BULGARIAN ACADEMY OF SCIENCES

CYBERNETICS AND INFORMATION TECHNOLOGIES • Volume **20**, No 4
Sofia • 2020 Print ISSN: 1311-9702; Online ISSN: 1314-4081
DOI: 10.2478/cait-2020-0051

Linguistic vs Encyclopaedic Knowledge. Classification of MWEs from Wikipedia Articles

Z. Kancheva, I. Radev

Institute of Information and Comunication Technologies, Bulgarian Academy of Sciences, 1113 Sofia, Bulgaria
E-mails: zara@bultreebank.org, radev@bultreebank.org

Abstract: *This paper reports on the first steps in the creation of linked data through the mapping between the synsets of BTB-WordNet and the articles in Bulgarian Wikipedia. The task of expanding the BTB-WordNet with encyclopaedic knowledge is done by mapping its synsets to Wikipedia articles with many MWEs found in the articles and subjected to further analysis. We look for a way to filter the Wikipedia MWEs in the effort of selecting the ones most beneficial to the enrichment of BTB-WN.*

Keywords: *MWEs, Wordnet, Wikipedia, Semantic Mapping.*

1. Introduction

Research in the field of Natural Language Processing (NLP) shows that traditional language resources do not give the best performance in every NLP task when used by themselves. In recent years researchers started to put together various lexical resources in projects such as BabelNet [12], SemLink [15], Predicate Matrix [2] and Uby [3]. The task to build relations between linguistic and semantic resources and to use this kind of new data for generating knowledge graphs has proven to be beneficial for languages with less lexical resources.

With the development of huge electronic corpora and advancements in corpus linguistics Multi Word Expressions (MWEs) receive more and more attention from researchers. A paper [18] estimates that the number of MWEs in the lexicon of a person is more than 40%. MWEs are one of the main challenges from the perspective of NLP. Due to their distinctive nature, many areas in NLP such as parsing, machine translation, key phrase or index term extraction, and language acquisition research can benefit from tackling MWEs. That but MWEs not only are omnipresent in all text data and cannot be skipped in tasks such as word sense disambiguation, named entity linking and co-reference resolution.

There are many projects that aim to integrate the two types of knowledge – of the linguistic system and of the world [3], most commonly by the merge of a dictionary or WordNet with Wikipedia or Wiktionary, but still it is challenging to estimate how much and what types of encyclopaedic information is useful to add to

a language resource. The task is even more difficult if the focus is on the MWE distribution in the resulting dataset.

During the process of manual mapping between Bulgarian WordNet (BTB-WN) and Wikipedia with the use of the CLaRK system [20] we plan to introduce to BTB-WN all MWEs related to the mapped Wikipedia articles. Usually these MWEs are with a head word corresponding to the title of the Wikipedia article – for example, "Wine" vs "Red wine", "White wine", "Sparkling wine", etc. From a linguistic perspective this determines relation head-dependent. From semantic point of view, the relations are more diverse. In this mainly we determine sub-concepts, but by different features.

The structure of the paper is as follows: the next section outlines the related work. Sections 3 and 4 explore the different approaches for classification of MWEs. Section 4 discusses the open questions. Section 5 concludes the paper.

2. Related work

Some of the most outstanding works on the alignment of linguistic and encyclopaedic knowledge with WordNets are: BabelNet – combining multilingual WordNet and Wikipedia; Uby – combining WordNet, GermaNet, Wiktionary, Wikipedia, FrameNet and VerbNet for English and German; the mapping of the Princeton WordNet with the English Wikipedia [11]; and the mapping of the plWordNet onto the Princeton WordNet [17].

For the purposes of this research we work with the BTB-WN [13], which was built in several steps. It started as a translation of Core WordNet and was expanded with concepts from Bulgarian Treebank (BulTreeBank [14]), frequency list and currently Bulgarian Wikipedia. At the moment BTB-WN contains about 25 000 synsets – the last 15 percent of them came from the expansion with around 13 000 articles from Bulgarian Wikipedia in attempt to map it to the BTB-WN [21].

Currently the Wikipedia in Bulgarian has 259,927 content articles, which makes it about 23 times smaller than the English version, but it is still a very useful resource to extract world knowledge from. It contains data for concepts (similarly to WordNet) and instances of concepts – Notable named Entities (NEs) for persons, locations and events (often excluded in WordNet). Building knowledge graphs upon the relations between concepts and their instances and using these graphs to train, test and improve NLP systems is deemed to be very impactful in positive manner. Being a communal free to use and edit resource, Wikipedia is constantly expanding with new articles and reflects the creation of new inventions and products or the emergence of new celebrities and events.

Recent paper [9] presents an overview of MWEs in BTB-WN, where the MWEs are presented as several types of phrases by their headword: multiword Nouns (Noun + Noun; Adjective + Noun; Numeral + Noun); Verbs (Verb + Noun; Verb + Adverb; Verb + Prepositional Phrase); Adjectives (Adverb + Adjective; Adjective + Prepositional Phrase) and Adverbials (Preposition + Noun; Preposition + Adjective; Adverb + Adverb) in accordance with the classification developed within WG 4 of PARSEME COST Action (**https://typo.uni-konstanz.de/parseme**); and treated

afterward with a catena-based modeling. Working with the same resource we are using the same classification method in our work.

A similar approach in dealing with MWEs is presented in [7]. The paper reports on classification of MWEs based on morphosyntactic, structural and semantic criteria and using semi-automatic methods to compile a MWE dictionary for Bulgarian. The work discusses a repository of 86,373 "nominal" and "verbal" MWEs, based on the head word.

Researchers continue to develop MWE corpora. One very recent example for dedicated MWEs dataset is reported in [8]. This monolingual dataset contains annotated MWEs for Swedish with focus on compositionality and has proven to be beneficial for evaluation of computational models. Another recent attempt at dedicated MWEs dataset is shown in [4]. The authors share their efforts on creation of multi-lingual and bilingual MWE corpora containing bilingual MWE pairs for German-English and Chinese-English. They also report on the impact of the dataset on machine translation once more proving the need of resources with MWEs for the purposes of several NLP tasks.

