Parallel Web Display of Transcribed Spoken Bulgarian with Its Normalised Version and an Indexed List of Lemmas

Marina Dzhonova
Sofia University "Sv. Kliment Ohridsky"
Faculty of Slavic Studies
mdjonova@gmail.com

Kjetil Rå Hauge
Oslo University
ILOS
k.r.hauge@ilos.uio.no

Yovka Tisheva
Sofia University "Sv. Kliment Ohridsky"
Faculty of Slavic Studies
yovka.tisheva@abv.bg

Abstract

We present and discuss problems in creating a lemmatised index to transcriptions of Bulgarian speech, including the prerequisites for such an index, and why we consider an index preferable to a search engine for this particular kind of text.

1. Introduction

This article focuses on the possibilities for automatic tagging of corpus of oral communication in the modern Bulgarian language. What distinguishes the object of our article from the more well-known corpora of Bulgarian language is the nature of the texts included in it. This corpus is not composed of written texts, but includes data representative of oral communication in different communicative situations with the participation of speakers of Bulgarian of varying status. The texts in the corpus are transcriptions of audio or video recordings of oral communication. In this sense, the written texts in the corpus are secondary to the original speech acts. The uniqueness of corpora of this type is related both to the specifics of the linguistic factors involved (spoken language, literary pronunciation, etc.) and to the establishment of standards and conventions for recording and transcribing oral speech.

Oral speech is one of the forms through which the modern Bulgarian language is realized. It is also its most dynamic form, where new tendencies in the language are introduced and the validity of normative criteria are contested. For an all-encompassing description and study of the modern Bulgarian language it is necessary to know both its written form and the oral variant. This understanding is the basis of the BgSpeech initiative (Tisheva and Dzhonova 2011; Tisheva and Dzhonova 2014; Tisheva 2014; Hauge and Tisheva 2014; Hauge et al. 2016), which brings together Bulgarianists and Slavists with research interests in oral communication (see bgspeech.net for participants). The creation of a corpus that is representative for the contemporary state of Bulgarian oral speech is one of the long-term tasks of the team. Resources of this type represent the called-for parallel to corpora of written (standard) texts. The creation of a corpus of oral speech complements and enriches the knowledge about the modern Bulgarian language. The inclusion of data on oral communication broadens the representativeness of linguistic research.

Compliance with the literary norm is mandatory in all cases in which the written form for realization of the Bulgarian language is chosen. In oral communication the picture is different — norms of literary pronunciation, as well as grammatical and lexical norms become "more elastic", and their application to a great extent dispensable in different speech situations. The complex of linguistic means that are at disposal in oral speech follows the basic features of the national (official) language because it is part of it. But its phonetics and grammar do not fully follow all the specifics of the written literary language, nor do they comply with any existing dialect norm. Some of the peculiarities that are noted in the transcribed oral speech in the corpus are elisions, ellipses, abbreviated forms or phrases, overlapping utterances, incomplete utterances, repetition of constituents/phrases, colloquial constructions, pragmatic markers and discourse markers. The transcriptions also give information

Keywords: corpus, keywords, spoken language

about the paralinguistic means used by the speakers (pauses, gestures, mimics, phonetic paralinguistic means such as laughter, etc.), as they are an integral part of oral communication. Along with linguistic information, a mandatory condition for the corpus to be representative is to include non-linguistic information (socio-demographic features of the participants in the communication, data about the recording itself).

The special features of speech also call for a specific approach to designing this type of corpus. While the first level of annotation in corpora of written language are lemmatization and part-of-speech analysis, in corpora of spoken language part of the syntactic and pragmatic annotation is carried out as an integral part of the transcription of the recordings. This is necessary in order to determine the boundaries between the individual utterances in the organization of the transcription into a dialogical form. Simultaneous speaking, pauses, overlapping as well as non-verbal information and the communicative status of the utterances are noted by the transcriber in the initial processing of the texts. The same applies to the metadata that accompany every transcription — information about speakers and recordings is also provided by transcription. Part-of-speech and clausal annotation become the next stage for a corpus of oral speech.

