Petya Osenova and Kiril Simov

ALL ABOUT WORDS! AN INTEGRATED DICTIONARIES PORTAL FOR BULGARIAN

Abstract The paper introduces a web portal of integrated dictionaries for Bulgarian. The mapping among the resources is lemma-based. Two dictionaries are in the centre of this integration – an Inflectional dictionary of Bulgarian, since Bulgarian is a morphologically rich language, and a Wordnet of Bulgarian – BTB-Wordnet, since it adds the level of lexical meanings to the dictionary-enhanced knowledge. Also, various other types of dictionaries are being gradually added – with diachronic spellings, bilingual, specialised, etc.

Keywords Bulgarian; dictionaries; integration

1. Introduction

The tradition in dictionary making and consequently – dictionary access, favoured the following rule: for each language phenomenon – a separate dictionary, mainly on paper. Thus, there were dictionaries for spelling, synonyms, antonyms, derivational relations, explanatory dictionaries with senses, new words and senses, jargon, dialectal, of foreign words, etc. Then the time of thesauri and linked dictionaries came – wordnets and ontowordnets were created for many languages; also babelnets, verbnets, framenets, wikidata, wiktionaries, etc. However, very often the users cannot get oriented where to find the most appropriate dictionaries for their specific tasks, especially when they need to consult not only contemporary dictionaries, but also historical ones, or when they need more complex information accessible in more than one dictionaries. Within CLaDA-BG infrastructure[1], we have to provide access to dictionaries that are in active use by different researchers in their everyday work. For example, ethnographers and historians very often explore documents written in different time periods and in different language varieties of Bulgarian. Thus, they consult dictionaries published more than one hundred years ago. In order to provide access to such dictionaries one of the tasks is to scan, OCR and proofread old dictionaries of Bulgarian, published within the period from the second half of the nineteenth century to the middle of the twentieth century. In order to facilitate the access to these dictionaries, our solution was to construct their online representation and to connect them to the contemporary lexical resources.

Thus, we decided to provide a web-based portal for users in Bulgarian which integrates a number of dictionaries with interlinked content.[2] The portal is based on a vocabulary of lemmas shared among dictionaries. For each lemma its paradigm (if available) was stored and linked to the corresponding lexical entries

[1] https://clada-bg.eu/en/

[2] https://concordance.webclark.org/#/word/гледах

in the other dictionaries. The user can navigate within the vocabulary through: i) access to a list of lemmas starting with a specific prefix; ii) a specific lemma; iii) a word form which first undergoes lemmatization and then is searched in the vocabulary. Here a central role plays the *Inflectional dictionary of Bulgarian* (IDB).[3] It provides information not only about the lemma (with its part-of-speech), but also about its paradigm and grammatical features. The next important dictionary-thesaurus is *BTB-Wordnet*.[4] We put this resource in the center of our dictionary hub because it covers a carefully selected set of senses and thus provides a basis for clustering information on sense level from the other dictionaries within the system. In the current version of the portal the alignment has been performed on the basis of lemmas, but in future we consider also an alignment on sense level. Although the lemma level seems to be an easy mechanism for alignments, there are two main problems to be solved – at least in our case. The first problem is that some of the dictionaries are results from OCR-ed paper dictionaries or based on paper dictionaries, and others are already digitally born. Thus, the approaches to lemmatization very often differ – especially with respect to inclusion or exclusion of optional reflexive particles or to considering attributively used participles as a separate part-of-speech. The second problem is the dynamics in the development of the Bulgarian language spelling system in the period from the second half of the nineteenth century to the middle of the twentieth century. This fact caused some difficulties in word normalisation and the consequent lemmatization. Thus, intensive work was invested from our side to unify the lemma representation across all the available resources.

The structure of the paper is as follows: the next section outlines some conceptual foci of the related work. Section 3 gives an overview of the integrated dictionaries model. Section 4 outlines the implementation details of the web portal. Section 5 concludes the paper.

2. Related Work

The idea of integrating and interlinking various grammatical and lexical resources has been explored successfully for quite some time now for various languages and various tasks. Here we mention only a few of the related works to show what kind of approach we follow at the moment, and what our aims are in the near future.