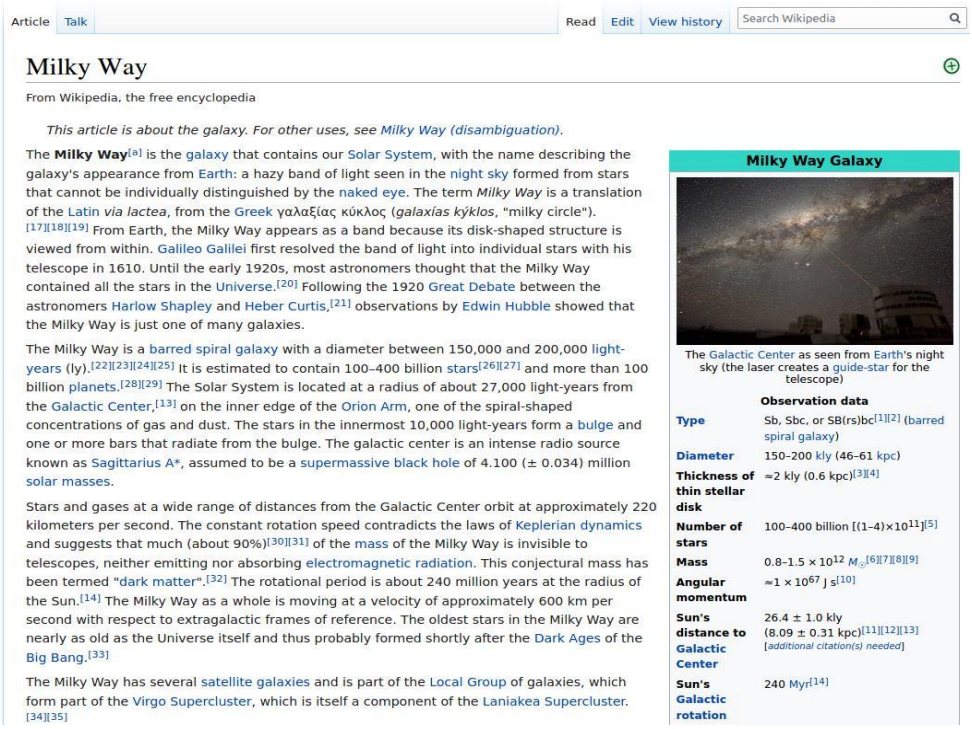
Another interesting work on MWEs is that of [19] where they experiment on different strategies for extracting headless MWEs (such as named entities, dates and others) from dependency parse trees. The work of [16] shows the benefits of using verbal MWEs in the task of metaphor identification and classification. They present a neural model that classifies metaphorical verbs with dependency parse trees and annotated verbal MWEs.

3.Wikipedia MWEs dataset

For the aims of this research we have semi-automatically extracted 13,173 MWEs from 14,512 Wikipedia pages. These pages contained over 30,000 links to other pages in Wikipedia from which manually we selected "true" MWEs, excluding person names, annual events (13th/14th Summer Olympics), titles of movies, books, music albums and songs. The initial set of 14,512 Wikipedia pages was the basis of mappings between Bulgarian Wikipedia and BTB-WN [9]. So far 1628 MWEs were preliminary added as synsets in the BTB-WordNet without being domain classified; 2187 MWEs have been domain classified in preparation to be added as synsets and 9358 MWEs are ongoing the process of domain classification.

It is important to outline (though, it was somehow predictable) that there are no proverbs, metaphorical expressions and verbal constructions among the extracted MWEs, because of the characteristics of the Wikipedia content, which most frequently concerns entities and events, constructed by nouns and adjectives. Typically, Wikipedia contains articles about famous geographical objects and terminology of different fields of science. MWEs that are proper names and terms may not be of the greatest interest for linguists, but they are valuable for our current research. Various NLP tasks need both linguistic and encyclopaedic knowledge, thus enriching BTB-WN with as much as possible synsets will be beneficial for our work. Also, this type of data can be used in further modelling of MWEs.

The intention of the research at this stage is to focus exactly on common domains and less on their specific subclasses (for now), so we will classify the MWEs on the one hand by the science branch that they belong to, and on the other hand – by their linguistic features.


Article Talk Read Edit View history Search Wikipedia

Milky Way

From Wikipedia, the free encyclopedia

This article is about the galaxy. For other uses, see Milky Way (disambiguation).

The **Milky Way**[a] is the galaxy that contains our Solar System, with the name describing the galaxy's appearance from Earth: a hazy band of light seen in the night sky formed from stars that cannot be individually distinguished by the naked eye. The term *Milky Way* is a translation of the Latin *via lactea*, from the Greek γαλαξίας κύκλος (*galaxías kýklos*, "milky circle").[17][18][19] From Earth, the Milky Way appears as a band because its disk-shaped structure is viewed from within. Galileo Galilei first resolved the band of light into individual stars with his telescope in 1610. Until the early 1920s, most astronomers thought that the Milky Way contained all the stars in the Universe.[20] Following the 1920 Great Debate between the astronomers Harlow Shapley and Heber Curtis,[21] observations by Edwin Hubble showed that the Milky Way is just one of many galaxies.

The Milky Way is a barred spiral galaxy with a diameter between 150,000 and 200,000 light-years (ly).[22][23][24][25] It is estimated to contain 100–400 billion stars[26][27] and more than 100 billion planets.[28][29] The Solar System is located at a radius of about 27,000 light-years from the Galactic Center,[13] on the inner edge of the Orion Arm, one of the spiral-shaped concentrations of gas and dust. The stars in the innermost 10,000 light-years form a bulge and one or more bars that radiate from the bulge. The galactic center is an intense radio source known as Sagittarius A*, assumed to be a supermassive black hole of 4.100 (± 0.034) million solar masses.