Practice around the world shows that oral communication data can be collected into separate, self-contained corpora of varying volume and degree of representativeness or included in representative national corpora as a sub-corpus under the main database of written texts. The resource under consideration here is not part of a larger corpus.

The resource we present here is organised in the form of a small parallel corpus. It presents two parallel (tabular) text records of the same audio source, where one represents the result of an editing/ normalizing process into the standard norm and the other the original transcriptions with the deviations from the norm indicated. This processing aims to facilitate both the extraction of data on the grammar and pragmatics of oral speech as well as the further automatic processing of the resource. Most of the texts represent unofficial colloquial speech, and in addition there are two interviews and one media text.

A search engine interface is an ideal tool for a user who wants to find out whether a certain item is a part of a given rather large set of items, for instance whether “floccinaucinihilipilification” is a word in English (it is, and it means ‘the action or habit of estimating something as worthless’ according to oxforddictionaries.com). But when the set of items is on a small scale and/or contains items that are confined to certain geographical or social entities or are scarce or of recent emergence, a gaping empty search field is of little help for the user, especially if the user is new to the field. Such sets of items are the vocabularies of dialects, professional or social jargon, neologisms, allegro forms, and colloquialisms. A better tool for such sets will be an index, that is, a list of all occurring forms with an explanation and/or an indication of the form’s place in the text.

The texts available at bgspeech.net are sets of comparably short transcriptions of spoken Bulgarian. Transcriptions of spoken language tend to contain a large number of spellings that reflect the actual pronunciation and thus differ from the standard spelling. Researchers on the hunt for data about, for instance, the use of subjunctions of cause, would search for, among other things, *защото* ‘because’, but might not consider the option of searching for an allegro form such as *штот*. This means that it could be necessary, depending on the degree of non-standard spellings in the transcripts, to produce normalised versions of them.

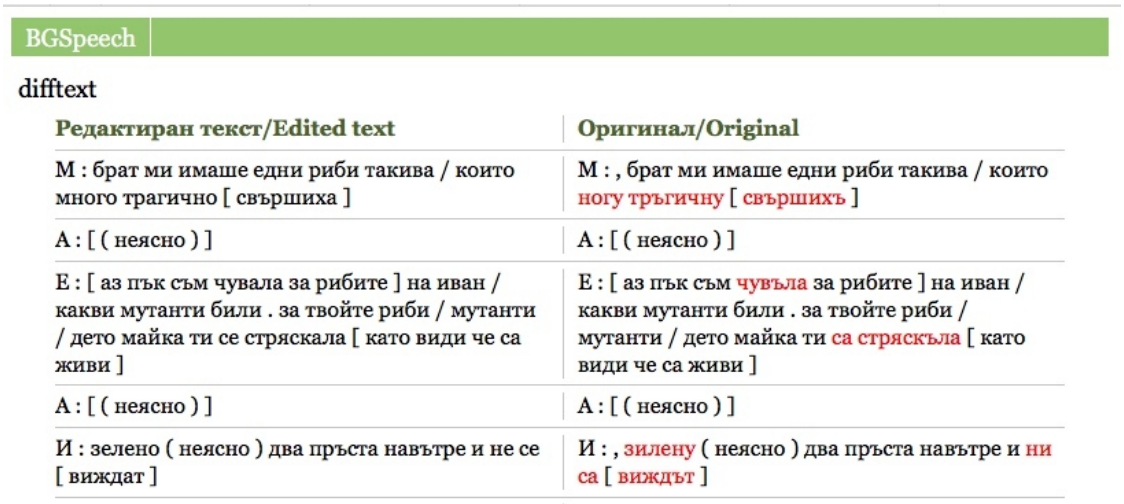
As described in a poster at CLIB2014, a part of the transcriptions are normalised, in the sense that in addition to the version with phonetic transcription there is a parallel version with the same texts normalised to the standard orthography. Normalisation has been effected through replacement of known pairs of semi-phonetic and standard spelling (1,358 in all), spellcheck with MacEst, developed by the Department of Computational Linguistics at the Institute for Bulgarian Language (hereafter DCL/IBL), and additional visual checking.

