Treating Multiword Expressions with a view to Morphologically Rich Languages

Svetlozara Leseva
Department of Computational Linguistics
Institute for Bulgarian Language – Bulgarian Academy of Sciences
zarka@dcl.bas.bg

Relevant UniDive working groups: WG1, WG2

1 Introduction

This abstract presents an effort towards a uniform description focusing on the linguistic representation of the structural, morphosyntactic, morphological, word-order and other features of Bulgarian MWEs with a view to their automatic recognition and annotation in running text. The proposal can be extended to the development of lexical resources of MWEs for other languages.

The importance of MWEs has been widely acknowledged by linguistics and computational linguistics alike (Sag et al., 2002; Baldwin et al., 2003), including the need for language processing systems providing access to resources in which information about MWEs is explicitly marked (Savary et al., 2019). This has resulted in international efforts such as PARSEME[1] (Ramisch et al., 2018), where special emphasis was placed on the properties of MWEs and on accounting for their variation across languages both in terms of semantic (non-)compositionality and morphosyntactic form and behaviour.

We aim at addressing the challenge of the description of MWEs in morphologically rich languages such as Bulgarian and other Slavic languages (Savary, 2008; Koeva, 2007) through implementing a set of comprehensive inflection types reflecting MWEs' internal structure, the morphosyntactic variability of their components, word order variations, intervening words and phrases, modification of MWE components, the possibility to leave out one or more of the MWE elements, etc. Such description facilitates the study of the interdependence between semantic non-compositionality and morphosyntactic fixedness, and can boost text analysis. The analysis is based on previous work on Bulgarian involving the semi-automatic compilation of a large MWE lexicon (Stoyanova et al., 2016).

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

2 A lexicon of Bulgarian MWEs

The lexicon includes a range of MWEs – over 10,000 nominal MWEs (including 5,000 NEs), 6,500 verbal MWEs (1,200 light verb constructions, 1,800 verbal idioms, 3,400 reflexives and others), and a small number of adverbial expressions. This work aims to establish a uniform framework for the description of their features.

2.1 Linguistic description of the MWEs

The linguistic description includes several layers of information: (i) lexicogrammatical characteristics of the head and the components (POS and lexical features); (ii) morphosyntactic characteristics – inflectional characteristics, morphosyntactic constraints imposed by the idiomatic meaning: e.g. *informatsionni tehnologii* 'information technologies' is always plural; in *imam zlatno sartse* 'have a heart of gold' the direct object must be indefinite and agrees in number with the subject); (iii) structural characteristics – the internal syntactic structure of the MWE and the number and type (head, dependents) of its constituents, as well as possible variations in their linear order; (iv) syntagmatic properties -– syntactic transformations and constraints, such as passivisation; optional components, such as modifier insertion/omission, e.g., *vzemam* ***vazhno*** *reshenie* 'make an **important** decision', possible derivations and other transformations; (v) the subcategorisation frame – subcategorised arguments, argument transformations and selectional restrictions (part of the description in (iv) and (v) is currently implemented).

2.2 Inflection types

We adopt a modular approach to the definition of inflection types. The inflection type of each MWE is defined as a sequence of elementary fields describing the properties of the individual MWE components. This allows us to model them independently by encoding information about: the components' syntactic categories (e.g. NP, PP), the MWE internal syntactic structure and the grammatical relations between the components couched in