Stars and gases at a wide range of distances from the Galactic Center orbit at approximately 220 kilometers per second. The constant rotation speed contradicts the laws of Keplerian dynamics and suggests that much (about 90%)[30][31] of the mass of the Milky Way is invisible to telescopes, neither emitting nor absorbing electromagnetic radiation. This conjectural mass has been termed "dark matter".[32] The rotational period is about 240 million years at the radius of the Sun.[14] The Milky Way as a whole is moving at a velocity of approximately 600 km per second with respect to extragalactic frames of reference. The oldest stars in the Milky Way are nearly as old as the Universe itself and thus probably formed shortly after the Dark Ages of the Big Bang.[33]

The Milky Way has several satellite galaxies and is part of the Local Group of galaxies, which form part of the Virgo Supercluster, which is itself a component of the Laniakea Supercluster.[34][35]

Milky Way Galaxy

The Galactic Center as seen from Earth's night sky (the laser creates a guide-star for the telescope)

Observation data

Type	Sb, Sbc, or SB(rs)bc[1][2] (barred spiral galaxy)
Diameter	150–200 kly (46–61 kpc)
Thickness of thin stellar disk	≈2 kly (0.6 kpc)[3][4]
Number of stars	100–400 billion [(1–4)×10^{11}][5]
Mass	0.8–1.5 × 10^{12} $M_{\odot}$[6][7][8][9]
Angular momentum	≈1 × 10^{67} J s[10]
Sun's distance to Galactic Center	26.4 ± 1.0 kly (8.09 ± 0.31 kpc)[11][12][13] *[additional citation(s) needed]*
Sun's Galactic rotation	240 Myr[14]

Fig. 1. Wikipedia article with astronomy MWE

One thing that we must keep in mind is the nature of Wikipedia itself. Being open to the wide public may lead to problems with the data it provides. We observed that one of the possible issues with Wikipedia is concerning orthography – everyone could write an article in it and there is no spelling check. This leads to a problem for the automatic data extraction – some of the excerpted lexical entries in our dataset are not MWEs, they are just misspelled. In the data we had (*bas kitara*, "bass guitar") which is not a MWE and should be spelled as one word according to the Bulgarian dictionary. Examples like this show that the manual check and selection of automatically extracted MWEs is a very important step of the processing, although it is very time-consuming.

4. Classification of MWEs

MWEs could be defined as "lexical units larger than a word that can bear both idiomatic and compositional meanings" [10]. Due to their nature there is no single generally accepted typology or classification of the MWE. Different researchers classify them at several levels – morphological, lexicological, syntactical and semantical; for example [22] where the paper uses different term for the same

linguistic phenomena – fixed expressions – and describes them as "combinations of two or more words that are typically used to express a specific concept. (...) these combinations are stored in the mental lexicon of native speakers and as a whole refer to a (linguistic) concept".

One of most detailed classifications for MWEs is that of [18]. It does not take into consideration the headword type of the MWEs like the approach of [9] and [7]; instead, it divides MWEs in two general types – lexicalised and institutionalized phrases. The first group is for phrases that have at least partially idiosyncratic syntax or semantics, or contain "words", which do not occur in isolation, and it has three subtypes:

1. Fixed expressions – fully lexicalized expressions that do not undergo morphosyntactic variation and internal modification (for example in short, ad hoc).

2. Semi-fixed expressions – these expressions undergo some degree of lexical variation and are further divided in three types:

- Non-decomposable Idioms – the only type of lexical variation observable in this group is inflection (kick the bucket) and reflexive form (wet oneself).
- Compound Nominals – these phrases inflect for number (car parks, parts of speech).
- Proper Names – the phrases in this group are syntactically highly idiosyncratic (San Francisco 49ers, Oakland Raiders), so they require different approach for analysis, depending on their instances.

3. Syntactically-flexible expressions – this subtype exhibits a much wider range of syntactic variability than the semi-fixed expressions and are divided by the types of variations possible:

- Verb-particle Constructions – these constructions consist of a verb and one or more particles (write up, look up).
- Decomposable Idioms – phrases of this subtype (for example let the cat out of the bag, sweep under the rug) are very challenging for analysis, because they are syntactically variable to varying degrees.
- Light Verbs – these constructions contain a noun used in a normal sense and a verb with bleached, rather than idiomatic meaning (make a mistake, give a demo).

The second type of MWEs in this classification is institutionalized phrases and it contains conventionalized phrases that are semantically and syntactically compositional, but statistically idiosyncratic (traffic light, fresh air).

The same approach to MWEs classification is shown in [1]. The authors also use the high-level classification, based particularly on the syntactic and semantic properties of MWEs and divide them into lexicalised phrases and institutionalised phrases.

Another approach on the differentiation of MWEs, that is not intended as a classification, but could give an interesting perspective on the subject is given in [5], where twelve groups are outlined:

- Proverbs (a bird in the hand is worth two in the bush, quotations (shaken, not stirred) and common-places (one never knows).
- Metaphorical Expressions (as sure as eggs is eggs).
- Verbal Idioms (to kick the bucket).

- Particle/Phrasal Verbs (to make up).
- Light Verb Constructions/Composite Predicates (to have a look).
- Syntactic/Quasi Noun Incorporation (German Auto waschen "to wash car").
- Stereotyped Comparisons/Similes (as nice as pie).
- Binomial Expressions (shoulder to shoulder).
- Complex Nominals (man about town).
- Collocations (strong tea).
- Fossilized/Frozen Forms (all of a sudden).
- Routine Formulas (Good morning).

Applying the above-mentioned classification of [18] to our dataset we can put most of the MWEs into the category of lexicalized phrases. There are no MWEs in the class of institutionalized phrases – they are very similar to compound nominals but because they are not lexicalized, they do not appear as heads of Wikipedia pages. Wikipedia pages contain NEs which act as fixed expressions for locations like Сан Франциско (*San Francisko*, "San Francisco") and Република Южна Африка (РЮА) (*Republika Uzhna Afrika* (*RUA*), "Republic of South Africa (RSA)") where the expressions are rigid and do not undergo modifications on the one hand. On the other hand, we can also find NEs for organizations acting like semi-fixed expressions where the fixed expressions can undergo morphosyntactic changes such as Европейска комисия (*Evropeiska komisia*, "European Commission") and Европейската комисия (*Evropeiskata komisia*, "the European Commission"). This category includes all of the compound nominals that constitute the terms from all kinds of different fields of science or everyday life. Considering the class of syntactically flexible expressions, we observed that in the dataset there are no verb-particle constructions, decomposable idioms and light verbs due to the nature of the Wikipedia dataset.

