Towards a Multimodal WordNet for Language Learning in Bulgarian

Petya Osenova[0000-0002-4484-5027], Kiril Simov[0000-0003-3555-0179]

Institute of Information and Communication Technologies, Bulgarian Academy of Sciences,
Acad. G. Bonchev Str., bl. 2, Sofia, Bulgaria
petya@bultreebank.org, kivs@bultreebank.bg

Abstract. In this paper we present some modifications to and extensions of a Wordnet for Bulgarian designed to make it more appropriate for applications in the area of language learning. However, in order to support education, we need to ensure the appropriate selection of sets of synonyms (synsets) from BTB-Wordnet, depending on the education level of the learners, and various types of exercises based on integration of the learning topic and semantic information within Wordnet. For this purpose, our focus is mainly on the combination of the lexemes (lemmas), with their meanings and examples, and the specially designed pictures as illustrations of those meanings within the synsets. We report on our preliminary results.

Keywords: Wordnet, Sub-lexicons, Language Learning, Images, Bulgarian.

1 Introduction

Language resources for general usage (such as Wordnet, Treebank and others) sometimes allow for fine-tuning with regard to a given application without changing their applicability to the initial goal. In this way, with a small amount of effort one resource could be used in several types of application.

In this paper we present some modifications and extensions of a Wordnet for Bulgarian – BTB-Wordnet – in order to better tune it for applications in the language learning arena. A ‘Wordnet’ is an electronic lexicon that contains semantic information about the lexemes in the language. The lexical entry is called a *synset* and is structured around a given sense. It comprises the following elements: a definition of the meaning expressed in natural language; a set of synonyms which share this sense; examples for the usage of these terms within this sense; relations to other synsets in the Wordnet. In the process of modifying this lexical resource for the purposes of education we had to adapt: (1) *the set of lemmas*. This is a necessary step because some of the lemmas for a given synset might not be appropriate for the status (intellectual or social) of the learner. Such problematic lemmas might be dialectal, slang terms, or register in other ways; they may also be terminologically too complex or socially very sensitive. (2) *the definitions*. The definition must explain the meaning of the concept as clearly as possible. It has to be formulated in a suitable language with words that are

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 14, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2024. ISSN: 1314-4006, eISSN: 2535-0366

understandable to the learner but at the same time, provide all the necessary information about the notion. (3) *the examples*. Some examples are assigned to the lemma in each synset but they also need to be adjusted in order to reflect the knowledge level of the learner. While the step in (1) is socio-intellectually-bound, those in (2) and (3) are related to the complexity of language and the cognitive capability at various learning levels. In this work we use the resource that was created within our group, BTB-Wordnet (Simov & Osenova, 2023).

It is well known that language learners can be classified into different groups. The two main ones are: (1) native speakers of a given natural language whose mastery varies across their age and learning curves; and (2) second language learners whose mastery also varies across their learning curves but in addition relates to their backgrounds, ambitions and motivation.

The paper is structured as follows: in section 2 the strategy for lemma selection is presented; in section 3 data manipulation in respect of parts of speech is discussed. Some preliminary validation of the relationship between definition and image is also outlined here. Section 4 describes the inclusion of images in the existing platform and the benefits thereof. In addition, some future lines of work are listed.

2 Data Selection

Our work started with the selection of a list of lemmas that correspond to various educational levels in the learning of Bulgarian lexica. Preparation of the set for native speakers turned out to be more difficult since, to our knowledge, there are no publicly available lists of minimum lexica that have to be learnt for each school subject at each grade level. For that reason, we started with excerpts from published spelling dictionaries that target certain learning levels. These are (Radeva & Burov, 2009) with 12,843 lexemes for the basic education level (1–4 grade) and (Mateeva & Karlova, 1993) with 8,109 lexemes for students in the 4th to 8th grades of high school. We also extracted the lexica from the Semantic Minimum Dictionary (Kasabov, 1990) which amount to 1,715 lexemes. All these words were added to the Bulgarian BTB-Wordnet. Thus, from a pedagogical point of view, we rely on the lexeme selection undertaken by the authors of the dictionaries mentioned.