There are several advantages connected with transcribing spoken language in standard orthographic form. In a report on a study of a Russian dialect, the authors explain why they forego transcriptions like “*[оп свojóй жы́з’н’е это хоц’у́ погувур’и́т’ / жы́с’ мojá прошлá н’е о́ц’ен’ вá́жно / жы́ла ф-так’и́jо го́ды т’ежо́лыjо / ə’ит’е́й у м’ен’á бы́ло п’е́т’еро / подн’alá ja ə’ит’е́й до войны́ / фторо́й сын пог’и́п на войн’е́]*” in favour of “*Об своей жизни это хочу поговорить. Жизнь моя прошла не очень важно, жила в такие годы тяжелые. Детей у меня было пятеро, подняла я детей до войны, второй сын погиб на войне.*” Five reasons are given:

transcription into standard language can be done quickly; it effectively solves the problem of normalization and standardization (as phonetic transcription systems used in different dialect corpora do not always coincide even for the same language); it makes the use of standard automatic annotation tools possible; it makes the data easily readable by non-linguist users; and loss of phonetic data in transcription may be made up for by aligning the transcription with the original audio, so they conclude: "All this boils down to the principle that, to make standard taggers applicable to the texts, we make as much phonetic adaptation as possible, rea- sonable and practicable without losing lexically, morphologically and syntactically relevant information (Waldenfels et al., 2014).

Furthermore, as the volume of transcribed texts in our case is comparatively small, an index, providing a full list of lemmas and forms, could provide a better overview of the vocabulary of the text than what one could attain by typing search terms into a search engine. In our case we have to do with 23 transcripts, varying in volume from 477 to 2,425 tokens and with a total of a little over 5,000 unique tokens, and a number of lemmas considerably smaller than that.

These transcripts are presented in a two-column view, normalised transcript to the left and the original to the right, with highlighting (red type) of the deviations that have been corrected as shown in fig. 1.



BGSpeech

difftext

Редактиран текст/Edited text	Оригинал/Original
М : брат ми имаше едни риби такива / които много трагично [свършиха]	М : , брат ми имаше едни риби такива / които ногу тръгичну [свършихъ]
А : [(неясно)]	А : [(неясно)]
Е : [аз пък съм чувала за рибите] на иван / какви мутанти били . за твойте риби / мутанти / дето майка ти се стряскала [като види че са живи]	Е : [аз пък съм чувъла за рибите] на иван / какви мутанти били . за твойте риби / мутанти / дето майка ти са стряскъла [като види че са живи]
А : [(неясно)]	А : [(неясно)]
И : зелено (неясно) два пръста навътре и не се [виждат]	И : , зилену (неясно) два пръста навътре и ни са [виждът]

Fig. 1: Two-column display

For the new display we are adding for each transcript a column to the left with a clickable list of lemmas and their attested wordforms, where a click on a wordform will lead to its instance in the text. Furthermore, there will be a fourth column on the right with an alphabetised list of all the corrected forms, that is, all the highlighted forms in the column with the original transcript. Each form in the list is clickable and will lead the the form in its context in the original text – see figs. 2 and 3. In addition, there will be a separate document with a full alphabetical list of all highlighted forms with links to the documents in which they occur.