4.1. Compositional classification

As already mentioned, we will apply the MWE classification of [9]. All of the extracted MWEs are nouns and most of them are from the type Adjective + Noun; smaller part of them are Noun + Noun and Numeral + Noun. MWEs in these domains can be divided in two groups Named Entities (NEs) and terminology concepts. Our main concern are the terms. We also include NEs of global scope such as the event of WW2 (and large-scale operations as D-Day) or Summer Olympics as a sports forum (but not its iterations).

Baldwin and Kim [1] review the major types of MWEs in terms of composition. They outline three classes – nominal, verbal and prepositional MWEs. They consider compositionality to be closely related to the notion of idiomaticity, where the degree of the features of the parts of a MWE combine to predict the features of the whole. The authors also argue that although "compositionality is often construed as applying exclusively to semantic idiomatic (hence by "non-compositional MWE", researchers tend to mean a semantically idiomatic MWE), in practice it can apply across all the same levels as idiomaticity."

4.1.1. Nominal MWEs

Baldwin and Kim [1] state that the primary type of this category for English is the noun compound MWE (such as golf club, computer science department). This type of MWEs is rarer in Bulgarian than in English. The Wikipedia content is specific, so it is hard to find MWEs of this type in the data extracted from it, but still some examples can be found – блок схема (*blok shema*, "flowchart"), делта блус (*detla blus*, "Delta blues"). However, the same construction as in English could be seen in some loanwords in Bulgarian – голф клуб (*golf klub*, "golf club"). In this class, they also consider a subset – compound nominalisations – in which the head of the phrase is deverbal (for example investor hesitation, stress avoidance). Examples of this type MWEs are not found in our dataset. There is another type of nominal MWEs – nominal compounds, where the modifier is a verb (connecting flight) or an adjective (open secret). The MWEs that are nominal compounds with adjective modifier are very frequent in our extracted data – terms from all kinds of sciences are examples of this type MWEs (екваториален климат (*ekvatorialen klimat*, "equatorial climate"), атомна орбитала (*atomna orbitala*, "atomic orbital"). There are no examples of the type with a verb modifier in the dataset. For languages like the Romance there is one more type – complex nominals, which have a preposition or other marker between the nouns (such as succo di limone "lemon juice" and porta a vetri "glass door"). This type is observed in the extracted MWEs – въже за скачане (*vazhe za skachane*, "jump rope"), Договор от Лисабон (*Dogovor ot Lisabon*, "Treaty of Lisbon"), лишаване от свобода (*lishavane ot svoboda*, "imprisonment").

4.1.2. Verbal MWEs

In this class four subtypes are outlined:

- **Verb-particle constructions.** The verb-particle constructions or particle verbs or phrasal verbs consist of a verb and an intransitive preposition (for example play around), an adjective (cut short) or another verb (let go).
- **Prepositional verbs.** This subtype consists of a verb and a transitive preposition (refer to). They could have a fixed or mobile preposition (refer to the book and to the DVD vs. come across the book and across the DVD).
- **Light-Verb Constructions.** This type contains a verb and a noun complement (take a walk). The authors list the most frequent verbs that occur in light-verb constructions for English – do, give, have, make, take.
- **Verb-Noun Idiomatic Combinations.** This type MWEs are constructed of a verb and a noun in direct object position and they are semantically idiomatic (shoot the breeze). Here two subtypes are considered – decomposable and non-decomposable verb-noun idiomatic combinations. For the decomposable "it is possible to associate components of the VNIC with distinct elements of the VNIC interpretation, based on semantics not immediately accessible from the component lexemes" (Baldwin and Kim [1]) – for example in the phrase "spill the beans" spill could be interpreted as reveal and the beans – as secret. The decomposable verb-noun idiomatic combinations – for example "kick the bucket" – cannot be interpreted in such way.

4.1.3. Prepositional MWEs

This group has two subtypes: determinerless prepositional phrases which are formed by a preposition and a singular noun without a determiner (on top, by car) and complex prepositions (on top of, in addition to). In our Wikipedia dataset there are no verbal and prepositional MWEs, although these types are widespread in Bulgarian.

4.2. Idiomaticity of MWEs

B a l d w i n and K i m [1] also show the idiomaticity as one of the major properties of MWEs. As such property it can be used for classification of MWEs in our dataset. It refers to the similarity or deviation from the basic properties of the component lexemes. MWEs are often idiomatic at multiple levels applying to the lexical, syntactic, semantic, pragmatic, and/or statistical level. The idiomaticity of MWEs can be grouped in five subtypes:

- Lexical Idiomaticity – it occurs when one or more of the components of an MWE are not part of the lexicon of the given language. For example, "ad hoc".
- Syntactic Idiomaticity – it occurs when the syntax of the MWE is not derived directly from that of its components. For example, "by and large".
- Semantic Idiomaticity – is the property of the meaning of a MWE not being explicitly derivable from its parts. For example, "middle of the road".
- Pragmatic Idiomaticity – the condition of a MWE being associated with a fixed set of situations or a particular context. For example, "good morning".
- Statistical Idiomaticity – it occurs when a particular combination of words has high frequency, relative to the component words or alternative phrases of the same expression. For example, binomials such as "black and white television".

The most common type of idiomaticity in our dataset is the semantic idiomaticity. This is due to the fact that a lot of the Wikipedia pages are about NEs where: a) the meaning of the MWE cannot be derived from the meaning of its lexemes in examples such as Млечен път (*mlechen pat*, "Milky way"); b) the name follows some taxonomy rules unknown to the layman such as the names of cosmic objects like 6 Хеба (*6 Heba*, "6 Hebe"); or c) the name comes from foreign languages and the meaning is lost in translation or due to some cultural practices like the cases of Буенос Айрес (*Buenos Aires*, "Buenos Aires") – "good winds" and Адис Абеба (*Adis Abeba*, "Addis Ababa") – "new flower".

