Linguistic vs. encyclopedic knowledge. Classification of MWEs on the base of domain information

Zara Kancheva
IICT-BAS
zara@bultreebank.org

Ivaylo Radev
IICT-BAS
radev@bultreebank.org

Abstract

This paper reports on the first steps in the creation of linked data through the mapping of BTB-WordNet and the Bulgarian Wikipedia. The task of expanding the BTB-WordNet with encyclopedic knowledge is done by mapping its synsets to Wikipedia pages 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.

1. Introduction

The state of the field shows that language resources used alone do not perform well in each and every NLP task. In recent years researchers started to align various lexical resources in projects such as BabelNet (Navigli and Ponzetto, 2012), SemLink (Palmer, 2009), Predicate Matrix (de Lacalle et al., 2014) and Uby (Gurevych et al., 2012). Building relations between linguistic and semantic resources and using this kind of new data to generate knowledge graphs has its benefits for languages with less lexical resources.

With the development of huge electronic corpora and advancements in corpus linguistics MultiWord Expressions (MWEs) receive more and more attention from researchers. A paper (Sag et al., 2002) estimates that the number of MWEs in the lexicon of a person is more than 40%. MWEs are omnipresent in all text data and can not be skipped in tasks such as word sense disambiguation, named entity linking and coreference resolution.

Some tendencies in contemporary linguistics have changed drastically and the observation of (Kiefer, 1988) that "theoretical linguists and lexicography seem each to go their own ways, they do not seem to show much interest in other's preoccupations" are no longer factual. There are several projects that aim to integrate linguistic and encyclopedic knowledge, most commonly by the merge of a dictionary or WordNet with Wikipedia or Wiktionary.

One of the main challenges occurring from the integration of the two types of knowledge - of the linguistic system and of the world (Kecskes, 2013) - is related with the question how much and what types of encyclopedic information is useful to add to our language resource? A lot of work is already done on the mapping of dictionaries and WordNets with Wikis, but it is interesting and challenging to focus on the MWE distribution in the resulting dataset.

In the process of manual mapping with CLaRK system (Simov et al., 2004) between Bulgarian WordNet (BTB-WN) and Wikipedia 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. Section 3 explores different domains of encyclopedic knowledge in Wikipedia. Section 4 concludes the paper.

2. Related Work

Among the most outstanding works on the alignment of linguistic and encyclopedic 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 (McCrae, 2018); and the mapping of the plWordNet onto the Princeton WordNet (Rudnicka et al., 2017).

For the purposes of this research we work with the BTB-WN (Osenova and Simov, 2018), which was build in several steps. It started as an translation of Core WordNet and was expanded with concepts from Bulgarian Treebank (BulTreeBank (Osenova et al., 2012)), 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 (Simov et al., 2019).

Currently the Wikipedia in Bulgarian has 259 927 content pages, 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 (Laskova et al., 2019) presents an overview of MWEs in BTB-WN, where the MWEs are presented as several types of phrases by their head-word: multiword Nouns (Noun+Noun (N); Adj+N; Numeral+N); Verbs (Verb+N; Verb+Adv; Verb+PP); Adjectives (Adv+Adj; Adj+PP) and Adverbials (Prep+N; Prep+Adj; Adv+Adv) in accordance with the classification developed within WG 4 of PARSEME COST Action[1]; and treated afterward with a catena-based modeling. Working with the same resource we will use the same classification method in our work.

A similar approach in dealing with MWEs is presented in (Koeva et al., 2016). 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 86373 'nominal' and 'verbal' MWEs, based on the head word.

MWEs could be defined as "lexical units larger than a word that can bear both idiomatic and compositional meanings" (Masini, 2005). (Sprenger, 2003) 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".

