ПЛОВДИВСКИ УНИВЕРСИТЕТ „ПАИСИЙ ХИЛЕНДАРСКИ“ – БЪЛГАРИЯ
НАУЧНИ ТРУДОВЕ, ТОМ 53, КН. 1, СБ. А, 2015 – ФИЛОЛОГИЯ,
PAISII HILENDARSKI UNIVERSITY OF PLOVDIV – BULGARIA
RESEARCH PAPERS, VOL. 53, BOOK 1, PART A, 2015 – LANGUAGES AND LITERATURE

ON THE AUTOMATIC RECOGNITION OF BULGARIAN VERB IDIOMS

Maria Todorova
Institute for Bulgarian Language, Bulgarian Academy of Sciences

The paper presents a formalized description of Bulgarian verb idioms aiming at their preprocessing in text. We present a uniform lexicalized description of 1,000 Bulgarian verb idioms, covering categorical, pragmatic and grammatical information. The method for formal representation of idioms includes a morpho-syntactic dictionary covering both paradigmatic characteristics of verb idioms and a system of rules covering their syntagmatic characteristics. The linguistic information follows the DELA formalism, created by M. Gross.

***Key words*:** verb idioms, electronic dictionaries, Bulgarian language, Natural language processing, language technologies

1. Introduction

The term idiom is used for a fuzzy category whose definition and investigation are unclear even nowadays (Nunberg et al. 1994). What is obvious is that idioms are phrases or sentences that involve some degree of lexical, syntactic, and/or semantic idiosyncrasy. The study of verb idioms and their inflection with a view to their automatic identification is a new and relatively unexplored field, especially in Bulgarian Computational Linguistics. Their automatic identification requires lexical resources with description of their wordforms. Fellbaum (2005) underlines that the specific behaviour of idioms is a signal for the need of their distinct formalized treatment in a computational lexicon. This task is not trivial, as in most cases regular grammar rules are not applicable for the class. Verb idioms, especially in morphologically rich languages, are characterised with inflectional irregularities and lexical and syntactic flexibility.

We propose a formalized representation for the encoding of the grammatical and syntactic behaviour of verbal idiomatic expressions. It combines dictionaries and a set of local rules for the automatic acquisition of verb idioms in text. The description is applicable in all tasks related to automatic processing of texts and contribute to the correct automatic identification of the grammatical and lexical meaning of any particular lexical unit**.**

2. Verb idioms with a view to their formal representation

The peculiarities of idioms origin from their graphical form and from their semantic characteristics. They consist of two or more component words and they have a constant referent. This reflects their functional characteristics – idioms represent different levels of morphological, syntactic, distributional or semantic irregularities and at the same time this is the reason for their homonymy with free expressions.

With a view to the automatic recognition we consider for verb idioms all idioms with verbal head. Verb idioms have rich inventories of synthetic and analytical verb forms combined with a complex and flexible word order and different structural peculiarities, such as mandatory components, discontinuous components, etc. The component structure of verb idioms results in the variations of components order and insertions. For example *paham nosa si* (literally – to put my nose in sth. 'to be very curious') can be transformed in *paham si nosa*, or *nosa si paham* and at the same time allows modifier insertion *paham si (lubopitniya) nos*. Morphological irregularities in comparison with free expressions can be illustrated from the lack of singular forms of *broim se na prasti* (literally – you can count us on fingers 'we are a few') or the fixedness in 3-rd person of *blizo e do uma* (literally – it's near the mind 'sth. is understandable easily'). Another challenge that poses the description of verb idioms is the determination of their lemmas. This task is open not only for Bulgarian (Todorova 2009), but in world practice (Savary 2005). Functional specifics of idioms reflect in incorrect recognition in the processes of lemmatization, tagging and sense definition.

3. Creation of the dictionary

The construction of formalized morpho-syntactic dictionary of idioms poses some specific tasks as the extraction, selection and normalization of lexical units. We extracted 1,000 Bulgarian verb idioms according to the frequency of their verbal head from a database of 27, 900 idioms excerpted from reliable dictionaries of Bulgarian idioms. The preprocessing and normalisation of lexical units is described in details in Todorova (2015). The unification of the lemmas of selected verbal idioms was performed manually. The next step was to provide coverage of all regular and irregular forms in the description.