Herold et al. (2012) introduce a complex typed link structure to ensure the proper relations among large modern and historical dictionaries with differing lexicographic traditions. The mapping has been performed through an aligned lemma list where the strategies for achieving entry equivalence are presented. Our approach is similar and we ensure the mappings through a lemma list, unified across resources.

[3] The link is with an example: the paradigm of the word form *гледах* (gledax-1PERS.SG.AORIST/IMPERFECT, I looked/was looking): https://concordance.webclark.org/#/word/%D0%B3%D0%BB%D0%B5%D0%B4%D0%B0%D1%85

[4] The link is with an example – the word *глад* (glad, 'hunger'): https://concordance.webclark.org/#/wordnet/глад?synID=14281&pos=n

In Tiberius et al. (2021, p. 60), the ELEXIS dictionary matrix has been discussed among other initiatives. It is defined as “universal repository of linked senses, meaning descriptions, collocations, phraseology, translation equivalents, examples of usage and other types of lexical information found in existing lexicographic resources, monolingual, multilingual, modern, historical etc.” In addition, the constituting elements of a common vocabulary have been presented. Our model is not sense-based yet but it comprises most of the characteristics of the common vocabulary such as an entry, a headword, a part-of-speech, a sense, a sense structure, a definition, an example, information about the inflected forms. More details of our model are given in the next section.

In Ahmadi et al. (2020, p. 3234), a sense-alignment strategy is presented for 15 languages, among which Bulgarian. The lemmas were selected through specially designed criteria. The aim was “to provide semantic relationships between two sets of senses for the same lemmas in two monolingual dictionaries.” For Bulgarian these were the BTB-Wordnet resource and the Wiktionary. In our integrated portal the mapping has not been performed through semantic relations yet but at the same time it refers to semantic resources and thus is semantics-aware.

3. Overview of the Integrated Dictionaries Model

The main components of the system are as follows: *Words in context* (this is the related concordancing service where one can search for word examples) and *All about words* (where information is provided about the grammatical features of words; about the synsets, they are part of; about their definitions, links to various dictionaries, etc.). If the user starts with the *All about words* component, they can write the word of interest in the search box. Even when it is not a lemma but a word form, the result will contain information about its lemma(s) as well. The integrated structure includes the following information: a) a part-of-speech; b) a grammatical paradigm, if any; c) a grammatical description; d) all the senses available in the wordnet; e) examples from the included corpora; f) relation to the existing etymological dictionary; g) relation to the existing terminological dictionaries, if applicable; h) relation to a ‘diachronized’ version of the word, if different from the contemporary spelling. Let us illustrate our integrated dictionary access with the word дете (dete, child, ‘a child’). From the large inflectional dictionary, the following information displays: де|те, ~тето, ~ца, ~цата. The vertical marker points the place in the lemma where it changes when it gets its word forms: a definite singular form (child-the ‘the child’), a non-definite plural form (children) and a definite plural form (children-the, ‘the children’). Then a piece of grammatical description is available from this dictionary about the form that has been selected by the user. In this case the description says: съществително нарицателно, среден род, единствено число, нечленувано ‘noun, common, neuter, singular, non-definite’. The BTB-Wordnet provides 4 meanings and 9 examples. At the moment the examples are listed separately from the meanings in the user interface, but it is easy to show the links between the examples and the meanings.