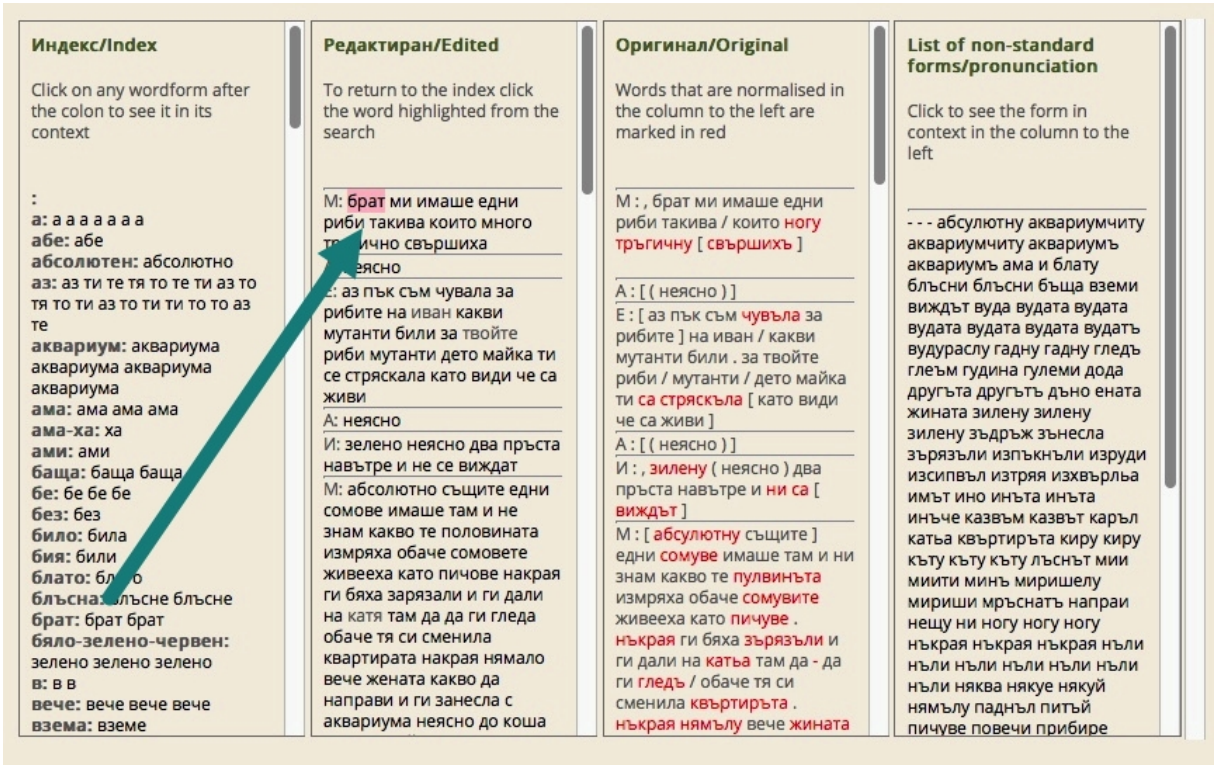


Fig. 2: Index of occurring lemmas with clickable word forms and links to the normalised text

Fig. 3: List of non-standard forms with clickable links to original transcription

2. Method

Our basic tool for lemmatisation is the Bulgarian morphological dictionary used in the production of the declination/conjugation patterns in Popov et al., 1998 and Popov et al., 2003, provided for us by Kiril Simov. The textual format of the dictionary was massaged into a more compact form using Applescript, and searches were made with database speed in the text editor BBEdit. An alternative would have been to use the lemmatiser provided by the Department of Computational Linguistics at the Bulgarian Academy of Sciences (http://dcl.bas.bg/dclservices/index.php), but lack of time for establishing a script for analysing the web pages sent in return from the lemmatiser made us go for a simpler solution. A custom-made script then traverses the normalised part of the HTML file, looking every wordform's lemma up in the morphological dictionary, producing a new document with each lemma connected with every one of its wordforms. A second script then produces HTML code for the new column from that document. The HTML and CSS coding patterns are borrowed from code by David J. Birnbaum and David Galloway at *The annotated Afanas'ev library* (http://aal.obdurodon.org/about.php).

3. Issues

3.1. Multiword Units

A considerable problem is posed by multiword units of the type *еди-кой/еди-чий си*, *кой/чий да е/било*, *който/чийто и да е/било*, all meaning 'whoever/whoseever'. Without special markup in the text to be lemmatised, each part of the unit will be lemmatised according to its single-form homograph. This problem has been addressed in a doctoral dissertation at IBL/BAS (Stoyanova, 2012), but there are still remaining problems — in IBL/BAS' lemmatiser, *било* in *когото и да било* is not recognised as a part of a multiword unit, and neither as a form of *съм* 'to be', but as a form of *бия* 'to beat':

```
<text>
<item><P> X <P> X</item>
<item><S> X <S> X</item>
<item>няма Vs няма VBIAr3s</item>
<item>да C да C</item>
<item>говоря Vs говоря VLITr1s</item>
<item>с R с R</item>
<item>когото Ps който PROasm</item>
<item>и C и C</item>
<item>да T да T</item>
<item>било Vs бия VLITxsno</item>
<item></S> X </S> X</item>
<item/>
</text>
```