3.1. Formalizing the morpho-syntactic properties of verb idioms

The formalized description of verb idioms' paradigm, we propose, combines their paradigmatic and syntagmatic features. The theoretical background of our dictionary is the conceptual framework for morphosyntactic description of MWEs, proposed by Koeva (2006). We applied it with

respect to Bulgarian verb idioms. The description of Bulgarian verb idioms' morphology is proposed in Todorova (2009; 2015) and resolves some grammatical issues such as: the unification of verb idioms' lemma; the inflectional paradigm of individual paradigmatic types and the possible idiomatic paraphrases. A system of inflectional types for verb idioms was formulated with a view to inflective dictionary. Those types cover both verb idioms' word-formation and word order specifics.

The unambiguous determination of the idiom lemma requires the definition of uniform rules. We apply the definition of lemma, as the most unmarked paradigmatic form of the language unit's real usage (Koeva 2008: 25) in a principle of minimalism and neutrality of abstract lemma. The principle is introduced in the dictionary: *any idiom constituent is presented in the most unloaded with grammatical features form, for which when combined with other constituents idiomaticity of the expression remains.* The idiomatic construction is presented in the form containing only constituents mandatory for idiomaticity.

As the paradigm of an idiom is a set of all its real usable word forms, the idiom word form is unique sequence of components with a unique grammatical meaning, assigned to idiom lemma. The description of paradigmatic characteristics has the following grouping <lexical unit – lemma, structure class, structure subclass, inflection type, inflection subtype>. For example i*zlizam (izlizam.VIT15:R1s) ot kojata(kojata.NFsdk) si, VC-PREP_Nk_si. (to be outrageous).*

One of the main tasks with a view to the graphical form of verb idioms and the representation of their syntagmatic peculiarities is their grouping in structural types and in formal paradigmatic subtypes respectively. Recent researches into multiword expressions (MWEs) focus the description of verbal MWEs on their components and structure (Villavicencio et al. 2004; Gregoire 2010). The idiomatic paradigm includes all quantitative and positional changes of the idiom components. Those are constructive and combinatorial characteristics as insertion, replacement and optionality of components.

From the unified lemma we identify structural classes considering the number, linear order and the part of speech of idioms' components. Word order specifics and categorial characteristics of idioms' components determine the structural types. The inflection of non-head components determines structural subtypes. The inflective peculiarities of the head verb defines idiomatic inflectional types.

In the table below we represent the most frequent structural types in the dictionary. The components within the verb idiom structure are grouped according to the degree of inflection regularness: frozen form,

semi-frozen, non-frozen. We also envisage some positions within the verb idiom as a part of structure – the position of possible modifier, the position of an argument, possessive positions.

structural type	description	example	occur-rences
V-Nk	verb component with full paradigm and a noun component with frozen form	*hvarlyam kotva (to settle)*	154
V~N2~PREP_ Nk	verb component with full paradigm, object position, preposition and a noun component with frozen form	*pravya neshto na sol (break something to small pieces)*	83
V-(Nk _POSsi)	verb component with full paradigm, a noun component with frozen form and possessive modifier position	*paham nosa si (to be insolently curious)*	75
V-Nk-(na_N2)	verb component with full paradigm, a noun component with frozen form and an object position	*podavam raka na nyakogo (help to someone)*	72
V-PREP-Nk	verb component with full paradigm, preposition and a noun component with frozen form	*umiram ot stud (to feel very cold)*	65
V-PREP_(NON1_ Nk)	verb component with full paradigm, preposition, possessive modifier position and a noun component with frozen form	*padam v nechii ochi (lose authority for someone)*	52
V- (NON1_Nk)	verb component with full paradigm, possessive modifier position and a noun component with frozen form	*vdigam nechie kravno (make someone angry)*	40
V-Ak-Nk	verb component with full paradigm, adjective with frozen form and a noun component with frozen form	*vdigam byalo zname (give up)*	26

4. Dictionary format

The specific paradigmatic features of idioms determine several ways for their formalization. One option is a list of all possible paradigmatic realizations of verb idioms, but such a resource is too voluminous and difficult to elaborate manually, especially in language with rich inflectional paradigms. M. Gross (1996) proposes lexico grammatical approaches as the most ap-

propriate framework for the formal presentation of MWEs forms. They include a list of base forms and group of rules valid for certain phrasal groups. These rules are called local grammars. The creation of inflective local grammars and dictionaries can be based on different mathematical formalisms, as final state transducers (Kartunen et al. 1992; Kartunen 1993) and databases (Copestake et al. 2002). Formalized description of idioms can be built also with unification grammars (Sag. et al. 2002; Villavicencio et al. 2004). Different approaches and platforms for generation of MWE forms have been proposed, such as the parameterised equivalence class method of the DUELME database (George et al. 2013) and linear string description in the POLENG formalism (Gralinski et al. 2010).