There is no single generally accepted typology of the MWE, different researchers classify them at several levels - morphology, lexicology, syntax and semantics. One of most detailed classifications is that of (Sag et al., 2002). It does not take into consideration the head-word type of the MWEs like the approach of (Laskova et al., 2019), (Koeva et al., 2016), 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:

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

[1] https://typo.uni-konstanz.de/parseme/

  - 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.

- 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*).

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 (Hüning and Schlücker, 2015), where twelve groups are outlined:

- Proverbs (*a bird in the hand is worth two in the bush*, quotations (*shaken, not stirred*) and commonplaces (*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*).

3. Domain Specific MWEs

For the aims of this research we have semi-automatically extracted 13173 MWEs from 14512 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 14512 Wikipedia pages were selected on the basis of mappings between Bulgarian Wikipedia and BTB-WN (Laskova et al., 2019). 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. An effort to use Wikipedia categories to match the WMEs to the respective domains was made, but they were too noisy and unstructured. 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.

Domain	**No MWEs**
Chemistry	30
Physics and Astronomy	89
Biology and Medicine	130
Geography	801
Social	960
Other	177
Total	2187

Table 1: Classified MWEs from Wikipedia

As already mentioned, we will apply the MWE classification of (Laskova et al., 2019). All of the extracted MWEs are nouns and most of them are type of Adj+N; smaller part of them are NN and Numeral+N. MWEs in these domains can be divided in two groups Named Entities (NEs) and terms / 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).

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 encyclopedic 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.

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 generalise the topics of the content, so they would form a very detailed and hard to apply classification of MWEs. Additionally, every article could and very often does belong to more than one category (for example the article for *Ammonium nitrate* appears in three categories - *Ammonium compounds, Nitrates, Explosive chemicals* and does not directly point to the more general category *Chemistry*. 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.

3.1. Physics and Astronomy Domain

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 (6 Хеба (*6 Heba*, “6 Hebe”). Scientific laws, theories, principles very often include the name(s) of its inventor and thus they are constructions of noun, preposition and surname - Уравнения на Максуел (*Uravneniya na Maksuel*, “Maxwell’s equations”), Принцип на Паули (*Printsip na Pauli*, “Pauli exclusion principle”).