The lexical entries of the integrated lexicons have the following structure: (1) *Inflectional dictionary of Bulgarian*. The lexical entry in this lexicon consists of the following elements: < Lemma, POS, Paradigm, Inflectional class, Clitics >, where the *Lemma* coincides with the basic form of the word; *POS* is the part-of-speech of the lemma; *Paradigm* is a list of all word forms for the lemma where each word form is associated with a tag, representing its grammatical features; *Inflectional class* is an index to all lemmas that share the same inflectional rules; *Clitics* provides a list of obligatory or optional reflexive clitics for the lemma (usually for the verb lemmas). Currently, IDB contains around 81 238 lemmas with 1 485 026 word forms. We use an XML structure for the representation of the lexical entries. (2) *BTB-Wordnet* (Simov & Osenova, 2023) is a Wordnet for Bulgarian encoded in WN-LMF[5]. BTB-Wordnet contains 35 833 synsets and more than 50 000 lemmas. BTB-Wordnet is aligned to the Open English Wordnet.[6] Also, the lemmas in BTB-Wordnet are aligned to the lemmas in IDB. Each lemma within a given synset is related to examples that exemplify the corresponding sense. Currently, we have gathered nearly 80 000 examples. Our goal is to assign at least 3 examples per lemma within the synsets. (3) *The Bulgarian part of the PONS bilingual dictionaries.* This part was created in-house and the publisher left the copyright of the Bulgarian vocabulary to us. In this case the lexical entries have the following structure: < HeadWord, POS, Grammatical information, Inflectional Representation, Senses, Phraseology >. The lexical entry format is inspired by the Concede lexical model.[7] This dictionary contains about 45 000 lexical entries and covers more than 60 000 senses. For IDB and BTB-Wordnet (the core lexicons of the integrated lexicons portal) we map the head words with the lemmas in IDB and BTB-Wordnet. Here this mapping was done already during the creation of the Bulgarian vocabulary.

Thus, currently we use the lemmas from each added dictionary as a means for the integration. It should be noted that some lemmas are represented in just one of the dictionaries. In order not to lose data we maintain a joint list of lemmas covering all the dictionaries in the portal. When the user searches for a certain lemma, only the information from the corresponding dictionaries is displayed in the user interface. The lemmas in each dictionary are used for keys which are stored in the joint database of all dictionaries.

Similarly, we define an appropriate XML representation for the elements of the lexical entries in the corresponding dictionary. In this way, we have a database representation of the dictionary which allows us to check and edit the dictionary content. Thus, for each dictionary, we first construct an appropriate representation in XML, through which the appropriate structure of the lexical entries is reflected. After having represented the lexical entries of some dictionaries, we aim at unifying their lemmas. In order to maintain digitised dictionaries published before 1945 and within the old Bulgarian orthography, we also created a list of lemmas according to this orthography with corresponding lemmas in the contemporary vocabulary.

[5] https://github.com/globalwordnet/schemas

[6] https://en-word.net/

[7] https://www.researchgate.net/publication/2403685_The_Concede_model_for_lexical_ databases

Usually, the mapping is done by looking up in the list of old lemmas as a first step. If the candidate lemmas are not presented within the list, we add them to it, convert them to the new spelling (via a set of rules) and a specialist checks the result.

In the extension phase of the web portal, we envisage that each dictionary has its own webpage and thus might be consulted separately from the rest of the portal. At the moment, such a separate webpage is implemented for the BTB-Wordnet only. The switch to the webpage is available from the list of meanings presented within the *All about words* component.

Two other contemporary dictionaries for Bulgarian to be included in the near future are: the Valency dictionary (Osenova, Simov, Laskova & Kancheva, 2012) and the Hurtlex dictionary. At the moment their verb vocabulary is being mapped to the wordnet lemmas and meanings, and the resources are being further edited and cleaned. The Valency dictionary represents the combinatorial potential of the Bulgarian verbs. It is also related to the BTB-Wordnet and to the English Verbnet. The Hurtlex dictionary started from the multilingual one presented in Bassignana, Basile & Patti (2018). The Bulgarian part was downloaded and cleaned. Then each lemma was connected to the appropriate sense in BTB-Wordnet. When missing there, the specialists added these senses themselves.

Besides these dictionaries, currently we have scanned, OCR-ed and converted to XML presentation some historical dictionaries. These dictionaries are also aligned to the joint list of lemmas. The dictionaries are: (1) *Najden Gerov. Dictionary of the Bulgarian Language with interpretation of speeches in Bulgarian and Russian, volumes 1 to 5 (published at the end of 19th century and the beginning of 20th century).* It is an explanatory dictionary, based on materials collected by the author himself for a period of 50 years. Besides the definitions, the dictionary contains information about the usages of the words as well as related proverbs, songs and others. (2) *Stoyan Romanski. Spelling dictionary of the Bulgarian literary language. 1933.* (3) *Stefan Mladenov. Etymological and orthographic dictionary of the Bulgarian literary language. 1941.* All these dictionaries are written in the old Bulgarian spelling system. In order to support the mapping between the lemmas in these dictionaries and the contemporary ones we maintain a representation of the joint lemma list in both spellings of Bulgarian. The joint list of lemmas is based on the *Inflectional lexicon of Bulgarian*. Thus, it contains the whole paradigms of most of the lemmas. The representation of the list in the old spelling system is presented (as mentioned above) in a separate dictionary of Bulgarian.