Our description is created using the graph-based morpho-syntactic generator of MWE Multiflex (Savary 2009) which combines simple words morphology and MWE forms generation. It is one of the applications based on the DELA format and is incorporated in the Unitex[1] system. The format DELA (Dictionnaires électroniques du LADL) (Kortua and Silberstein 1990, Silberstein 1993a; 1993b), developed in the laboratory of automatic linguistic processing (LADL) at the French National Center for Scientific Research (CNRS) focuses on extensive morphological analysis of lexical units through automatic matching of words in the text with the full list of possible grammatical annotations.

The automatic word formation of a verb idiom is performed from a lemma and inflective grammars representing different inflective types. The lemmas in the dictionary and their inflective grammars are related by means of the inflective grammar name. Each inflective grammar correlates with specific inflective type through which the paradigm forms are generated and the description of the grammatical categories and characteristics is supplemented. The dictionary description is combined with a system of local syntactic rules (section 5.2 and 5.3). They allow automatic generation of the possible syntagmatic variations of Bulgarian verb idioms and their identification in a text.

5. Content of the dictionary

Our formalized description includes a dictionary of simple words (idiomatic components), dictionary of verb idioms, inflective grammars and local syntactic grammars presented as graphs.

The total number of verb idioms in the dictionary is 1000. They are divided into 30 structural types and are described with 50 inflective types in the grammar. The formal description includes a conventional idiomatic

[1] http://www-igm.univ-mlv.fr/~unitex/

lemma. Within lemma each component is labeled with the respective inflectional type (if variable) or fixed grammatical values (if not variable). Word order features of idioms are represented by rules for linear preceding of the components.

5.1. Inflection of the verb idioms' components

Verb idioms' components are listed in DELAS dictionary as simple words. The morphophonemic variations in the possible word forms are represented with inflectional grammars, describing verbal head's word-forms within the idiom. The assignment of the grammar to the idioms component provides the generation of its word forms. As illustrated on figure 1 for *izkarvam* in *izkarvam ot kojata (make s.o outrageous)*. When the component paradigm is limited in comparison with its free word counterpart, we include it in the dictionary as a new word homonymous with some of the free words and their forms.