In our case, we are slightly better off than the IBL/BAS lemmatiser, because we know exactly which texts we are going to lemmatise and can groom them to our requirements in advance, and not only that, we can also adjust the morphological dictionary, where each of these multi-word units is represented as one lemma. So we do the following, expressed in pseudocode:

```
for each lemma in morphological dictionary
    if the lemma is multiword
        for each wordform in the lemma's set of wordforms
            search for the wordform in texts to be lemmatised
            replace " " with "_" in text
            replace " " with "_" in the lemma and wordforms in morphological dictionary y
        end
    end
end
```

We now end up with a situation where all multiword items have been converted to singleword items, both in the morphological dictionary and in the text to be lemmatised, and all that remains is to go on with the job of lemmatising and by all means remember to convert all underscores back into spaces before we publish it.

3.2. Lemmatisation Errors

There is a definite need for disambiguation of homographs in the lemmatisation process — is *говори* a form of the verb *говоря* 'to speak' or of the noun *говор* 'speech; dialect'? An educated guess for the right answer can be made by checking the immediate context of the word: if the preceding word form is an adjective in the plural (as in *западните говори* 'the western dialects'), there is a considerable chance that the form belongs to the lemma *говор*, while if it is followed by a preposition (*говори за* 'speaks of', *говори с* 'speaks with), there is a similar chance that it is a form of the verb *говоря* (Simov et al., 2013).

Our lemmatisation was "quick-and-dirty" — we let the script accept the first hit for any given word form, expecting to do a clean-up job afterwards for cases like *прави* in *май беше където ти прави прическата* being classed as a of the adjective *прав* 'right, correct' instead of the verb *правя* 'to do'; or *иска* in *каза нали че иска да е при мене* as an articled form of the masculine noun *иск* 'claim, action' rather than as a form of the verb *искам* 'to want'.

IBL/BAS' lemmatiser, mentioned above, will do a better job with these, relating both *прави* and *иска* to their proper lemmas (although mislabelling the particle *май* as a noun, but that was a tricky one, with no left context):

```
<text>
<item><P>  X   <P> X</item>
<item><S>  X   <S> X</item>
<item>май Ns  май NCMNsom</item>
<item>беше Vs  съм VLINd3s</item>
<item>където  D   където D</item>
<item>ти   Ps  аз  PHi2s</item>
<item>прави   Vs  правя  VLITe3s</item>
<item>прическата  Ns  прическа   NCFsdf</item>
<item></S> X   </S>   X</item>
<item/>
</text>

<text>
<item><P>  X   <P> X</item>
<item><S>  X   <S> X</item>
<item>каза Vs  кажа   VLPTe2s</item>
<item>нали T   нали   T</item>
<item>че   C   че  C</item>
<item>иска Vs  искам  VLITe3s</item>
<item>да   C   да  C</item>
<item>е Vs  съм VLINr3s</item>
<item>при  R   при R</item>
<item>мене Ps  аз  PHyt1s</item>
<item></S> X   </S>   X</item>
<item/>
</text>
```

However, the DCL lemmatiser did not excel in all cases. While our method (or lack of it) assigned the plural noun form *движения* 'movements' to the verb *движа* 'to move', the DCL lemmatiser, even with (or perhaps mislead by) two plural adjectival forms in the left context, proposed the adjective *движен* 'moved':

```
<text>
<item><P>  X   <P> X</item>
<item><S>  X   <S> X</item>
<item>виж Vs  видя   VLPTI2s</item>
<item>че   C   че  C</item>
<item>такива  Pp  такъв  PDAp</item>
<item>елементарни    Ap  елементарен   Apo</item>
<item>движения   Np  движенAqsmo</item>
<item></S> X   </S>   X</item>
<item/>
</text>
```