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 do-

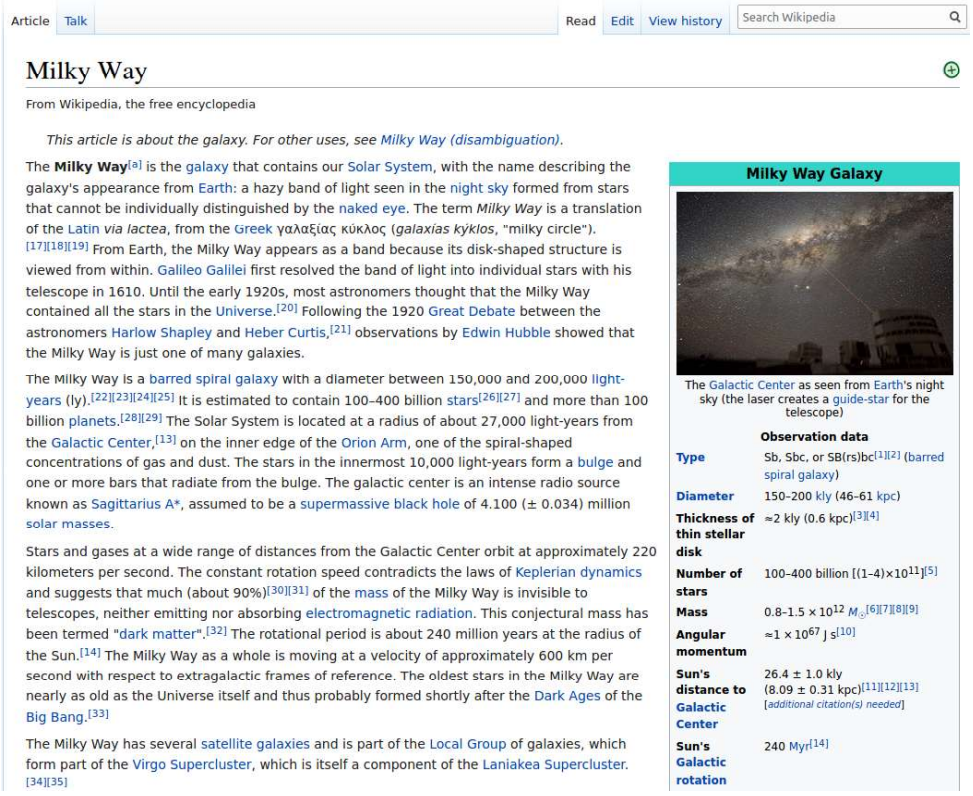

Figure 1: Wikipedia page with astronomy MWE

main are: слънчев вятър (*slanchev vyatar*, "solar wind"), магнитно поле (*magnitno pole*, "magnetic field"), горен/долен/странен кварк (*goren/dolen/stranen kvark*, "up/down/strange quark"). There were observed two types of measure units - Adj+N (конска сила (*konska sila*, "horsepower") and N+Prep+N (километър в час (*kilometar v chas*, "kilometre per hour").

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.

3.2. Chemistry Domain

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 глюкуронова киселина (*glyukuronova kiselina*, "glucuronic acid"), but there are also other types of terms with the same structure such as периодична система (*periodichna sistema*, "periodic table") and ковалентна връзка (*kovalentna vrazka*, "covalent bond"). 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 (*Protses na Fisher-Tropsh*, "Fischer–Tropsch process").

3.3. Geography Domain

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 южен полюс (*yuzhen polyus*, "South pole") and geopolitical locations (LOC-GPE) as countries, regions, departments, provinces, cities, kingdoms, counties, colonies

(for example Лос Анджелис (*Los Angelis*, “Los Angeles”); 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 Peninsula”); 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 (умерен климат (*umeren klimat*, “temperate climate”) and natural phenomena and disasters such as storms, volcanic eruptions, etc. (Ел Ниньо (*El Ninyo*, “El Niño”), Вранчанско земетресение (*Vrachansko zemetresenie*, “Vrancea earthquake”).

The geography domain is also very rich in terms that are not named entities: тектонска плоча (*tektonska plocha*, “tectonic plate”), морско равнище (*morsko ravnishte*, “sea level”).

3.4. Biology and Medicine Domain

The MWEs in these domains most frequently are constructed of two components. Here there are names of species of animals, plants and mushrooms: червена лисица (*chervena lisica*, “Red Fox”); бял бряст (*bial briast*, “european white elm”), бяла мухоморка (*byala muhomorka*, “destroying angel”); of body organs or diseases - костен мозък (*kosten mozak*, “bone marrow”); бели кръвни тела (*beli kravni tela*, “white blood cells”), метаболитен синдром (*metaboliten sindrom*, “metabolic syndrome”); branches or subfields of biology and medicine - молекулярна генетика (*molekulyarna genetika*, “molecular genetics”), ветеринарна медицина (*veterinarna meditsina*, “veterinary medicine”); and other types of domain specific terms - застрашен вид (*zastrashen vid*, “endangered species”), вечнозелено растение (*vechnozeleno rastenie*, “evergreen plant”).