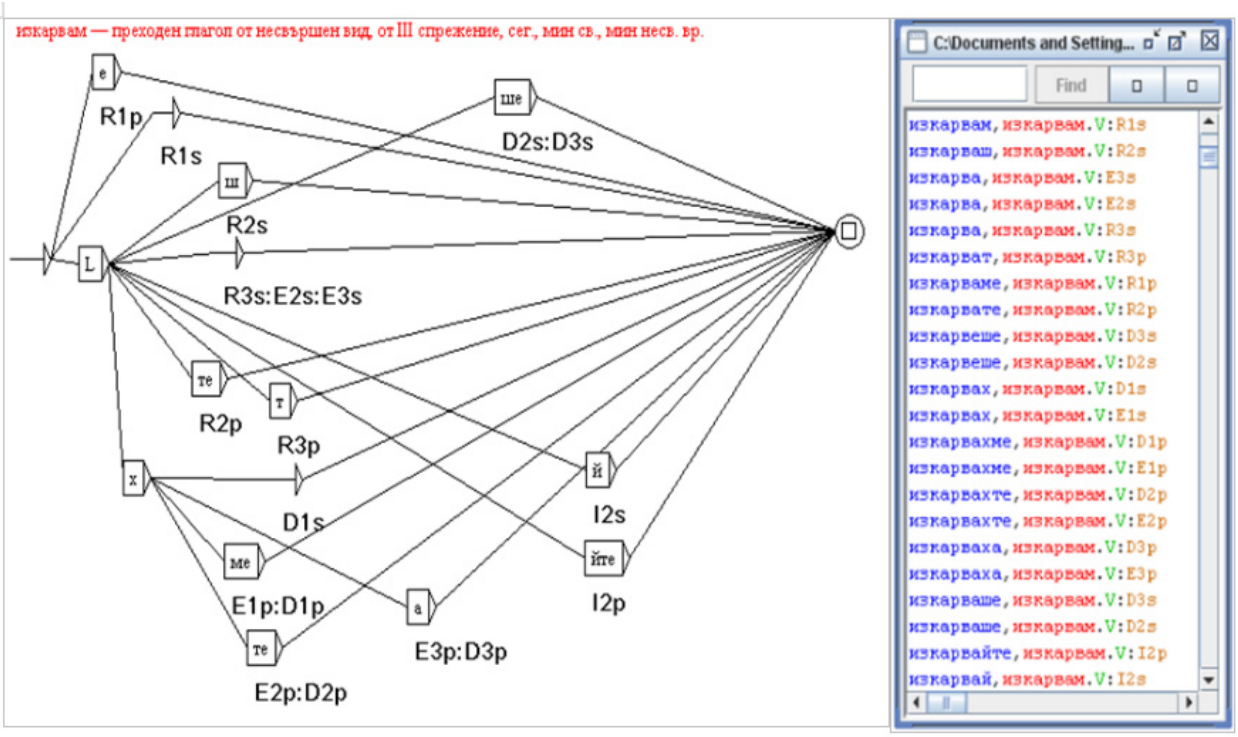

Figure 1. Generation of idioms' component forms

5.2. Verb idioms' components structure

Word order variations of idioms are described as MWE inflectional grammar, as illustrated on figure 2. for *broya(broya.V11:R1s) zvezdite(zvezdite.N6:fpdk),VC_V-Nk.* where each component is represented by a variable with definite position.

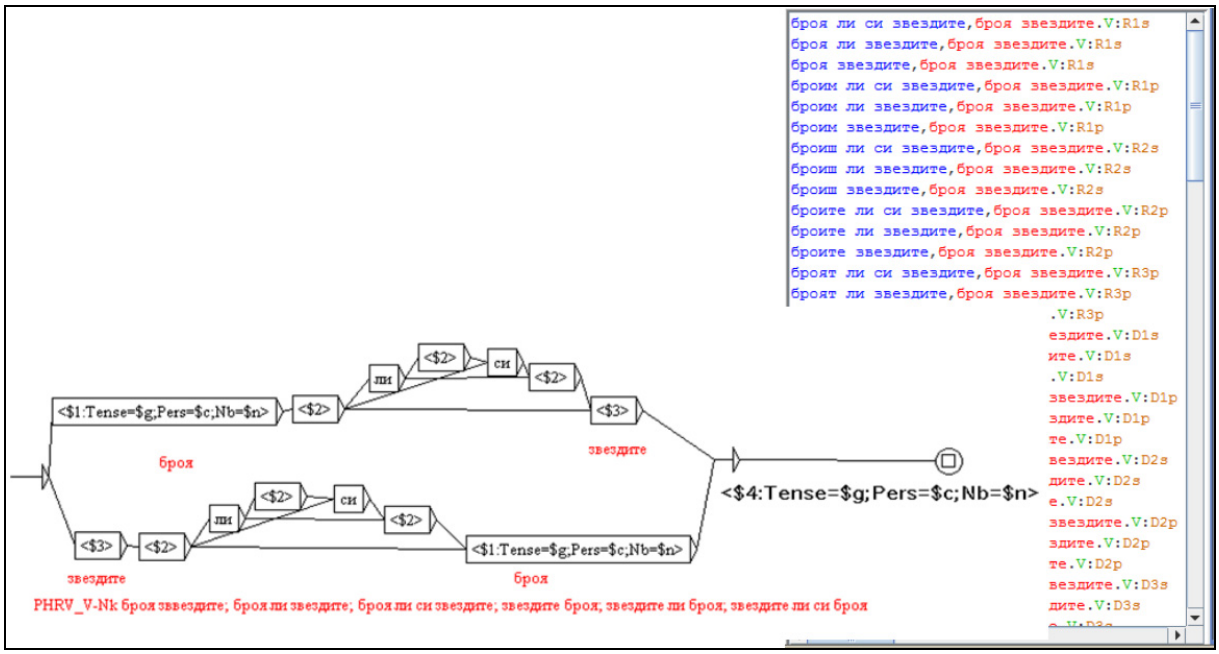

Figure 2. Idioms' structural variations

The variable number labels the component's position within lemma. They depend on structural subtype and cover: a) a fixed word order – *pravya na mat i maskara* (humiliate); and b) determined word order (possibility of insertion of specific classes or syntactic groups) – *imam (golyama) belya na glavata (to be in (a big) trouble)*.

