



WG1 | WG2

# Treating Multiword Expressions with a view to Morphologically Rich Languages

Svetlozara Leseva | [zarka@dcl.bas.bg](mailto:zarka@dcl.bas.bg) | <https://dcl.bas.bg/>  
Department of Computational Linguistics  
Institute for Bulgarian Language, Bulgarian Academy of Sciences



## Objective

- A uniform linguistic description focusing on:
  - the representation of the structural, morphological, morphosyntactic, word-order, and other features of Bulgarian MWEs;
  - extension to MWEs for other languages;
  - with a view to the automatic recognition and annotation of MWEs in running text.

## A layered linguistic description

Modularity: the inflection type of each MWE is defined as a sequence of elementary fields describing the individual MWE components in a way that allows us to model independently them and thus the MWEs':

- lexicogrammatical characteristics;
- morphosyntactic features: inflectional characteristics, morphosyntactic constraints;

*informatsonni tehnologii* (pluralia tantum)

'information technologies' = plural

(Az) imam **zlatno sartse** – (Nie) imame **zlatni sartsa**

(direct object agreeing in number with the subject)

'(I) have a heart of gold' – '(We) have hearts of gold'

- structural characteristics:
  - internal syntactic structure;
  - the list of head and dependents;
  - possible variations in their linear order.
- syntagmatic properties:
  - syntactic transformations and constraints;
  - optional components and possible insertions:

*vzemam reshenie* – *vzemam **vazhno** reshenie*

'make a decision' – 'make an **important** decision'

- possible derivations and other transformations:

*vzemam reshenie* – *vzemanе na reshenie*

'make a decision' – '(the) making of a decision'

- (if applicable): subcategorisation frame, argument transformations and selectional restrictions.

## A lexicon of MWEs

- The lexicon includes:
  - 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.

## Validation of MWE forms in BulNC

- The MWE forms are automatically generated and:
  - validated against the Bulgarian National Corpus (the attested forms are marked by a ✓); OR
  - heuristically confirmed (not marked); OR
  - non-confirmed (in gray):

лимфен възел	Мед.	NHM24B	фикс.	eng-30-05430095-n
лимфен възел		NMso	✓	
лимфният възел		NMsl	✓	
лимфния възел		NMsh	✓	
лимфни възела		NMpb	✓	
лимфни възли		NMpo	✓	
лимфните възли		NMpd	✓	

лимфен възел:1; лимфна жлеза:1; всяко едно от малките телца, изградени от лимфна тъкан, разположени на интервали по хода на лимфните съдове; откриват се в много части по тялото, например в слабните, в аксилата, зад ухото; действат като филтри за лимфата, като не позволяват на чуждите частици да попаднат в кръвообращението, произвеждат и лимфоцити

лимфна система	Мед.	NHF11	фикс.	eng-30-05396366-n
лимфна система		NFso	✓	
лимфната система		NFsd	✓	
лимфни системи		NFpo	✓	
лимфните системи		NFpd	✓	

Nominal MWEs (*limfen vazel* 'lymph node', *limfna sistema* 'lymphatic system') with morphosyntactic and semantic description, link to WN synset and validation in BulNC.

## Description of verbal MWEs

BG: <i>удрям джакпота</i> – EN: <i>hit the jackpot 'succeed by luck'</i>	
Synset ID / MWE ID	eng-30-02524739-v / bg_2291
MWE lemma / Abstract lemma	<i>удрям джакпота</i> / <i>удрям джакпот</i>
Morphosyntactic features	<i>удрям.V_IMPERF_r1s джакпота.Nsh</i>
Head and head inflection type	<i>удрям.V_IM TT_S3_01</i>
Head restrictions	none
Dependent and dep restrictions	<i>джакпота</i> / 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.: <i>удрям <b>големия</b>/Ash джакпот/Ns0</i>
External elements	regular (question particle   subj   AdvP..)
PARSEME type	VID
Register and connotation	Colloquial; -0.125 +0.25
Derivational relations	<i>удряне на джакпота</i>

Description of verbal MWEs.

\*In pink: directions for future work.

## 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).