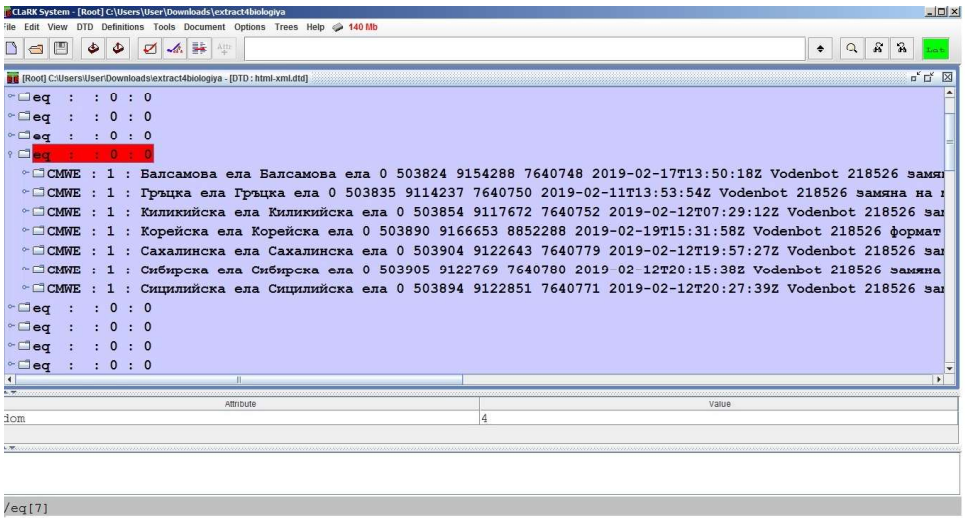

Figure 2: The Wikipedia category “Firs” with articles for the different species in CLaRK system[2]

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 Синдром на Турет (*Sindrom na Turet*, “Tourette syndrome”); subtypes of disease like рак на дебелото черво (*rak na debeloto chervo*, “Colorectal cancer”); terms like оцеляване на най-приспособения (*otselyavane na nai-prisposobeniya*, “survival of the fittest”).

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;

[2]http://bultreebank.org/en/clark/

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 (Sag et al., 2002) 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 occurance - 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 (бретонски език (*bretonski ezik*, "Breton language"), тюркски езици (*tyurkski ezitsi*, "Turkic languages") Many holidays and currencies appear in this group too - Рождество Христово (*Rozhdestvo Hristovo*, "Feast of the Nativity"), суринамски долар (*surinamski dolar*, "Surinamese dollar").

3.6. "Other" Domain

This group contains heterogeneous MWEs, that can not 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. 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 батарейна палуба (*batareina palouba*, ("gun deck"), бойна рубка (*boina rubka* , ("conning tower"), минен трал (*minen tral*, ("mine roller"). 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).

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 гладкоцевно оръжие *gladkotsevno orazhie*, "smoothbore"), but could be more complex in some cases - ръчен противотанков гранатомет (*rachen protivotaknov granatomet*, "rocket-propelled grenade"), междуконтинентална балистична ракета (*mezhdukontinentalna balistichna raketa*, "intercontinental ballistic missile"). The tendencies in the MWEs for vehicles, machines and their components, musical instruments are quite the same like the before mentioned groups - асинхронен двигател (*asinhronen dvigatel*, "asynchronous motor"), бронирана кола (*bronirana kola*, "armoured car"), бас китара (*bas kitara*, "bass guitar"). 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 choval*, ("sleeping bag"); автомобилна гума (*avtomobilna guma*, ("automobile tyre")

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 червено вино (*cherveno vino*, "red wine"), пейл ейл (*peil eil*, "pale ale"), бяло саламурено сирене

(*byalo salamureno sirene*, "white brine cheese").

Another group is related to mathematics and IT: аксиоматичен метод (*aksiomatichen metod*, "axiomatic system"); 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").

4. Conclusion

Aligning lexical resources like WordNet with encyclopedic knowledge from Wikipedia has proven very beneficial in the NLP field. This is even more true about relatively small sized resource that is BTB-WN. It has already been expanded once by 15% with general concepts from Wikipedia and now we are working on MWEs specialized expansion with another 15%, but this may be an underestimate.

Although doing this kind of work manually is very time consuming our experience shows that in the case of Bulgarian Wikipedia attempting to do this kind of domain classification automatically using only the data (in the form of its categories and hierarchy) from Wikipedia is not beneficial enough.

