Proceedings of the 2007 International NooJ Conference

Edited by

Xavier Blanco and Max Silberztein

Cambridge Scholars Publishing

Proceedings of the 2007 International NooJ Conference, Edited by Xavier Blanco and Max Silberztein

This book first published 2008

Cambridge Scholars Publishing

12 Back Chapman Street, Newcastle upon Tyne, NE6 2XX, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-4438-0053-8, ISBN (13): 978-1-4438-0053-2

TABLE OF CONTENTS

Morpho-Syntactic Properties of Bulgarian Verbal Idiomatic Expressions

Maria Todorova
Institute of Bulgarian Language – Bulgarian Academy of Sciences
maria@ibl.bas.bg

ABSTRACT

This paper is focused on the Bulgarian verbal idioms and the problems related to the representation of their structure. The presented work is aimed at the identification, recognition and translation of Bulgarian idiomatic expressions for the purposes of WSD and machine translation. The main goal is to provide a sufficient framework combining internal idiomatic structure, inflectional and semantic information for the description of the wide variety of idiomatic expressions.

1. Introduction

NLP applications directed to the multilingual natural language processing such as machine translation, information retrieval, text summarization, paraphrasing, etc. encounter difficulties with the border between compositional and non-compositional word meaning, and the word ←→ phrase translation, illustrated by idioms and collocations. To the best of our knowledge at present there is no lexical resource that represents entirely the syntactic, morphological and semantic properties of Bulgarian collocations and idioms. In this paper will be explored some peculiarities of idioms variation. We refer to Bulgarian data as evidence for our claims. The focus of interest here are some structural and functional properties of verbal semi-idioms, one of the subtypes of the semi-decomposable items. Our aim is to elaborate detailed classification of their pragmatic behavior and syntactic representations which generalize the level of their flexibility and the relevant selection restrictions. We use NooJ local grammars to describe some productive patterns of idiomatic lexical items and some combinatorial constraints. The grammars capture certain idiom types using lexical, syntactic and / or semantic criteria of regularity. The semi-decomposable idioms are formally described by means of syntactic graphs in combination with local grammars.

2. Linguistic Resources

Our investigation is based on the linguistic resources elaborated at the WSD project of the Department of Computational Linguistics - lexical semantic database BulNet[1] and the Bulgarian semantic corpus BulSemCor [Koeva et. all 2006].

Wordnet covers a large number of MWEs [Fellbaum, 1998]. Idiomatic expressions are multiword expressions incorporated as separate items in the English database. Our analyses are based on the idiomatic expressions extracted from BulNet and from the Bulgarian semantic corpus. The lexical semantic database BulNet is a resource for defining the meaning of semantically annotated lexical items. The specific properties of idiomatic expressions give rise to problems as distant components, elliptic use of heads of the expression, variable word order structure etc., regarding their annotation and incorporation in BulNet. In the process of annotation of the Bulgarian semantic corpus 1068 compound words were selected and manually subclassified into 4 groups – named entities, gramaticalized phrases, compounds and idiomatic expressions. In order to test and investigate the hypothesis a database of idioms was created from four of the most popular dictionaries of Bulgarian idioms. The database consists of over 200 000 lexical units, including dialect, colloquial and literary ones. A list of dictionary units was derived using the idiomatic database and a test corpus version consisting of 159 texts was created from Internet. The idiomatic database is used both as source for dictionary units and as test corpus. The analysis of usage data provides new insights concerning idiomatic expressions and leads to a framework for the description of idioms which accounts for the relation between syntactic structure, meaning and modification types. The table represents the current state of lexical resources and the extracted data.

	Lexical units	**MWE**	**Idiomatic expressions**	**Verbal Semi-idioms**
Lexical resources	BulSemCor	approx. 1068 units	approx 60 units	approx 35 units
	BulNet	approx. 12 600 units	approx 250 units	approx 130 units
	Bulgarian Idiomatic Dictionary List	over 200 000 units		approx. 600 units

Table 1. derivation of verbal semi-idioms from linguistic resources – work in progress

3. Idiomatic Subclass in Focus - Semi-idioms

The General Class of Multi-Word Expressions (Non-free Phrases [Melchuk, 1995]) represents a wide scope of variability, ranging from closed sets, which could be enlisted, to open sets, based on specific properties. Their idiosyncratic behavior is characterized by "a lack of compositionality, manifested at different levels of analysis - lexical, morphological, syntactic, semantic,

1 http://dcl.bas.bg/BulNet/general_en.html

pragmatic and statistical" [Baldwin, 2006]. In order to determine separate idiomatic subcategories we use the formal representation of Theory of Lexical Functions and Meaning text theory (MTT) [Mel'cuk, 1995] which represents the linguistic sign with a set of atomic features <'X"; /X/; SX, SIT, ConceptR>. The signified is denoted with 'X", /X/ denotes the signifier of the linguistic sign, with SX is named the syntactic sign - the set of all necessary data on the co-occurrence of the sign, and with "O" is denoted the linguistic union operation. ConceptR (conceptual representation) and SIT (a situation) are used for the representation of semantic properties of the linguistic sign that are not observed here.