5.3. Verb idioms' syntagmatic description

The transformations and coordination of idiom components is represented by syntactic grammars and variable grammatical values. They represent typical paraphrases with a view to idiomatized position – *vrememeto mi nastapi (it's my moment) -> nastapi mi vremrmto -> moeto vreme nastapi.*

The coordination features are represented by unified variables „$" where the grammatical limitations are set as a variable value. The endocentric coordination (fig. 3) of a component is governed by the idiom's head. Exocentric coordination (fig. 4) is the coordination of an idiom component with the word in the idiomatized subject or object position.

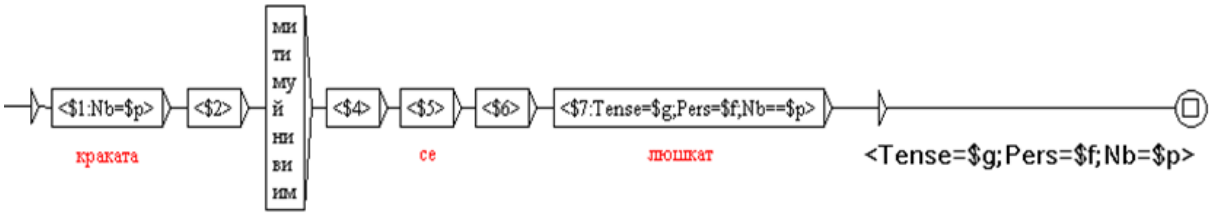

Figure 3. Endocentric coordination

Figure 4. Exocentric coordination

6. Evaluation

In order to check the coverage and the frequency of dictionary entries in a corpus, it was applied on a part of the literature texts from Bulgarian National Corpus (BulNC[2])[3].

In 18.150 texts with 15,780,435 words were found totally 80.385 examples of verb idioms. With frequency over 1000 are 15 verb idioms, 80 verb idioms from the dictionary has frequency between 100 and 1000. With frequency between 11 – 100 are 163 verb idioms and 235 verb idioms has frequency between 1 – 10. There are no examples for 507 verb idioms.

As we can see from the results half of the dictionary entries have very broad coverage in the texts. This makes us conclude that the formalized description is quite reliable although dependable on the linguist's subjective opinion.

We consider that the Multiflex tool is expressive and efficient with respect to Bulgarian verb idioms. Of course as any other formal frame it determines some limitations and the need of compromises and clever decisions in some cases. Those are the overgeneration of the limited paradigms of idioms' components which are homonymic with some free word forms. The incorporation of possessive and objective syntactic transformations, as well as using subgraphs for pronoun forms within idoms' inflectional structure is also problematic.

The lack of occurrences for half of the dictionary entries in the corpus excerpt from Bulgarian literature we used arises questions in a few directions: what is the degree of actualness of dictionaries of Bulgarian idioms, based on Bulgarian literature classics (19-th and the beginning of 20-th century), nowadays and the need of precise corpus selection for testing concrete idioms.

Another conclusion with a view to semantic disambiguation of verb idioms arises from the fact that the most frequent verb idiom examples in the tested corpus are homonymic free phrases: *vdigna glava (to be proud/ to raise head)* with 6737 occurrences; *ostavyam na mira (leave alone/ leave to the peace)* with 2322 occurrences; *treska trese nyakogo (to have*

[2] http://dcl.bas.bg/bulnc/
[3] The test was performed by Ivelina Stoyanova

fever/ to be nervous) with 2033 occurrences and *padna na kolene (to beg/to fall on knees)* with 1526 occurrences. Still there is no criteria to conclude which part of the occurrences are idiomatic and what is the percentage of free phrases among them.

7. Conclusions and Future work

Idioms are important both for the creation of specialized or wide coverage computational lexicons, and for the development of natural language processing (NLP) systems (Sag et al., 2002). The paper describes a knowledge-based method towards formalized description of Bulgarian verb idioms in an electronic dictionary. It is using the DELA formalism and the Multiflex tool application. The future improvement of the dictionary includes enlargement of dictionary entries, types and grammars and incorporating the verb idioms in the BulNet structure. Though the verb idiom forms, represented in dictionary are generated automatically from the inflectional types, the manual description of types is considerably slow and laborious. So in we plan to combine the formalized frame description with statistical heuristics on a corpus in order to get more data and to escape from subjectiveness in determining idioms' word formation. Further improvement of the verb idioms recognition is planned in testing the dictionary in tagging of large texts and distinguishing between idioms and homonymic free phrases in the semantic disambiguation.

REFERENCES

Copestake et al., 2002: Copestake, A., et al. Multiword expressions: linguistic precision and reusability. // *Proceedings of the 3-rd International Conference on language resources and evaluation* (LREC 2002), 2002, 1–7.