For the second group of learners, who study Bulgarian as a second language we made took extracts of lemmas from textbooks for foreign learners of Bulgarian at levels A1–A2 (Kurteva et all, 2016) and B1–B2 (Kurteva aet all, 2017). These are textbooks published for foreigners regardless of the purpose of studying Bulgarian. We have also included similar lists of lemmas, but from textbooks having a target group of learners – refugees in Bulgaria. In this case parts of the vocabularies include some specific lemmas (Andonova et all, 2014a; Andonova et all, 2014b). These were also added to the BTB-Wordnet.

The addition of all these lemmas to the vocabulary of BTB-Wordnet (when they were missing) was done by the creation of a new synset for each new sense per lemma. In this way we ensure that BTB-Wordnet covers the full set of meanings for each of the lemmas in the vocabulary. To apply this in teaching we also have to select the appropriate meaning. For example, ‘table’ as *a piece of furniture* is more basic than a ‘table’ as *a set of data arranged in rows and columns*.

At present there are only a few more than 800 images as there are challenges in portraying some of the meanings.

In the next section we describe the steps taken in manipulating these lists of lemmas in order to make them usable in different teaching scenarios.

3 Data Manipulation

Let us discuss, in more detail, our approach to resource manipulation in respect of education. Our focus is on the interrelation of definitions, examples and their illustration in images. Thus, we may enhance multimodality and provide a strategy for verifying the lexical data. It should be noted that the so-called ‘ImageNet’ (Deng et al., 2009) is available. This resource is aligned with the English wordnet synsets and could easily be transferred to other languages if they have Wordnets aligned to English Wordnet. However, ImageNet’s mission is different - it provides visuals for image recognition tasks and these are not owned by the ImageNet developers. In our case, the images that cover some basic notions are created for the purposes of learning. For this reason, the illustrations of the different synsets have to follow that same style of presentation. ImageNet is not such a resource. However, in future we also plan to use pictures that are freely available. They will be selected through image recognition techniques.

Initially, the definitions of notions for drawing illustrations were selected from the textbooks for foreign learners at levels A1–A2. Also, we focused on nouns only. The reasons were as follows: i) at these levels more concrete meanings are expected (in comparison with abstract ones which are more complex) and ii) nouns are easier to present as images.

However, later on, verbs and adjectives were also selected for image representation together with some more abstract nouns. We did not opt to use ready-made pictures that are authorship-free but hired an artist who prepared illustrations specifically for this task. We also decided not to choose colourful images but rather graphical illustrations. The artist received the following information: the lemma, its semantic class (e.g. concrete/abstract noun) and the definition. All this information came from the BTB-Wordnet. The artist was not instructed what to draw. He had the freedom to apply his own approach. Let us discuss the pictures with respect to the parts of speech.

3.1 Nouns

Here is an example of a concrete noun: {аптека *noun.artifact* **LEMMA**: аптека **DEF:** Заведение, в което се приготвят и продават лекарства (farmacy *noun.artifact* **LEMMA**: farmacy **DEF:** A place where medicines are prepared and sold)}.

Please note that in the **LEMMA** slot synonyms can be added, if available. For example, let us consider an abstract noun: {злополука *noun.event* **LEMMA:** катастрофа, злополука, произшествие **DEF:** Нещастен инцидент, причинен от сблъсък между две или повече превозни средства, често придружен с човешки жертви или наранявания (disaster *noun.event* **LEMMAS:** catastrophe, disaster **DEF:** An incident caused by a collision between two or more vehicles, often accompanied by human casualties or injuries). In Figure 1 an illustration of a shoe is given which is a clear and easy case. Here is the synset: {обувка *noun.artifact* **LEMMA:** обувка, чепик, патък, обущък **DEF:** Изделие от кожа или друга материя, обикновено с твърда подметка, което се обува и се носи на краката при ходене, за да ги предпазва от студ, влага или нараняване (shoe *noun.artifact* **LEMMA:** shoe **DEF:** An artefact of leather or other material, usually with a hard sole, that is put on and worn on the feet when walking to protect them from cold, moisture, or injury)}.