The provided combined functionality is somewhat similar to the word sketch oriented approach as defined by (Kilgarriff & Tugwell, 2001). However, the resource described here is not meant for lexicographers only, but also for any interested parties – specialists in Digital Humanities, students, interested parties, etc. We view the word sketch through its contexts of usage as well as through its morphosyntactic profile, its meanings, its diachronic stages, its terminological load, its register, domain, etc.

4. Web Portal Implementation Outline

There exists a CLaDA-BG dictionary system (Angelov, Simov, Osenova & Kancheva, 2022) which is used for dictionary creation and editing. It also includes some of the dictionaries present in the web portal of integrated dictionaries (such as BTB-Wordnet, Explanatory Dictionary of Bulgarian, Inflectional dictionary, and some others). The resources in the web portal can be used for easy searches and consultations. The portal displays resources we have the publishing rights for. One drawback is that at the moment the two services – the editing and the web ones – are separately implemented. Thus, as future work we plan to unify them. The first step is the design and implementation of a shared database. Currently, for the server we are using MariaDB[8], Spring Boot framework[9] and Java 17.

As it was described above, for each entry in a given dictionary one or more keys are constructed. The keys are matched to the joint list of lemmas of the inflectional dictionary and the wordnet. This list is extended with new lemmas when they are not present in the two central dictionaries. In this way all the dictionaries are arranged around this joint list of lemmas. If the key has no lemma to correspond to, then it is inspected by a specialist and added to the joint list of lemmas. It is also marked for future analysis. Such lemmas could be of two main sorts: (1) a contemporary lemma for words that are not presented within both - the inflectional dictionary and the wordnet; and (2) it is a lemma of an old word, which is not actively used in contemporary Bulgarian due to some old spelling, or of a word which is not used at all and thus not recognised by most Bulgarians at the moment. Such lemmas are marked as archaic.

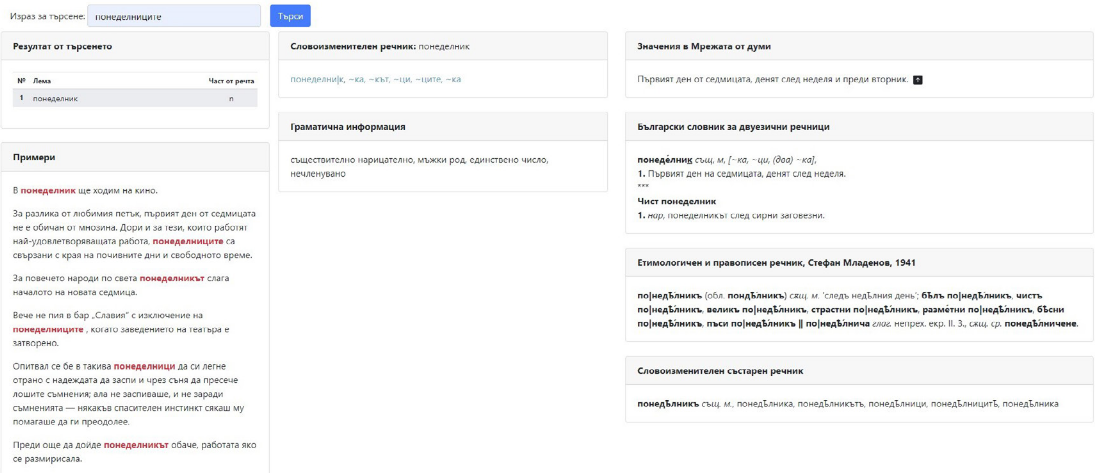
After the assignment of keys for each lexical entry in a given dictionary, the data is loaded into a relational database which includes an index on the lemmas. For each lexical entry the following information is stored in the databases: (1) a visual representation of the lexical entry; (2) examples related to the lexical entry; (3) the dictionary from which the lexical entry originates. For the user interface we rely on Bootstrap for UI[10] and plain Javascript.

