Towards an Integrated Dictionary Portal:

the Dictionary of Bulgarian Language, the Bulgarian WordNet, the Grammatical Dictionary of Bulgarian and the Dictionary of Old Bulgarian

Svetla Koeva, Ivelina Stoyanova

Department of Computational Linguistics, IBL – BAS
E-mail: svetla@dcl.bas.bg, iva@dcl.bas.bg

1. Introduction

In recent years a large number of Bulgarian dictionaries and other language resources and tools have been made available online aimed at researchers (such as lexicographers), teachers and students in Bulgarian language and literature, as well as the wide public. Efforts are now focused on widening the usage of those electronic resources and providing an integrated environment which allows the users to access several resources simultaneously in an easy and efficient manner.

Dictionary portals gain in popularity due to the need to provide easy and uniform access to a range of language resources. Engelberg and Mueller-Spitzer (2009) outline the main criteria for distinguishing the types of dictionary portals: (1) types of access provided, (2) the implemented cross-reference structures between dictionaries, (3) the ownership relation between the portal and the contributing dictionaries, and (4) the layout of the portal. These are closely related to the number and variety of available resources to be integrated, and the needs of the target users.

Most popular dictionary portals provide multi-layered information such as definitions, spelling, pronunciation, synonyms (e.g., [1]), and interface to multilingual resources (e.g. Google Dictionary integrated into the search engine via the define operator). Internet portals for Slavic languages also provide various linguistic resources and linguistic information, e.g. for Croatian [2], Czech [3], Slovenian [4].

Several Bulgarian resources developed at the Institute for Bulgarian Language have been integrated and made accessible via the web portals Infolex [5] and Neolex [6] providing access to an extensive lexicographic database (dictionary of synonyms, dictionary of antonyms, dictionary of phraseology, dictionary of neologisms).

2. Dictionary Resources

One of the main objectives of the Institute for Bulgarian Language is the study and description of the Bulgarian language and the codification of its literary norms. As a result of decades of lexicographical work, a large number of dictionaries and reference books have been published. For the purposes of various research projects, a number of electronic corpora and databases have been compiled to support the work and facilitate the access to unique language resources in specialised fields. In recent years considerable effort has been invested into making the language resources available online through digitisation of printed dictionaries, providing

interface and search facilities to electronic databases, and integration between various resources.

2.1 Dictionary of Bulgarian Language

The most significant lexicographical work of the Institute for Bulgarian Language to date is the multi-volume Dictionary of Bulgarian (DB). In the period from 1977 to 2015, a total of 15 consecutive volumes of the DB were printed (from letter A to letter P), and work has begun on the last volumes. In the meantime, the revision and updating of the first 8 volumes has started, where the revised volumes 1, 3 and 4 have already been printed.

The multi-volume DB reflects the current state of the Bulgarian vocabulary and the development of the lexical and semantic system from the second quarter of the 19th century until modern times (Cholakova 2010). The DB includes headwords, word senses and examples attested in the literature of the Bulgarian National Revival and of the period following the Liberation, as well as in fiction, science and popular-science, journalistic publications and periodicals of the 20th and the 21st century, and in the spoken language. The dictionary has been available on the Internet for two years [7] and it is the most popular and widely used language resource of the Institute.

2.2 Bulgarian WordNet

The Bulgarian WordNet (Koeva 2010) is a large lexical-semantic network developed after the model of Princeton WordNet. Currently, it contains over 120,000 synsets (approximately 63,000 of which are manually checked) interconnected through a rich set of semantic, morpho-semantic, derivational and extralinguistic relations (a total of over 256,000 links). WordNet is one of the most complete and consistent lexical resources (the literals in the Bulgarian WordNet are much greater in number – over 135,000, than the word list in a standard spelling dictionary). The rich linguistic information provided by the Bulgarian, Romanian and Princeton WordNet is accessible online [8].

2.3 Grammatical Dictionary of Bulgarian

The dictionary is based on the spelling dictionaries developed at the Institute for Bulgarian Language. Its inflectional types are described in the finite state technology framework. The dictionary is used for the development of the Bulgarian Spell Checker [9] and the Online Consultations Service [10]. Comments and recommendations for the Online Spell Checker are collected via an online form [11].

2.4 Dictionary of Old Bulgarian

The Dictionary of Old Bulgarian presents the vocabulary of classical Old Bulgarian in the original epigraphic monuments and in Old Bulgarian manuscripts from the 10th-11th century (Ivanova-Mircheva et al. 1999; 2009). The Dictionary is available through the Electronic library of the Institute for Bulgarian Language [12]. Currently, a database and an interface are being developed, to provide easy access to the Dictionary and to add references between the old Bulgarian words and their modern

counterparts.

3. Towards an Integrated Dictionary Portal

The main function of dictionary portals is to provide access to a set of dictionaries in a way that the dictionary users obtain the information they need in an easy and efficient manner. Our work is focused on building an integrated dictionary portal which links the following resources and linguistic information: the Grammatical Dictionary of Bulgarian (information about the correct spelling of the word and its grammatical characteristics), the Dictionary of Bulgarian Language (thesaurus information), the Bulgarian WordNet (definition and links to synonyms and other semantically related words, English translation), the Dictionary of Old Bulgarian (information about the original form of the word in Old Bulgarian). Further, more resources will be integrated: Infolex and Neolex for referring to synonyms, antonyms, neologisms, and phraseology; brief reference book of Bulgarian grammar [13], etc.

