work represents statistical data obtained from small corpus (of around 200 sentences) focusing on the frequency of occurrences of Bulgarian tense forms:

Tense form	Frequency of occurrences
Aorist	40,5%
Imperfect	20%
Praesens	19,5%
Futurum	10%
Perfect	4,5%
Futurum praeteriti	2%
Plusquamperfect	0,5%
Other verbal forms	3%

Table 2: Frequency of occurrences of tense forms for Bulgarian

Since this fact would affect the constructed statistical model, we have made several improvements of the initial working data. The initial conception of implementing whole texts in the corpus was dismissed and single not logically connected sentences were introduced to the initial working data.

After the preparation of the training data is completed the deficient tense forms will be artificially constructed and introduced to the training set. The success rate of this method will be assessed during the manual correction of the annotated data based on the model constructed by the training data.

5. Future work and research aims

After the training data is manually annotated, evaluated and completed with artificially constructed tense forms it will be used to train an annotation model on the TreeTagger. The PoS tags are intended to be the primary input data. The output data will be the tagset of the second annotation layer assigned to chunks of tokens from the input layer. The targeted volume of the corpus is 5,000 aligned pairs of sentences with the two layers of annotation. The current workflow aims at creating a corpus with frequency of tense form occurrences close to the presented in Table 2. As can be seen from the data presented in Table 2, the three most frequent tense forms represent more than 75% of the total occurrences. This fact will result in artificially constructed and translated tense forms for the purposes of creating scientifically representative corpus.

The targeted volume of 5,000 entries was determinated after analyzing and preparing the suitable resources and after summarizing the available literature on the issue. On one side, although the major part of the preselected linguistic resources (BulEnAC) represent a perfect prerequisite to start the project at the stage of annotating the second layer, it is an automatically annotated resource and thus contains unresolved annotation issues, that have to be resolved beforehand. On the other side, the smaller part of the resources consists of linguistic data that can provide a reasonable diversity of temporal forms to amplify the data. The targeted volume was also determined after considering that most of the temporal forms practically have zero frequency in present-day Bulgarian (Kucarov, 2007). Aiming at collecting equal numbers of examples for all tenses would be labor-intensive and statistically inaccurate since the constructed corpus won`t consist of adequate representative data. The targeted volume of the corpus aims at presenting enough translation variations of the Bulgarian temporal forms in English at a satisfactory level for future scientific researches based on the corpus data and the methodology for its construction.

The aims of this project and consequently the creation of the described corpus are to create a statistical translation model for verb forms, which will be based on reliable linguistic data. The statistical model will be able to provide an answer to the initial questions of this research: in what manner the grammatical information is transferred between Bulgarian and English; what type of grammatical information is transferred and what type is lost during the process of translation and why; how close are the verbal morphological categories of both languages and in what manner are they related; in what manner the combination of certain grammatical categories in Bulgarian influences the translation in English. Most of these questions already have elaborate theoretical explanations which will be empirically demonstrated.

The constructed corpus, the gathered scientific data and the constructed statistical language and translation models are envisioned to be freely available linguistic resources for various scientific purposes. The training models and the working files for the TreeTagger and WebAnno will be published together as part of the ready-to-use linguistic resource.

Acknowledgements

This article is part of project No. 72-00-40-221/10.05.2017: „Българо-английски граматични паралели с оглед на машинния превод. Обогатяване на статистически модел за превод от български на английски с лингвистична информация” (Bulgarian-English grammatical parallels with respect to machine translation. Enriching statistical translation mode from Bulgarian to English with linguistic information) sponsored by “Програма за подпомагане на млади учени и докторанти на БАН – 2017” (Support program for young scientists and PhD students at the Bulgarian Academy of Sciences – 2017).

References

Bond, F., Oepen, S., Siegel, M., & Flickinger, D. (2005). Open source machine translation with DELPH-IN. *Open-Source Machine Translation Workshop at the 10th Machine translation Summit*, (pp. 15-22).

Iliev, G. (2014). *Ezikovo motivirana optimizatsiya na mashiniya prevod.* Department of Computational Lingusitcs, IBL-BAS. Retrieved from http://roboread.com/doc/Iliev_G_Dissertation.pdf

Ivanova, K. (1968). Varhu vzaimotnosheniyata na glagolnata prefiksaciya i kategoriyata prehodnost/neprehodnost v savremenniya balgarski knizhoven ezik. *Slavistitshen sbornik*, 156-157.

Kabakciev, K. (2000). *Aspect in English: A `Common-Sense' View of the Interplay Between Verbal and Nominal Referents.* Springer.

Koeva, S., Rizov, B., Tarpomanova, E., Dimitrova, T., Dekova, R., Stoyanova, I., . . . Genov, A. (2012). Bulgarian-English Sentence- and Clause-Aligned Corpus. *Proceedings of the Second Workshop on Annotation of Corpora for Research in the Humanities (ACRH-2).*

Kucarov, I. (2007). *Teoretitshna gramatika na balgarskiya ezik. Morfologiya.* Plovdiv: University press Paisii Hilendarski.

Lazarov, T. (2016). Analysis of the Resources for Statistical Translation Model of the Verb Forms from Bulgarian to English. *Bulgarian language*, 96-102.

Lazarov, T. (2017). Functional grammatical parallels with regards of translation of verb forms from Bulgarian to English. *Proceedings of the International Jubilee Conference of the Institute for Bulgarian Language (Sofia, 15 – 16 May 2017).*

Nedelcheva, S. (2012). Bulgarian Ingressive Verbs: The Case of Za- and Do-. *Godishnik of University of Shumen Konstantin Preslavsky*, pp. 72-89.

Riezler, S., & Maxwell, J. T. (2006). Grammatical machine translation. *Human Language Technology Conference - North American Chapter of the Association for Computational Linguistics Annual Meeting (HLT-NAACL'06).* New York; NY.

Santorini, B. (1991). *Part-of-Speech Tagging Guidelines for the Penn Treebank Project.* University of Pennsylvania.

Schmid, H. (1995). Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop.* Dublin.

Simov, K., Osenova, P., & Slavcheva, M. (2004). BulTreeBank Morphosyntactic Tagset. In *BulTreeBank Technical Report BTB-TR03.*

Sinapova, L., & Dochev, D. (1999). Analyzing Bulgarian and English Collocations. *Problems of Engineering Cybernetics and Robotics.*

Vassileva, A. (2003). Tense variations in Bulgarian narratives and their translational equivalents in English. Plovdiv University “Paisii Hilendarski”. Retrieved from http://georgesg.info/belb/doktoranti/vasileva/MA2.htm

Yimam, S., Castilho, E., & Gurevych, I. (2014). Automatic Annotation Suggestions and Custom Annotation Layers in WebAnno. *Proceedings of ACL-2014, demo session.* Baltimore, MD, USA.

Appendices

```
#Text=012bg
1-1 0-4     Фонд          Ncmsi          _   _
1-2 5-7     ще            Tx             Vfutur[1]    *->1-1
1-3 8-16    подкрепя      Vpitf-r3s      Vfutur[1]    *->1-2
1-4 17-26   Балкански     A-pi           _   _
1-5 27-36   кинодейци     Ncmpi          _   _
1-6 37-39   .             sent           _   _

#Text=012en
1-1 0-4     Fund               NN        _   _
1-2 5-6     will               MD        Vfs[1]   *->1-1
1-3 7-14    support            VV        Vfs[1]   *->1-2
1-4 15-21   Balkan             JJ        _   _
1-5 22-38   cinematographers   NNS       _   _
1-6 39-40   .                  sent      _   _
```

Appendix A. Example of the file format.