[8] https://mariadb.org/

[9] https://spring.io/projects/spring-boot/

[10] https://getbootstrap.com/

Figure 1 presents the user interface to the *All about words* module.



понеделниците

Резултат от търсенето

№	Лема	Част от речта
1	понеделник	n

Примери

В **понеделник** ще ходим на кино.

За разлика от любимия петък, първият ден от седмицата не е обичан от мнозина. Дори и за тези, които работят най-удовлетворяващата работа, **понеделниците** са свързани с края на почивните дни и свободното време.

За повечето народи по света **понеделникът** слага началото на новата седмица.

Вече не пия в бар „Славия“ с изключение на **понеделниците** , когато заведението на театъра е затворено.

Опитвал се бе в такива **понеделници** да си легне отрано с надеждата да заспи и чрез съня да пресече лошите съмнения; ала не заспиваше, и не заради съмненията — някакъв спасителен инстинкт сякаш му помагаше да ги преодолее.

Преди още да дойде **понеделникът** обаче, работата яко се размирисала.

Словоизменителен речник: понеделник

понеделни|к, ~ка, ~кът, ~ци, ~ците, ~ка

Граматична информация

съществително нарицателно, мъжки род, единствено число, нечленувано

Значения в Мрежата от думи

Първият ден от седмицата, денят след неделя и преди вторник.

Български словник за двуезични речници

понеде́лник *същ, м, [~ка, ~ци, (два) ~ка],*
1. Първият ден на седмицата, денят след неделя.
***
Чист понеделник
1. *нар,* понеделникът след сирни заговезни.

Етимологичен и правописен речник, Стефан Младенов, 1941

по|недѢлникъ (обл. **пондѢлникъ**) *сжщ. м.* 'следъ недѢлния день'; **бѢлъ по|недѢ́лникъ**, **чистъ по|недѢ́лникъ**, **великъ по|недѢ́лникъ**, **страстни по|недѢ́лникъ**, **разме́тни по|недѢ́лникъ**, **бѢсни по|недѢ́лникъ**, **пъси по|недѢ́лникъ** || **по|недѢ́лнича** *глаг.* непрех. екр. II. 3., *сжщ. ср.* **понедѢ́лничене**.

Словоизменителен състарен речник

понедѢлникъ *същ. м.*, понедѢлника, понедѢлникътъ, понедѢлници, понедѢлницитѢ, понедѢлника

Fig. 1: The user interface to the *All about words* module

The input for the module is a search string. It can be any Cyrillic string. The search is processed in the following way: (1) first the string is analysed by the API for a word forms analysis. The API returns one or more lemmas for the word form. In this case, the lemmas are presented in the list of lemmas on the window in an alphabetical order; (2) if the string is not a (known) word form and cannot be analysed by the API, the system checks whether it is a prefix of a lemma. If the string is a prefix of such a lemma (or more than one), then the lemmas that start with the string are listed in the lemma box on the user interface. (3) If the string is not a word form or a prefix, then a message is displayed that there is no result. When the list of lemmas is displayed, the information related to the first lemma is presented. In the middle of the screen, the paradigm of the lemma is given (in case the lemma is included in the inflectional dictionary). Here the paradigm is given for the noun понеделник (ponedelnik, ‘Monday’). Below the paradigm information, the user might consult the grammatical features associated with each of the word forms. On the right side of the screen, the lexical entries from the different dictionaries are presented. First, the synsets from the wordnet are shown (in Figure 1 just one synset is presented, followed by the icon which allows for switching to the website of the wordnet). There more information about the synset is outlined. Then the lexical entries from the other dictionaries are shown. The last one in the list is from the etymological dictionary and it demonstrates the lexical entry in the old Bulgarian spelling (from before 1945). On the left of the screen, under the list of lemmas, a list of the available examples is given. The examples in this case are the ones related to the synsets in the wordnet as it was mentioned above.