It is possible to use automatic methods in the future to produce synsets for NEs related to Bulgaria. For example all of the PERs and LOC-GPEs in Bulgarian Wikipedia can be added with the instance-of relation to the respective type of occupation/profession and settlement (village, town and city).

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.

Acknowledgements

This work was partially supported by the *Bulgarian Ministry of Education and Science under the National Research Programme "Young scientists and postdoctoral students" approved by DCM # 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

de Lacalle, M. L., Laparra, E., and Rigau, G. (2014). Predicate matrix: extending semlink through wordnet mappings. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 903–909, Reykjavik, Iceland, May. European Language Resources Association (ELRA).

Gurevych, I., Eckle-Kohler, J., Hartmann, S., Matuschek, M., Meyer, C. M., and Wirth, C. (2012). Uby - a large-scale unified lexical-semantic resource based on lmf. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 580–590, Avignon, France, April. Association for Computational Linguistics.

Hüning, M. and Schlücker, B. (2015). *Multi-word expressions*, volume 1, pages 450–467. De Gruyter Mouton, January.

Kecskes, I. (2013). *Encyclopedic Knowledge, Cultural Models, and Interculturality*, pages 81–104. OUP USA, December.

Kiefer, F. (1988). Linguistic, conceptual and encyclopedic knowledge: Some implications for lexicography. In Magay, T. and Zigány, J., Eds., *Proceedings of the 3rd EURALEX International Congress*, pages 1–10, Budapest, Hungary, September. Akadémiai Kiadó.

Koeva, S., Stoyanova, I., Todorova, M., and Leseva, S. (2016). Semi-automatic compilation of the dictionary of bulgarian multiwordexpressions. *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology, Workshop at LREC2016, Portorož, Slovenia*, May.

Laskova, L., Osenova, P., Simov, K., Radev, I., and Kancheva, Z. (2019). Modeling MWEs in BTB-WN. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 70–78, Florence, Italy, August. Association for Computational Linguistics.

Masini, F. (2005). Multi-word expressions between syntax and the lexicon: The case of italian verb-particle constructions. *SKY Journal of Linguistics*, 18:145–173, January.

McCrae, J. P. (2018). Mapping WordNet Instances to Wikipedia. In *Proceedings of Ninth Global WordNet Conference*, pages 62–69. The Global WordNet Association.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and applicationof a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Osenova, P. and Simov, K. (2018). The datadriven Bulgarian WordNet: BTBWN. *Cognitive Studies — Études cognitives*, 18.

Osenova, P., Simov, K., Laskova, L., and Kancheva, S. (2012). A treebank-driven creation of an ontovalence verb lexicon for Bulgarian. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 2636–2640, Istanbul, Turkey. LREC 2012.

Palmer, M. (2009). Semlink: Linking PropBank, VerbNet and FrameNet. In *Proceedings of the Generative Lexicon Conference*, page 9–15.

Rudnicka, E. K., Piasecki, M. T., Piotrowski, T., Łukasz Grabowski, and Bond, F. (2017). Mapping WordNets from the perspective of inter-lingual equivalence. *Cognitive Studies — Études cognitives*, 17(1373):1–17.

Sag, I., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. (2002). Multiword expressions: A pain in the neck for nlp. In *Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 1–15. Springer-Verlag, February.

Simov, K., Simov, A., Ganev, H., Ivanova, K., and Grigorov, I. (2004). The CLaRK System: XML-based Corpora Development System for Rapid Prototyping. *Proceedings of LREC 2004*, pages 235–238, May.

Simov, K., Osenova, P., Laskova, L., Radev, I., and Kancheva, Z. (2019). Aligning the Bulgarian BTB WordNet with the Bulgarian Wikipedia. In *Proceedings of the 10th Global WordNet Conference*, pages 290–297. Oficyna Wydawnicza Politechniki Wrocławskiej.

Sprenger, S. (2003). *Fixed expressions and the production of idioms*. Ponsen and Looijen BV, Wageningen.