In the set of nouns there were various challenges on how to present, for example, the units of time like days of the week, months, etc. For days and months some wording was used in the pictures. But for units like a minute or an hour such a strategy was not possible.

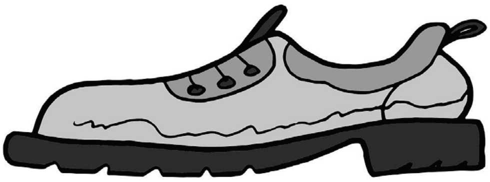


Fig. 1. The picture of the lemma shoe.

It should be noted that in the synset also dialectic lemmas are present. In this way the language heritage is preserved and the knowledge about it is expanded.

3.2 Verbs

Verbs, mainly those for ‘motion’, ‘contact’ and ‘change’ are easier to draw in general as they often denote specific actions. For example, {летя *verb.motion* **LEMMA:** летя, хвърча, хвъркам **DEF:** За птица, насекомо и други - движа се, нося се във въздуха (to fly *verb.motion* **LEMMA:** fly **DEF:** For birds, insects and others – moving, floating in the air)}. On the other hand, the following verb is abstract and difficult to present as an image: разглеждам *verb.cognition* **LEMMA:** разглеждам, проучвам, разгледам, проуча, преглеждам, прегледам **DEF:** Изучавам от всички страни, всестранно се запознавам с някого или нещо (look into *verb.cognition* **LEMMA:** examine, study, review, look into **DEF:** to study from all sides, to get to know someone or something from all sides)}.

In Fig. 2 the picture of the verb *write* is presented. Here is the synset: {пиша *verb.creation* **LEMMA:** пиша, написвам, напиша, изписвам, изпиша, прописвам, пропиша **DEF:** Чертая, изобразявам букви, цифри и други знаци, обикновено с молив, писалка или на компютър (write *verb.creation* **LEMMA:** to write **DEF:** to draw, depict letters, numbers, and other signs, usually with a pencil, pen, or computer)}.

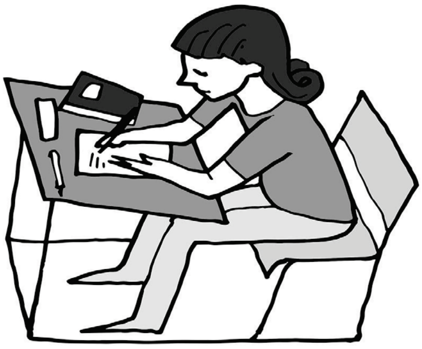


Fig. 2. The picture shows the activity of writing.

3.3 Adjectives

Illustrations of adjectives are very hard to draw and, in most cases, impossible. Thus, we decided to start with the antonymic pairs since the contrast might help in deriving the meaning of the antonyms, and many of the basic adjectives included in the above-mentioned textbooks have their antonyms given. In many textbooks for the learning of a foreign language at earlier levels, adjectives are represented in antonymic pairs. Thus, if we want to represent the adjective *гладен* (hungry), we also provide its opposite – *сит* (full). The adjectives in our resource do not have semantic categories since such categories were missing in the English wordnet. In future we plan to add semantics. For that reason, for the time being, one general semantic category was used only - adj.all. Compare: {adj.all **LEMMA:** гладен, ненахранен **DEF:** Който усеща, изпитва глад, който има нужда, потребност от храна (adj.all **LEMMA:** hungry, unfed **DEF:** One who feels, experiences hunger, who needs food.) **vs.** adj.all **LEMMA:** сит **DEF:** Който не чувства глад, който се е нахранил.(adj.all **LEMMA:** full **DEF:** He who does not feel hungry, who has eaten)}. On Fig. 3 the picture is given of the pair *бърз – бавен* (fast – slow). The adjectives are exemplified by a snail and a rabbit:

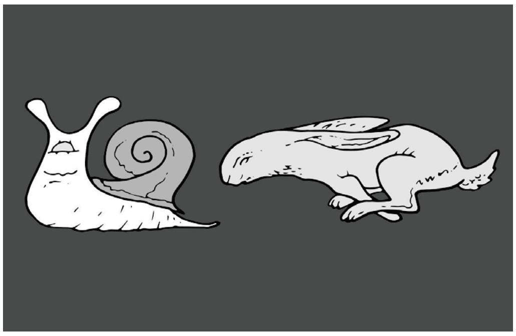
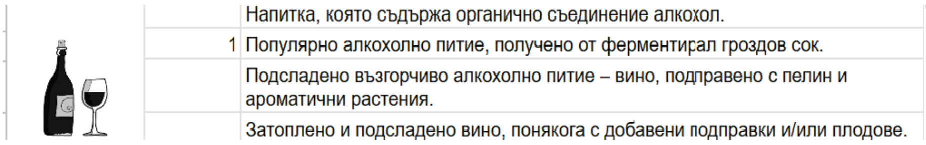


Fig. 3. The snail illustrates the idea of being slow while the rabbit – of being fast.

3.4 Preliminary Validation

In order to test how self-explanatory the pictures are, we decided to apply a reverse mechanism. We put each picture in a file, together with four possible definitions and asked 6 students to assess these options, selecting the one they thought corresponded to the image. 122 tasks were assessed. All 6 students were required to assess the same dataset independently of each other. They were encouraged to provide comments in cases where they faced problems. An excerpt of the validation task is shown in Fig. 4. It can be seen that the second option from the top has been chosen as the correct one. It is marked with number 1 for validation. The definitions supplied contain the correct one (as provided to the artist) as well as others that might be very close to, or far from, the correct meaning. The ones that are very close in meaning, or follow the same hierarchy are the most difficult to address. The algorithm for selection of alternative definitions is working over the graph structure of the Wordnet as defined by (Hirst & St-Onge, 1996) for constructing lexical chains. This algorithm can be used to construct tasks with harder alternative definitions (i.e. semantically closer to the true one) or easier ones that are simpler to distinguish from the true one. In this way tasks could be tuned to the needs of learners.

Here the definitions, in descending order are as follows: *A drink that contains the organic compound alcohol*; *A popular alcoholic drink made from fermented grape juice* (selected as the correct one); *Sweetened bitter alcoholic drink – wine seasoned with wormwood and aromatic plants*; *Warmed and sweetened wine, sometimes with added spices and/or fruit*:



		Напитка, която съдържа органично съединение алкохол.
	1	Популярно алкохолно питие, получено от ферментирал гроздов сок.
		Подсладено възгорчиво алкохолно питие – вино, подправено с пелин и ароматични растения.
		Затоплено и подсладено вино, понякога с добавени подправки и/или плодове.

Fig. 4. The image is presented on the left, while on the right all definition options are given.

As can be seen, the first definition is rather general. The second is the neutral description of wine, while the third and fourth are specific types of wine, either when initially prepared (the third), or later in the kitchen – with temperature and content manipulated (the fourth option).

The deviation from definitions given as correct, was 20.36 %. Behind this metric, various factors are hidden. On the one hand, the assessor may have been given the option to choose only from close definitions. Alternatively, some pictures turned out to be ambiguous. For example, one picture showed a road with houses around it. Among the definitions there is one for a road and one for a town. In this instance, one of these is selected and then commented upon. Thus, the combination of the image with definitions might cause some confusion in certain cases. In order to avoid such problems, tutors will have the opportunity/ability to modify the optional definitions in order to make the task appropriate for students.