Both approaches failed miserably with the 1st person verb form *отчета* ‘to account for’ in *нали бях си насъбрала пари да и ги отчета*. Bypassing, or in our case, not even reaching to the same-stemmed noun *отчет* ‘account’ , they suggested *отче* ‘father (in the religious sense)’:

```
<text>
<item><P>  X   <P> X</item>
<item><S>  X   <S> X</item>
<item>нали T   нали   T</item>
<item>бях  Vs  съмVLINe1s</item>
<item>си   P   себе   PFHzt</item>
<item>насъбрала   Vs  насъбера   VLPTxsfo</item>
<item>пари Np  пара   NCFpof</item>
<item>да   C   да  C</item>
<item>иC   и   C</item>
<item>ги   Pp  аз  PHza3p</item>
<item>отчета   Np  отче   NCNpon</item>
<item></S> X   </S>   X</item>
<item/>
</text>
```

4. Conclusion

The use of the method described in this test case shows what problems may arise when trying to use presently available programs for the automatic processing of Bulgarian text data. Normalisation of the word forms in the text is a necessity, as the available programs and morphological dictionaries include only data from the written language, and the remaining conversational syntactic structure may restrict the automatic annotation of the text. It is also obvious that a degree of manual assistance will be necessary in any case. The lessons learned so far will be applied to the tagging of the other speech data we have at our disposal and will hopefully facilitate user access to our data.

References

Hauge et al. 2016: Hauge, K., Tisheva, Y., Džonova, M. (2016). BgSpeech i predstavyaneto na ustnata rech v Balgarskiya natsionalen korpus. *Problemi na ustnata komunikaciya*. T. 10. Chast 2. Veliko Tarnovo: UI „Sv. sv. Kiril i Metodij”, 175–186.

Hauge, Tisheva, 2014: Hauge, K. Ro, Tisheva, Y. (2014). Paralelen korpus s danni za balgarskata razgovorna rech – struktura i prilozhenie. *Ezikovi resursi i technologii za balgarski ezik*. Sofija: Akademichno izdatelstvo „Prof. Marin Drinov”, 142–153.

Popov et al., 1998: Popov, D., Simov, K., Vidinska, S. (1998). *Rečnik za pravogovor, pravopis, punktuacija*. Sofija: Atlantis.

Popov et al., 2003: Popov, D., Simov, K., Vidinska, S. (2003). *Pravopisen rečnik na bălgarskija ezik*. Sofija: Nauka i izkustvo.

Simov et al., 2013: Simov, K., Ivanova, G., Mateva, M., Osenova, P. (2013). Integration of dependency parsers for Bulgarian. *The Twelfth Workshop on Treebanks and Linguistic Theories*, 145–156. Sofia.

Stoyanova, 2012: Stoyanova, I. (2012). .*Avtomatichno razpoznavane i tagirane na sastavni leksikalni edinitsi v balgarskiya ezik*. Disertaciya za prisazhdane na obrazovatelnata i nauchnata stepen „doktor". Sektsiya po kompyutarna lingvistika, Institut za balgarski ezik, Balgarska akademiya na naukite.

Tisheva, Dzhonova, 2011: Tisheva, Y., Dzhonova, M. (2011). Korpus s ustna balgarska rech – struktura i spetsifika. *Balgarski ezik* 3, 34–53.

Tisheva, Dzhonova, 2014: Tisheva, Y., Dzhonova, M. (2014). Balgaristikata – mezhdu fishovete i multimediynite korpusi. *Balgarski ezik, literatura i e-obuchenie*. Plovdiv: „Rakursi" OOD, 24–35.

Tisheva, 2014: Tisheva, Y. (2014). Ezikovi bazi danni, korpusi i elektronni resursi za balgarskata ustna rech. *Littera et Lingua* 11, 1–2. http://slav.uni-sofia.bg/naum/lilijournal/2014/11/1-2/ytisheva

Waldenfels et al., 2014: Waldenfels. R. von, Daniel, M., Dobrushina, N. (2014). Why Standard Orthography? Building the Ustya River Basin Corpus, an Online Corpus of a Russian Dialect. http://www.dialog-21.ru/digests/dialog2014/materials/pdf/WaldenfelsR.pdf.