Fellbaum 2005: Fellbaum, C. Theories of human semantic representation of the mental lexicon. // Cruse, D. A. (Ed.). *Handbook of Linguistics and Communication Science*, Berlin, Germany: Walter de Gruyter, 1749–1758.

George et al. 2013: George, M., G. Francopoulo. Model Description. // Francopoulo, G. (ed.). *LMF Lexikal Markup Framework*. Wiley Online Library. Chap. 2, 19–40.

Gralinski et al. 2010: Filip, Gralinski, Krzysztof Jassem, and Michał Marcinczuk. An environment for named entity recognition and translation. // *Proceedings of the 13th Annual Meeting of the European Association for Machine Translation* (EAMT'09), Barcelona, 88–96.

Gregoire 2010: Gregoire, Nicole. DuELME: a Dutch electronic lexicon of multiword expressions. // *Language Resources & Evaluation* (2010) 44, 23–39.

Gross 1996: Gross, M. Lexicon-grammar. In: BROWN, K., and J. Miller, eds. // *Concise Encyclopedia of Syntactic Theories*. Oxford: Pergamon Press, 1996, 224–259.

Kartunen et al., 1992: Kartunen, L., R. M. Kaplan, and A. Zaenen. Two-level morphology with composition [Online report]. // *Proceedings of the 14-th Библиография 181 International Conference on Computational linguistics* (COLING'92). Vol. 1. Association for Computational Linguistics, 1992. 141–148. <http://bit.ly/14ue8re.>

Kartunen 1993: Kartunen, L. Finite-state lexicon compiler. ISTL-NLTT-1993-04-02 [Technical Report]. Palo Alto: Xerox Corporation Research Center, April 1993.

Koeva 2006: Koeva, Sv. Inflection Morphology of Bulgarian Multiword Expressions. // *Computer Applications in Slavic Studies*, Bîyan Penev Publishing Center, Sofia, 201–216, 2006.

Koeva 2008: Коева, Св. Българският ФреймНет. Семантико-синтактичен речник на българския език – концептуален модел. // Коева, Св., съст. *Българският ФреймНет: Семантико-синтактичен речник на българския език*. София: БАН, 2010, 5–57.

Kortua, Silberstein 1990: COURTOIS, B., and M. Silberstein. Dictionnaires electroniques du francais. // *Langue françai*se, 1990, 87(1), 3–4. <bit.ly/1yf0O4M>

Nunberg et al. 1994: Nunberg, G., I. Sag, and T. Wasow. Idioms. // *English. Language.* 1994, (70), 491–538.

Sag et al. 2002: Sag et al., 2002. Multiword expressions: a pain in the neck for NLP [Online report]. // *Proceedings of the 3rd International Conference on intelligent text processing and computational linguistics (CICLing–2002)*. Berlin, Heidelberg: Springer, 2002. 1–15. <http://bit.ly/1KJc6ou.>

Savary 2005: Savary, A. A formalism for the computational morphology of multiword units. // *Archives of control sciences*, 2005, 15(3), 437–449.

Savary 2009: Savary, A.: Multiflex: A Multilingual Finite-StateTool for Multi-Word Units. // Maneth, S. (ed.) *Implementation and Application of Automata*. LNCS, vol. 5642, 237–240. Springer, Heidelberg (2009)

Silberstein 1993a: Silberstein, M. *Dictionnaires électroniques et analyse 180 automatique de textes: le système INTEX*. Paris: Masson, 1993.

Silberstein 1993b: Silberstein, M. Les groupes nominaux productifs et les noms composés lexicalisés. // *Lingvisticæ investigationes*.1993, 17 (2), 405–425.

Todorova 2009: Тодорова, М. Лексикализирана граматика на български глаголни фразеологизми. // *Български език*. 2009, (3), 70–83.

Todorova 2015: Todorova, 2015. Todorova, Maria. *Typology and Properties of Multiword Expressions in Bulgarian.* Verb idioms. PHD Thesis, Sofia 2015. <http://dcl.bas.bg/Disertacia_M.Todorova/M.Todorova_disertacia.pdf>

Villavicencio et al. 2004: A. Villavicencio, A. Copestake, B. Waldron, F. Lambeau. The lexical encoding of MWEs. // *Proceedings of the ACL 2004 workshop on multiword expressions: Integrating processing*. Barcelona, Spain, 2004, 80–87.