The four dictionaries (available independently on the website of the Institute [14]) are now linked into a dictionary net (Engelberg and Mueller-Spitzer 2009) [15]. The user can select among the four dictionaries to search for a word. From the Grammatical Dictionary of Bulgarian all word forms of the queried word are displayed with information about their grammatical characteristics. The lemma of the queried word also links to any of the other resources it appears in: the Dictionary of Bulgarian, the Bulgarian WordNet and the Dictionary of Old Bulgarian. Further, from the portal there are links to the Online Language Consultations Guide [16], the Facebook Language Consultations page [17] and an online interactive language game [18] (testing knowledge from various areas of the Bulgarian language). The users can provide feedback and can ask questions about Bulgarian vocabulary, grammar, phraseology and spelling using the web question form of the Online Language Consultations Guide and on the Facebook page where answers are provided within an hour.

Further development of the portal will aim at extending the cross-references between the dictionaries by creating a set of virtual microstructures upon query which will combine data from the different sources to present a complex unified multi-layered description of the headword. Additionally, the Bulgarian WordNet allows the extension of the portal to a bilingual (and possibly multilingual) portal with focus on Bulgarian.

4. User Feedback on the Language Resources

Our work on the Dictionary Portal is part of the initiative of the Institute to provide public access to the language resources and technologies which can be applied in various areas, such as research, education, public services, etc. It is closely related to the Language Consultation Service (available as an Online guide, on Facebook and Twitter) answering queries and providing information and advice regarding language-related problems, common mistakes and difficulties in language usage, specific language phenomena, etc.

The feedback we collect from users of the language resources developed at the Institute for Bulgarian Language is aimed at identifying: (a) the most problematic areas of Bulgarian spellings, grammar and lexical system, which need to be widely

addressed and clarified; (b) language phenomena that the wide public is interested in, such as the usage of foreign words or dialect expressions, the appearance of new words, etymology of words, etc.; (c) the needs of the education in Bulgarian language.

We collect user data in the following ways:

(1) Questions and comments from users received by phone or email, on the Online Language Consultations Guide, the Facebook Language Consultations page or Twitter. Questions regard mostly spellings and grammar and are answered within one hour. Frequently asked questions are then answered in details in separate posts in the Online Guide with explanations, links to dictionaries and examples.

(2) Popular topics or problematic language-related questions discussed or observed on conventional and social media in general - questions and comments on TV, radio, peculiar use of words and language, effect of political and social events on language. These are also published on the blog.

(3) Comments received from teachers in Bulgarian language during qualification courses or special school-based events carried out by the Institute. Teachers share their experience and the need for any specific resources to facilitate the lessons.

(4) Comments received from the wide public during the special events organised to promote the Bulgarian language and culture dedicated to national holidays and the European Day of Languages.

Additional source of usage data comes from Google Analytics, which identifies the Dictionary of Bulgarian as the most frequently used resource. However, the questions received are predominantly focused on spellings and grammar rules which suggests that the other resources need to be more widely promoted. The portal will also serve that purpose since when a user accesses one of the resources, they will also receive information about the others.

5. Conclusions

The Institute has compiled a large number of electronic resources and published a vast amount of printed dictionaries. Effort in recent years resulted in making many of these resources available online which supports the Language Consultation Service by providing extensive and multi-layered information about language usage. Our work in the near future is oriented towards optimising user interface to our resources by integrating the resources into a uniform dictionary portal which: provides easy and efficient access to the high quality dictionary content developed at the Institute; offers the opportunity to compare various types and sources of information; collects user data and user feedback, and meets their most frequent needs (answering questions regarding the spelling and meaning of words, giving usage examples, etc.); provides mobile user interface and plain text interface (UTF-8) allowing access for text-to-speech applications.

6. References

Cholakova, K. Predgovor kam parvoto izdanie na Rechnik na balgarskiya ezik. *Rechnik na balgarskiya ezik, tom parvi, vtoro dopalneno i preraboteno izdanie*. Sofia. 2011: Akademichno izdatelstvo "Prof. Marin Drinov", ET "EMAS", 2011. [In Bulgarian]

Engelberg, Stefan and Carolin Mueller-Spitzer. Dictionary Portals. In: Rufus H. Gouws, Ulrich Heid, Wolfgang Schweickard & Herbert Ernst Wiegand (Eds.): *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography.* Berlin/New York: de Gruyter, 2009.

Ivanova-Mircheva, Dora (Ed.). 1999. *Dictionary of Old Bulgarian*. Vol. 1 (A-I). Sofia: Valentin Trayanov Publishing House. [In Bulgarian]

Ivanova-Mircheva, Dora (Ed.). 2009. *Dictionary of Old Bulgarian*. Vol. 2 (O-X). Sofia: Valentin Trayanov Publishing House. [In Bulgarian]

Koeva, Svetla. Bulgarian WordNet – current state, applications and prospects. – In: *Bulgarian-American Dialogues*, Prof. M. Drinov Academic Publishing House Sofia, 120-132, 2010.

7. Online sources

[1] http://dictionary.reference.com/

[2] http://hjp.znanje.hr/

[3] http://prirucka.ujc.cas.cz/

[4] http://www.fran.si/

[5] http://ibl.bas.bg/infolex/

[6] http://ibl.bas.bg/neolex/bg/

[7] http://ibl.bas.bg/rbe/

[8] http://dcl.bas.bg/bulnet/

[9] http://dcl.bas.bg/est/checker.php

[10] http://dcl.bas.bg/est/dict.php

[11] http://dcl.bas.bg/est/comments/

[12] http://dcl.bas.bg/lib/Starobalgarski_rechnik_tom1/

[13] http://znam.bg/com/action/showArticle?encID=693&article=619863583

[14] http://ibl.bas.bg/en/

[15] http://ibl.bas.bg/dictionary_portal/

[16] http://ibl.bas.bg/ezikovi_spravki/

[17] https://www.facebook.com/ezikovi.spravki/

[18] http://ibl.bas.bg/ezikova_igra/