We can see lexical idiomaticity in Bulgarian Wikipedia MWEs due to terms becoming loanwords such as "quark" in странен кварк (*stranen kvark*, "strange quark") or the transliteration of names that are coming from other European languages, for example, "Fischer and Tropsch" in Процес на Фишер-Тропш (*Protses na Fisher-Tropsh*, "Fischer-Tropsch process").

The same reasons can be applied to syntactic, pragmatic and statistical idiomaticity where NEs and loanwords tend to behave rigidly in the Bulgarian language, but the rest of MWEs follow the language rules.

4.3. Domain classification

We would like to propose our approach using domain knowledge from Wikipedia for MWEs classification. This way we will enrich our notion of MWEs with

encyclopaedic information and we will be able to analyse them from other than linguistic perspective. Many domains are represented in Wikipedia, but here we outline the most prominent ones divided in six conditional groups: Physics and astronomy, Chemistry, Geography, Biology and medicine, Social, Other.

Table 1. Domain classified MWEs from Wikipedia

Domain	Number of MWEs
Chemistry	30
Physics and Astronomy	89
Biology and Medicine	130
Geography	801
Social	960
Other	177
Total	2187

4.3.1. Chemistry

This is the domain with the fewest amount of MWEs – only 30 and they are maybe the most homogeneous group. Most of the MWEs here are chemical compounds, which are traditionally built of Adj+N (for example солна киселина (*solna kiselina*, "hydrochloric acid"), but there are also other types of terms with the same structure such as атомен номер (*atomen nomer*, "atomic number") and полипептидна верига (*polipeptidna veriga*, "polypeptide chain"). We also have concepts like селитра (*selitra*, "saltpeter") and its sub-types (hyponyms): натриев нитрат (*natriev nitrat*, "sodium nitrate"); амониев нитрат (*amoniev nitrat*, "ammonium nitrate"); калиев нитрат (*kaliev nitrat*, "potassium nitrate"). The most complex in lexical structure MWEs in this domain are the terms that contain preposition and proper name like Принцип на Льо Шателие-Браун (*Printsip na Lyo Shatelie-Braun*, "Le Chatelier's principle") and Процес на Фишер-Тропш (*Protsesna Fisher-Tropsh*, "Fischer–Tropsch process").

4.3.2. Physics and astronomy

Wikipedia contains many MWEs (for example Fig. 1) which are NE to astral objects like asteroids, comets and planets of the type 3 Juno and 81P/Wild, that follow taxonomic patterns and are considered infinite. Such MWEs are interesting but of little importance to the expansion of the BTB-WN. They are typically constructed of noun and number: asteroids and spaceflight programs tend to contain the name of an Ancient Greek or roman gods or mythological characters 16 Психея (*16 Psiheya*, "16 Psyche"), 12 Виктория (*12 Viktoriya*, "12 Victoria"). Scientific laws, theories, principles and inventions very often include the name(s) of its inventor and thus they are constructions of noun, preposition and surname – Закон на Хенри (*Zakon na Henri*, "Henry's law"), Константа на Планк (*Konstanta na Plank*, "Planck constant").

Other MWEs that are much more valuable are the terms for physical phenomena and units of measure, because they do not contain numbers and proper nouns. Examples for this type of MWEs in the domain are: кинетична енергия (*kinetichna energiya*, "kinetic energy"), ядрена реакция (*yadrena reaktsiya*, "nuclear reaction"),

електрически заряд (*elektricheski zaryad*, “electric charge”). There were observed two types of measure units – Adj+N конска сила (*konska sila*, “horsepower”) and N+Prep+N метър в секунда (*metar v sekunda*, “metre per second”).

Because of the constant new findings and inventions of the modern science this domain is one of the most productive in Wikipedia, so it could be considered as a regular source for MWEs extraction.

4.3.3. Biology and medicine

The MWEs in these domains most frequently are constructed of two components. Here there are names of species of animals, plants and mushrooms: бяла мечка (*byala mechka*, “polar bear”); синя боровинка (*sinya borovinka*, “western blueberry”), дяволска гъба (*dyavolska gaba*, “Devil’s bolete”); of body organs and parts or diseases – гръбначен стълб (*grabnachen stalb*, “vertebral column”); нервна тъкан (*nervna takan*, “nervous tissue”), западнонилска треска (*zapadnonilska treska*, “West Nile fever”); branches or subfields of biology and medicine – популационна генетика (*populatsionna genetika*, “population genetics”), ветеринарна медицина (*veterinarna meditsina*, “veterinary medicine”); and other types of domain specific terms – клетъчна стена (*kletachna stena*, “cell wall”), двудомно растение (*dvudomno rastenie*, “dioecious plant”).

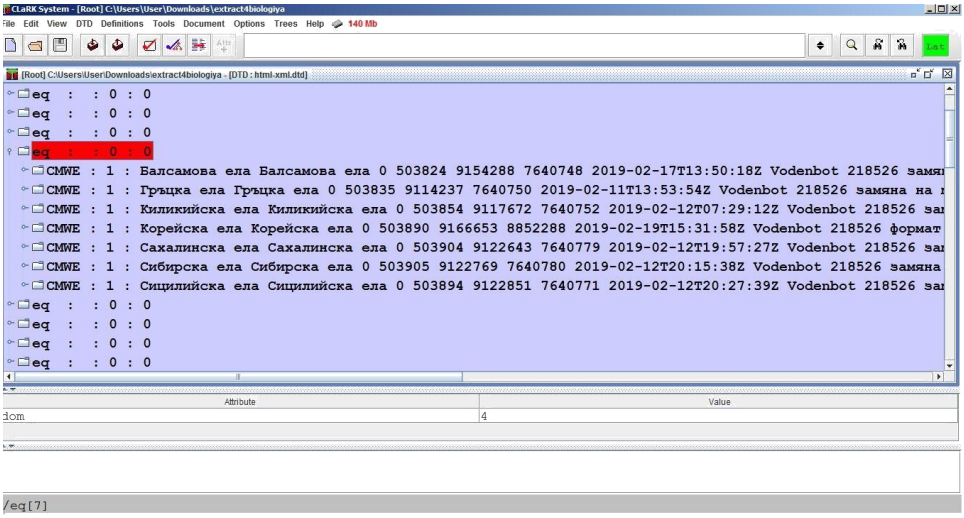


Fig. 2. The Wikipedia category “Firs” with articles for the different species in CLaRK system (**http://bultreebank.org/en/clark/**)

Some exceptions from the two component structure are: the terms for some disorders that include the name of their discoverer (as it is with the principles in the physics domain) such as Болест на Уилсън (*Bolest na Uilsan*, “Wilson’s disease”); subtypes of disease like рак на белия дроб (*rak na beliya drob*, “lung cancer”); terms like оцеляване на най-приспособения (*otselyavane na nai-prisposobeniya*, “survival of the fittest”).