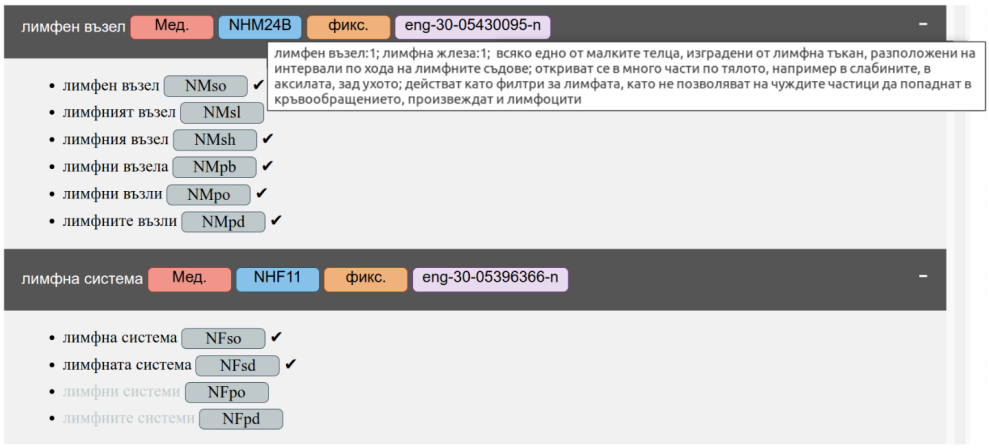

Figure 1: An example of the nominal MWEs *limfen vazel* 'lymph node' and *limfna sistema* 'lymphatic system'

BG: *удрям джакпота* – EN: *hit the jackpot 'succeed by luck'*	
Synset ID / MWE ID	eng-30-02524739-v / bg_2291
MWE lemma / Abstract lemma	*удрям джакпота / удрям джакпот*
Morphosyntactic features	*удрям.V_IMPERF_r1s джакпота.Nsh*
Head and head inflection type	*удрям*.V_IM_TT_S3_01
Head restrictions	none
Dependent and dep restrictions	*джакпота* / fixed; **N**(umber) = s; **D**(efiniteness) = h
Syntactic structure	Constituent: V N(P) | UD: V + obj
Semantic frame	Success_or_failure (Agent, Goal | Role)
Subcategorisation: subject	N(P)_subj | UD: nsubj
Subcategorisation: complements	Goal: PP | UD: obl & P = в/във; Role: PP | UD: obl & P = като
Possible modifiers of the head	regular
Possible modifiers of dependent	regular; A(P); Ex.: *удрям* ***големия/Ash*** *джакпот/Ns0*
External elements	regular (question particle | subj | AdvP...)
PARSEME type	VID
Register and connotation	Colloquial; −0.125 +0.25
Derivational relations	*удряне на джакпота*

Figure 2: The VMWE *udryam dzhakpota* 'hit the jackpot' (the fields in grey are currently being implemented)

the UD framework (de Marneffe et al., 2021); the (in))variable morphological categories, etc.

Each elementary inflection type can be divided into subtypes on the basis of the morphophonemic changes that take place in the paradigm, e.g. *gol**ya**m*-M.SG – *gol**e**mi*-PL ('big'). The elementary type defines the number and type of the forms generated. It is further extended with additional inflection components depending on the basic structural type of the MWE: (i) possible word-order changes; (ii) possible modifiers of certain components; (iii) obligatory non-lexicalised components within the MWE; (iv) discontinuous components and admissible external elements that can occur between the MWE components.

The nominal MWEs currently included in the lexicon are described in terms of 6 main structural types (e.g., A N — *byala mechka* 'polar bear'; N PP — *More na spokoystvieto* 'Sea of Tranquility', etc.), 31 inflection types and a total of 81 subtypes, which are being further enriched. Verbal MWEs are typically modelled as complex inflection types based on their structural types (e.g., V NP -– *obrashtam nova stranitsa* 'turn a new page'; V PP -– *prashtam za zelen hayver* 'send on a wild-goose chase'). We have described 11 structural types, 91 inflection types with over 1,000 subtypes.

2.3 Visualisation

The lexicon is available as a computational and as a human-readable resource viewable online[2].

The following information is displayed for each nominal MWE (Fig. 1). The red box contains the domain (Medicine, Technical, Chemistry, etc.) to which the MWE pertains. NEs are marked in a green box. The blue box contains the basic part of the inflection type which reflects the paradigm

[2]Nominal MWEs: https://dcl.bas.bg/mwe-dictionary-data/; verbal MWEs: https://dcl.bas.bg/derivation-vmwe/.

of the MWE. The orange box shows information about the word order of the MWE components ('fixed', i.e. invariable for the respective MWEs). Links to other resources where the item was found – in particular BulNet / WordNet (on hovering over the ID, information about the synset is displayed) and Wikipedia (with information about the Wikipedia article) – are shown in pale pink.