The pictures cover some basic and some more complicated lemma meanings. For the latter case various approaches have been used where applicable. The most popular of these is the combination of an image with numbers or words. For example, the month of January is presented as a calendar list where number 1 is shown.

Needless to say, the pictures will be used in conjunction with the dictionary and text appropriate to the student level. Thus, some clarity will be achieved only by this contextualization. Our validation aimed at finding the difficulties of human perception when an image and a few meanings are presented. We believe that such experiments may be useful to create better learning scenarios.

4 Conclusions

Pictures will be incorporated, as an additional level, into the existing platform for the 'meaning' exercises and creation of tests. It should be noted that a successful application depends on the quality of the Wordnet (definitions, examples, hierarchy) and on the embedded algorithm. As already mentioned, some meanings are very difficult to present in an image since they are too abstract or ambiguous. Thus, one line of future work will be to invent a process that does not strictly follow level stratifications – words, texts, images – but combines them in the best way for successful cognitive perception. Several types of relationships will be established in respect of comprehension, among which: i) an image is self-explanatory without the need to check the meaning of the word; ii) an image needs some numbers or words to be included in order to be comprehensible; iii) the image cannot effectively represent the meaning, it can only illustrate it.

Cultural heritage also includes dialect language data. These data are being included in BTB-Wordnet. Through images learners of Bulgarian can extend their knowledge of Bulgarian vocabulary and concepts beyond the information in normative dictionaries.

Acknowledgements.

This work was financially supported by the Bulgarian Ministry of Education and Science, DO1-301/17.12.21: Bulgarian National Interdisciplinary Research e-

Infrastructure for Resources and Technologies in favour of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH.

References

Andonova, M., Sabeva, R., & Zagorova, Z. (2014a). *Bulgarian for refugees: A1.* Caritas Sofia.

Andonova, M., Sabeva, R., & Zagorova, Z. (2014b). *Bulgarian for refugees: A1.* Caritas Sofia.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248-255). IEEE. https://doi.org/10.1109/CVPR.2009.5206848

Hirst, G., & St-Onge, D. (1996). *Lexical chains as representations of context for the detection and correction of malapropisms.* The MIT Press.

Kasabov, I. (1990). *Semantichen rechnik-minimum* [Semantic dictionary-minimum]. St. Kliment Ohridski University Press.

Kurteva, G., Bumbarova, K., & Bachvarova, S. (2016). *Zdraveyte! Uchebnik po balgarski za chuzhdentsi (nivo A1 - A2)* [Hello! Textbook Bulgarian for foreigners: (A1-A2)]. New Bulgarian University.

Kurteva, G., Bumbarova, K., & Bachvarova, S. (2017). *Zdraveyte! Uchebnik po balgarski za chuzhdentsi (nivo B1-B2)* [Hello! Textbook Bulgarian for foreigners (B1-B2)]. New Bulgarian University.

Mateeva, A., & Karlova, R. (1993). *Pravopisen rechnik na savremenniya balgarski ezik za uchenitsi ot 4 do 8 klas* [Spelling dictionary of the modern Bulgarian language for students from 4th to 8th grade]. Zhelev-Gegov.

Radeva, P., & Burov, S. (2009). *Ucheben pravopisen rechnik na balgarskiya ezik za uchenitsite ot nachalniya kurs* [Educational spelling dictionary in the Bulgarian language for elementary school students]. Slovo.

Simov, K., & Osenova, P. (2023). Recent Developments in BTB-WordNet. In *Proceedings of the 12th Global Wordnet Conference* (pp. 220-227). Global Wordnet Association. https://aclanthology.org/2023.gwc-1.27

Received: March 24, 2024
Reviewed: April 12, 2024
Finally Accepted: April 28, 2024