4.3.4. Geography

This is the second largest domain and contains Wikipedia pages mostly for NEs, that may be put in two groups: locations (LOC) such as mountains, deserts, lowlands, bodies of water, islands, archipelagos, capes, hemispheres, etc., (for example Бяло море (*Byalo more*, "Aegean Sea")), and geopolitical locations (LOC-GPE) as countries, regions, departments, provinces, cities, kingdoms, counties, colonies (for example Ню Касъл (*Nyu Kasal*, "New Castle"); Обединеното кралство (*Obedinenoto kralstvo*, "The United Kingdom")).

There are several instances of peninsula with NEs – Скандинавски полуостров (*Skandinavski poluostrov*, "Scandinavian Peninsula"); Корейски полуостров (*Koreiski poluostrov*, "Korean Penin-sula"); Баха Калифорния (*Baha California*, "Baja California").

Lots of NEs that are settlements in Bulgaria will be added as instances of the synsets for village, town, city.

In this domain we also include climate zones and types тропичен саванен климат (*tropichen savanen klimat*, "tropical savanna climate") and natural phenomena and disasters such as storms, volcanic eruptions, etc., Ел Ниньо (*El Ninyo*, "El Nino"), Вранчанско земетресение (*Vranchansko zemetresenie*, "Vrancea earthquake").

The geography domain is also very rich in terms that are not named entities: дъждовна гора (*dazhdovna gora*, "rainforest"), морско равнище (*morsko ravnishte*, "sea level").

4.3.5. Social domain

This is the largest and most prominent domain that contains concepts related to society and humans, thus making it the most heterogeneous. Here these types of MWEs can be found: sport events and teams; wars, battles and crisis; pacts, contracts and unions; armies and legions; languages and linguistic terms; famous buildings; the parts of the Bible; holidays; art styles; institutions and organizations.

The longest MWEs in this domain are the different types of institutions and organizations (of course not all of them are so complex – Върховен съд (*Varhoven sad*, "Supreme court") such as Българска народна македоно-одринска революционна организация (*Bulgarska narodna makedono-odrinska revolyutsionna organizatsiya*, "Bulgarian people's Macedonian-Adrianople revolutionary organization"), Координационен комитет за контрол на износа (*Koordinatsionen komitet za kontrol na iznosa*, "Coordinating committee for multilateral export controls").

Some Wikipedia pages contain information about hyponyms of a concept like: президентска република (*prezidentska republika*, "presidential republic") and парламентарна република (*parlamentarna republika*, "parliamentary republic") as sub-types (hyponyms) of република (*republika*, "republic").

As observed in (S a g et al. [18]) sports team names usually contain a place or organization name (for example Бостън Селтикс (*Bostan Seltiks*, "Boston Celtics"). The case with sports competitions and different types of festivals is similar – Токио 2020 (*Tokio 2020*, "Tokyo 2020"); Филмов фестивал в Кан (*Filmov festival v Kan*,

"Cannes Film Festival"). Events of wars and battles are built of at least two lexical elements and usually denominate the place or time/duration of their occurrence – Първа световна война (*Parva svetovna voina*, "World War I"); Битка при Вердюн (*Bitka pri Verdyun*, "Battle of Verdun"). Some of these concepts are annual and similar to the productivity of NEs in the astronomy domain and are skipped.

There are many MWEs for organizations in different fields – I Германски легион (*Parvi germanski legion*, "1st Germanic Legion") and occupations – министър на отбраната (*ministar na otbranata*, "minister of defence"). Languages and language families are always MWEs валонски език (*valonski ezik*, "Walloon language"), памирски езици (*pamirski ezitsi*, "Pamir languages") Many holidays and currencies appear in this group too – Рождество Христово (*Rozhdestvo Hristovo*, "Feast of the Nativity"), щатски долар (*shtatski dolar*, "United States dollar").

4.3.6. Miscellaneous domain

This group contains heterogeneous MWEs, that cannot be placed in the previous categories and are too little to be in separate groups. Most of them could be generalized as artefacts – there are products, inventions, man-made entities.

Nautical and aviation. A quite big part of this group consists of nautical and aviation terminology – types of ships, ship elements, military aircraft are very well presented in the Bulgarian Wikipedia. Here the MWEs always have two components – adjective and noun – like бронирана палуба (*bronirana paluba*, ("armoured deck"), бойна рубка (*boina rubka*, ("conning tower"), атомна подводница (*atomna podvodnitsa*, ("nuclear submarine"). There are exceptions like the names of fighter aircraft and bombers, that usually contain a proper name and numbers (Messerschmitt Bf 109, Avia B-135, Albatros C.III).

Weapons and ammunition. Another big group is formed by types of weapons and ammunition, which are also frequently constructed of Adj+N in Bulgarian (in English they usually are compound nouns) for example огнестрелно оръжие (*ognestrelno orazhie*, "firearm"), but could be more complex in some cases – ръчен противотанков гранатомет (*rachen protivotaknov granatomet*, "rocket-propelled grenade"), междуконтинентална балистична ракета (*mezhdukontinentalna balistichna raketa*, "in-tercontinental ballistic missile").