The forms of the MWE corresponding to the inflection type are visualised along with the relevant grammatical characteristics. A checkmark indicates whether a form is available in the Bulgarian National Corpus (Koeva et al., 2012); if not, it is coloured in light grey. If a form is possible, even if it is missing in the corpus, it is listed as existing. For instance, even if the long definite form of a masculine MWE (realised as a subject) may not be found, it does exist if the short definite form (realised as an object) is attested (i.e. the MWE changes for definiteness), or vice versa. Heuristic procedures such as this are used to predict the (non-)existence of certain forms.

Figure 2 is a visualisation of a verbal MWE listing the various fields used in the description of the components. The validation of verbal MWEs in corpora will be undertaken in due course.

3 Conclusion

The unified approach proposed includes several layers of information and applies a set of procedures for partial automatic description. The approach allows the enrichment of the description with language-specific features. The ongoing refinement of the model has been informed by the findings of joint work on verbal MWEs in Bulgarian and Romanian (Leseva et al., 2020).

Acknowledgements

This research is carried out as part of the project *Semantic Resources and Language Technologies (Lexical-Semantic Networks and Language Models)* of the Institute for Bulgarian Language, Bulgarian Academy of Sciences (2023–2025).

References

Timothy Baldwin, Colin Bannard, Takaaki Tanaka, and Dominic Widdows. 2003. An empirical model of multiword expression decomposability. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, pages 89–96. ACL.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

Svetla Koeva. 2007. Multi-word term extraction for Bulgarian. In *Proceedings of the Workshop on Balto-Slavonic Natural Language Processing*, pages 59–66, Prague, Czech Republic. Association for Computational Linguistics.

Svetla Koeva, Ivelina Stoyanova, Svetlozara Leseva, Rositsa Dekova, Tsvetana Dimitrova, and Ekaterina Tarpomanova. 2012. The Bulgarian National Corpus: theory and practice in corpus design . *Journal of Language Modelling*, 0(1):65–110.

Svetlozara Leseva, Verginica Barbu Mititelu, and Ivelina Stoyanova. 2020. It takes two to tango – towards a multilingual MWE resource. In *Proceedings of the 4th International Conference on Computational Linguistics in Bulgaria (CLIB 2020)*, pages 101–111, Sofia, Bulgaria. Department of Computational Linguistics, IBL – BAS.

Carlos Ramisch, Silvio Cordeiro, Agata Savary, Veronika Vincze, Barbu Mititelu Verginica, Archna Bhatia, Maja Buljan, Marie Candito, Polona Gantar, Voula Giouli, Tunga Güngör, Abdelati Hawwari, Uxoa Iñurrieta, Jolanta Kovalevskaitė, Simon Krek, Timm Lichte, Chaya Liebeskind, Johanna Monti, Parra Escartín Carla, Behrang QasemiZadeh, Renata Ramisch, Nathan Schneider, Ivelina Stoyanova, Ashwini Vaidya, and Abigail Walsh. 2018. Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *The Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, pages 222–240.

Ivan Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *A. Gelbukh (Ed.) Proceedings of CICLing 2002*, pages 1–15. Springer-Verlag Berlin Heidelberg.

Agata Savary. 2008. Computational Inflection of Multi-Word Units. A contrastive study of lexical approaches. *Linguistic Issues in Language Technology*, 1(2)).

Agata Savary, Silvio Cordeiro, and Carlos Ramisch. 2019. Without lexicons, multiword expression identification will never fly: A position statement. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 79–91. Association for Computational Linguistics.

Ivelina Stoyanova, Svetla Koeva, Maria Todorova, and Svetlozara Leseva. 2016. Semi-automatic Compilation of a Very Large Multiword Expression Dictionary for Bulgarian. In *Proceedings of GLOBALEX 2016: Lexicographic Resources for Human Language Technology, Workshop at LREC-2016, Portorož, Slovenia, May 24, 2016*, pages 86–95.