5. Conclusions

In this paper, we presented a lemma-based model for integrating various dictionaries for Bulgarian in an online portal. We believe that this portal answers the needs of the users about getting knowledge from multiple dictionaries and within the context of other available lexical resources.

The lemma-based approach that we follow here is far from straightforward, especially in cases where diachronic data have been added. While the lemma-oriented mapping is very important for dictionary integration, it is not sufficient as a step. For that reason, as future work we plan to exploit some mixed models of dictionary mapping, i.e., lemma- and sense-based, example-based, user-oriented, etc.

References

Ahmadi, S., McCrae, J. Ph., Nimb, S., Khan, F., Monachini, M., Pedersen, B., Declerck, Th., Wissik, T., Bellandi, A., Pisani, I., Troelsgård, Th., Olsen, S., Krek, S., Lipp, V., Váradi, T., Simon, L., Gyorffy, A., Tiberius, C., Schoonheim, T., Ben Moshe, Y., Rudich, M., Abu Ahmad, R., Lonke, D., Kovalenko, K., Langemets, M., Kallas, J., Dereza, O., Fransen, Th., Cillessen, D., Lindemann, D., Alonso, M., Salgado, A., Sancho, J. L., Ureña-Ruiz, R.-J. Zamorano, J. P., Simov, K., Osenova, P., Kancheva, Z., Radev, I., Stanković, R., Perdih, A., & Gabrovsek, D. (2020). A Multilingual Evaluation Dataset for Monolingual Word Sense Alignment. In N. Calzolari, F. Béchet, Ph. Blache, Kh. Choukri, Ch. Cieri, Th. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 3232–3242). European Language Resources Association.

Angelov, Zh., Simov, K., Osenova, P., & Kancheva, Z. (2022). The CLaDA-BG Dictionary Creation System: Specifics and Perspectives. In T. Erjavec, & M. Eskevich (Eds.), *CLARIN Annual Conference Proceedings* (pp. 24–28).

Bassignana, E., Basile, V., & Patti, V. (2018). Hurtlex: A Multilingual Lexicon of Words to Hurt. In El. Cabrio, Al. Mazzei & F. Tamburini (Eds.), *Proceedings of the Fifth Italian Conference on Computational Linguistics* (CLiC-it 2018) (pp. 51–56). Accademia University Press.

Herold, A., Lemnitzer, L., & Geyken, A. (2012). Integrating Lexical Resources Through an Aligned Lemma List. In C. Chiarcos, S. Nordhoff, & S. Hellmann (Eds.), *Linked Data in Linguistics* (pp. 35–44). Springer.

Kilgarriff, A., & Tugwell, D. (2001). WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. In *Proceedings of the ACL Workshop on COLLOCATION: Computational Extraction, Analysis and Exploitation* (pp. 32–38). Association for Computational Linguistics.

Osenova, P., Simov, K., Laskova, L., & Kancheva, S. (2012). A Treebank-driven Creation of an OntoValence Verb lexicon for Bulgarian. In N. Calzolari, Kh. Choukri, Th. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Eighth International Conference on Language Resources and Evaluation* (pp. 2636–2640). European Language Resources Association (ELRA).

Simov, K., & Osenova. P. (2023). Recent Developments in BTB-WordNet. In G. Rigau, F. Bond, & A. Rademaker (Eds.), *Proceedings of the 12th Global Wordnet Conference* (pp. 220–227). Global Wordnet Association.

Tiberius, C., Krek, S., Depuydt, K., Gantar, P., Kallas, J., Kosem, I., & Rundell, M. (2021). Towards the ELEXIS data model: defining a common vocabulary for lexicographic resources. In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, & C. Tiberius (Eds.), *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference* (pp. 56–77). Lexical Computing CZ.

Acknowledgements

The reported work has been partially supported by CLaDA-BG, 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, funded by the Ministry of Education and Science of Bulgaria (support for the Bulgarian National Roadmap for Research Infrastructure). We would like to thank also the reviewers for their valuable and insightful comments.

Contact information

Petya Osenova
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences
petya@bultreebank.org

Kiril Simov
Institute of Information and Communication Technologies
Bulgarian Academy of Sciences, Bulgaria
kivs@bultreebank.org