Machines and vehicles. The tendencies in the MWEs for machines and their components, vehicles and musical instruments are quite the same like the before mentioned groups – асинхронен двигател (*asinhronen dvigatel*, "asynchronous motor"), бронирана кола (*bronirana kola*, "armoured car"), цилиндричен барабан (*tsilindrichen baraban*, "cylindrical drum"). Rarely constructions with preposition could be observed – автомобил с повишена проходимост (*avtomobil s povishena prohodimost*, "sport utility vehicle"), but we can see more everyday artefacts like: вятърна мелница (*viaturna melnica*, ("wind mill"); спален чувал (*spalen chuval*, ("sleeping bag"); автомобилна гума (*avtomobilna guma*, ("automobile tyre").

Musical instruments. Wikipedia contains a lot of pages for music, so in this subdomain we observe MWEs that are musical instruments like явански гонг

(*yavanski gong*, (“gong ageng”) and categories of instruments such as ударен инструмент (*udaren instrument*, (“percussion instrument”).

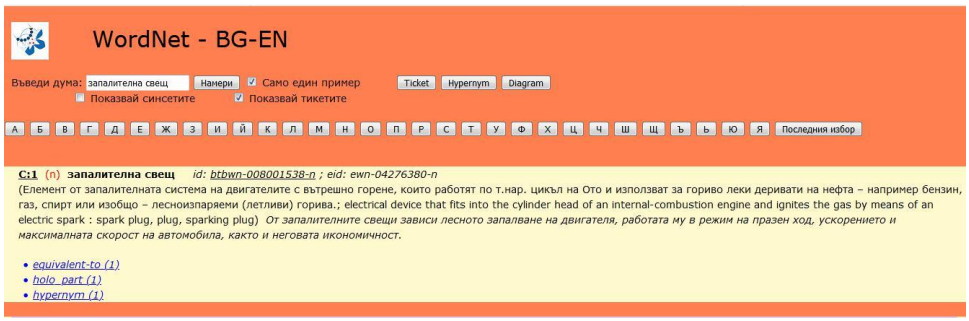


Fig. 3. Example of a MWE synset in BTB-WN

Food and drinks. Different types of food and drinks are included in the Other domain and they do not diverge in type and number of lexical elements from the already mentioned MWEs in this domain. Typical examples are стафидено вино (*stafideno vino*, “straw wine”), стар ейл (*star eil*, “old ale”), бяло саламурено сирене (*byalo salamureno sirene*, “white brine cheese”).

Mathematics and IT. Another group is related to mathematics and IT with MWEs like аксиоматичен метод (*aksiomatichen metod*, “axiomatic method”); закон за големите числа (*zakon za golemite chisla*, “law of large numbers”); уеб дизайн (*uoeb dizain*, “web design”); син екран на смъртта (*sin ekran na smurtta*, “blue screen of death”).

5. Automatic domain classification of MWEs

After the manual determination of MWEs, we have automatically divided them by their category in Wikipedia, which helps with the domain typology to a certain extent, but is definitely tricky. The categories in Wikipedia are thousands and tend to specify, rather than to generalize the topics of the content, so they would form a very detailed and hard to apply classification of MWEs. There was the idea of using categories from Wikipedia in English and Bulgarian for automated clustering of MWEs. Those categories proven to be too unstructured. As the examples bellow show: If we compare two MWEs from the same domain like Milky way and Solar wind in English Wikipedia we see that the first has three categories that follow different paths: 1) Milky Way – has three more categories; 2) Astronomical objects known since antiquity – has two more categories; 3) Barred spiral galaxies – has two more categories, but Solar wind two: 1) Solar phenomena – has two more categories and 2) Space plasmas – has three more categories.

The common categories between the two MWEs are: Astronomical objects known since antiquity – two steps from Solar wind and one from Milky way; Space science – two steps or four from Solar wind, depending on the path taken, and four from Milky way; Astronomical sub-disciplines – four steps from Solar wind and three from Milky way.

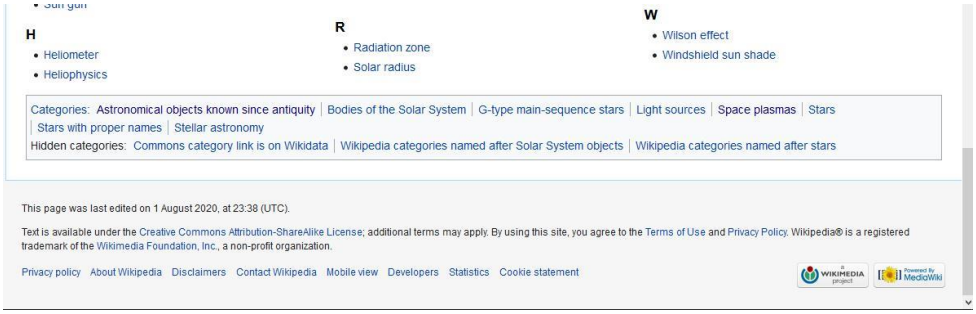


Fig. 4. Example of Wikipedia categories for the page Sun – two steps from Solar wind

In Bulgarian Wikipedia there are different categories and paths. On the one hand, as said before, the Bulgarian Wikipedia is 23 times smaller than the English Wikipedia, which leads to less categories per page. On the other hand, the existing categories take on different paths. For example: Solar wind is under Sun, which is under Solar system, which is under Milky way.

Most of the Wikipedia pages are marked with two and more categories and then some have additional Hidden categories. The end result is granularity and overlapping.

6. Conclusion

The integration of lexical resources like WordNet with encyclopaedic knowledge derived from Wikipedia proves to be very useful in the NLP field. It is even more beneficial for relatively small sized resources such as BTB-WN, which has been already expanded with 15% general concepts from Wikipedia and it is going through a MWEs specialized expansion with at least another 15%.

The manual selection and classification of MWEs is highly time consuming, but our experience shows that it is the better option in the case of extraction from Bulgarian Wikipedia – if the attempt is to do this kind of domain classification automatically using only the data from Wikipedia, its categories and hierarchy – it will not be beneficial enough.

Part of our plans for future work is to produce automatically synsets for NEs, which are related to Bulgaria. All of the PERs and LOC-GPEs in Bulgarian Wikipedia would be added to the synset of the corresponding occupation or settlement (village, town, and city) with the relation instance-of.

In regard to the domain distribution of the extracted MWEs it could be summarized that the fields of social sciences, sport and art and the geography domain are the most numerous.

Maybe the best approach to enrich a wordnet with MWE synsets would be to combine methods with accumulation of knowledge – linguistic and encyclopaedic. MWEs are complex and have several layers of properties or features. We are hoping that adding encyclopaedic knowledge for domains and context on top of linguistic knowledge for morphology, syntax and semantics is the way forward.

***Acknowledgements*:** This work was partially supported by the Bulgarian Ministry of Education and Science under the Na-tional Research Programme "Young scientists and postdoctoral students" approved by DCM No 577/17.08.2018 and by 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-CLaDA-BG, Grant number DO01-272/16.12.2019.

References

1. Baldwin, T., S. N. Kim. Multiword Expressions. – In: Nitin Indurkhya, Fred J. Damerau, Eds. Handbook of Natural Language Processing. No 2. Boca Raton, USA, Chapman and Hall/CRC, 2020, pp. 267-292.
2. De Lacalle, M. L., E. Laparra, G. Rigau. Predicate Matrix: Extending Semlink through Wordnet Mappings. – In: Proc. of 9th International Conference on Language Resources and Evaluation (LREC'14), European Language Resources Association (ELRA), Reykjavik, Iceland, 2014, pp. 903-909.
3. Gurevych, I., J. Eckle-Kohler, S. Hartmann, M. Matuschek, C. M. Meyer, C. Wirth. UBY – A Large-Scale Unified Lexical-Semantic Resource Based on LMF. – In: Proc. of 13th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Avignon, France, 2012, pp. 580-590.
4. Han, L., G. J. Jones, A. F. Smeaton. MultiMWE: Building a Multi-Lingual Multi-Word Expression (MWE) Parallel Corpora. – In: Proc. of 12th Conference on Language Resources and Evaluation (LREC'20), European Language Resources Association, Marseille, France, 2020, pp. 2970-2979.
5. Huning, M. B. Schlucker. Multi-Word Expressions. Vol. **1**. 2015, pp. 450-467.
6. Kecskes, I. Intercultural Pragmatics. New York, USA, Oxford University Press, 2013, pp. 81-104.
7. Koeva, S., I. Stoyanova, M. Todorova, S. Leseva. Semi-Automatic Compilation of the Dictionary of Bulgarian Multiword Expressions. – In: Proc. of GLOBALEX 2016: Lexicographic Resources for Human Language Technology, Workshop at LREC2016, Portoroz, Slovenia, 2016.
8. Kurfalı, M., R. Ostling, J. Sjons, M. Wiren. A Multi-Word Expression Dataset for Swedish. – In: Proc. of 12th Conference on Language Resources and Evaluation (LREC'20), LREC 2020, Marseille, France, 2020, pp. 4402-4409.
9. Laskova, L., P. Osenova, K. Simov, I. Radev, Z. Kancheva. Modeling MWEs in BTB-WN. – In: Proc. of Joint Workshop on Multiword Expressions and WordNet (MWE-WN'19), Association for Computational Linguistics, Florence, Italy, 2019, pp. 70-78.
10. Masini, F. Multi-Word Expressions between Syntax and the Lexicon: The Case of Italian Verb-Particle Constructions. – SKY Journal of Linguistics, Vol. **18**, 2005, pp. 145-173.
11. McCrae, J. P. Mapping WordNet Instances to Wikipedia. – In: Proc. of 9th Global WordNet Conference, the Global WordNet Association, Singapore, 2018, pp. 62-69.
12. Navigli, R., S. P. Ponzetto. BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. – Journal of Artificial Intelligence, 2012, pp. 217-250.
13. Osenova, P., K. Simov. The Datadriven Bulgarian WordNet: BTBWN. – In: Cognitive Studies. Etudes Cognitives. Vol. **18**. 2018.
14. Osenova, P., K. Simov, L. Laskova, S. Kancheva. A Treebank-Driven Creation of an Ontovalence Verb Lexicon for Bulgarian. – In: Proc. of 8th International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 2636-2640.
15. Palmer, M. Semlink: Linking PropBank, VerbNet and FrameNet. – In: Proc. of Generative Lexicon Conference, 2009, pp. 9-15.
16. Rohanian, O., M. Rei, S. Taslimipoor, L. A. Ha. Verbal Multiword Expressions for Identification of Metaphor. – In: Proc. of 58th Annual Meeting of the Association for Computational Linguistics, 2020, pp. 2890-2895.

17. Rudnicka, E. K., M. T. Piasecki, T. Piotrowski, Ł. Grabowski, F. Bond. Mapping WordNets from the Perspective of Inter-Lingual Equivalence. – In: Cognitive Studies. Etudes Cognitives. Vol. **17**. 2017, pp. 1-17.
18. Sag, I., T. Baldwin, F. Bond, A. Copestake, D. Flickinger. Multiword Expressions: A Pain in the Neck for Nlp. – In: Proc. of 3rd International Conference on Computational Linguistics and Intelligent Text Processing, 2002, pp. 1-15.
19. Shi, T., L. Lee. Extracting Headless MWEs from Dependency Parse Trees: Parsing, Tagging, and Joint Modeling Approaches, 2020.
20. Simov, K., A. Simov, H. Ganev, K. Ivanova, I. Grigorov. The CLaRK System: XML-Based Corpora Development System for Rapid Prototyping. – In: Proc. of LREC'04, 2004, pp. 235-238.
21. Simov, K., P. Osenova, L. Laskova, I. Radev, Z. Kancheva. Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia. – In: Proc. of 10th Global WordNet Conference, 2019, pp. 290-297.
22. Sprenger, S. Fixed Expressions and the Production of Idioms. Ponsen and Looijen BV, Wageningen, 2003.

Received: 02.11.2020; Second Version: 04.11.2020; Accepted: 09.11.2020 (fast track)