Semi-idioms, also called lexicalized metaphorical expression [Van der Linden, 1991] or analyzable idioms [Erbach, 1992], undergo variations and / or take semantics outside the expression, but do not follow the regular language rules and are restricted in a set of realizations. Verbal semi-idioms are a type of multi-word predicates - the head of the expression is a verb which occurs in the whole idiom paradigm. The idiomatic meaning is connected with the restrictions on subcategorization and combinatorial properties of the verb, represented in one of its meanings. The present analysis requires the idiomatic NP and the idiomatic verb to co-occur. The semantic properties of idioms explain some asymmetry in the grammatical properties of idiomatic phrases. The meaning of the whole expression is partly independent from the meaning of the components, that can have figurative sense or to vary in a semantic or lexical set of possible realizations. Formally they can be represented as AB = <'AOBOC"; /AOB/> | 'C"≠'A" & 'C"≠'B". The signified of the expression 'X", includes the signified of the two components 'A" and 'B" and additional signified 'C", different from 'A" and 'B". For example if in the expression *разбивам сърцето (break one's heart)*, *разбивам (break)* is considered for *A* and *сърцето(heart)* for B and the meaning '*disappoint someone's hopes and intentions for love*' is C, its obvious that the meaning of the expression includes the meanings of its components and additional meaning which arises from their specific combination. 'C" can be also expressed by signifiers /D/ or /R/, so that 'C"= /AOB/, /AOR/, /AOD/ | 'R", 'D" € 'B" and 'C"Ø 'A", 'C"Ø 'B", C"Ø 'R", C"Ø 'D". These are idiomatic expressions in which one of the components can vary in a set of semantically similar components. For example if in the expression *търся / диря / гледам под дърво и камък* literally *"seek/look for under a tree and stone"* meaning '*look for everywhere possible*' the verbal component can be expressed with different synonymic verbs.

4. Formal Classes of Verbal Semi-idioms:

A criterion for the identification of the semantic, morphological or phonological behavior of the idiom types are modifications by permutation, addition, replacement, etc. that cause a loss of the idiosyncrasy of the phrase [Bauer et al., 2004]. The morpho-syntactic behavior of semi-idioms is characterized with: restrictions on substitutability of each element; syntactic irregularity; meaning that cannot be predicted from a surface form; single-word paraphrasability; word order substitution and restricted paradigmatic productivity. The different kinds of variations of the non-fixed forms of idiomatic phrases concern their internal idiomatic structure (structural and morphological properties) and their external idiomatic structure (semantic information and syntactic properties). Following Sathi (2006) we outline the relevant phenomena that should be examined for the purposes of determination of the structural and grammatical properties of idioms in order to distinguish the relevant description subclasses:

a) The behavior of idioms on the syntagmatic axis - modification of idiom components by adjectives, adverbials, genitive, prepositional attributes etc.

b) The behavior of idioms on paradigmatic axis - substitution of idiom components by other words or phrases - by synonyms, hyponyms, and hyperonyms.

c) The degree of the morpho-syntactic flexibility of verb – argument realizations, passivization, pronominalization etc.

4.1. Internal structure

The internal structure of the idiomatic constructions or the description of their "surface structure constituents" is examined to distinguish the homonymous idiomatic and non-idiomatic phrases and to resolve the ambiguity between idiomatic constituents and their non-idiomatic counterparts, free words. The description of idiom architecture poses the problem of the definition and recognition of the boundaries of the "idiomatic chain" [O'Grady, 1998] and the flexibility of idiomatic composition. We outline the following criteria with a view to categorization of classes of semi-idioms with similar paradigmatic behavior:

- The number and order of elements in the inner structure of the expression.
- The inflective type of the head element of the expression. Idioms comprising two changeable elements combine grammatical features of different categories.
- The deviation and defectiveness of the paradigm of the idiomatic expressions.

4.1.1. Determination of relevant inflection subtypes according to the number and order of constituent elements

The list of verbal idiomatic units, derived and subclassified from BulSemCor, BulNet and Bulgarian Idiomatic Dictionaries (see section 2), represents a wide variety of entry structures, so a manual selection and association of each token with part of speech tag entries was made. On one had we aimed to exclude idiomatic expressions which are outside of our scope of interests as exclamations, sentences, proverbs etc. and on the other hand we needed to determine different formalized structures. We used NooJ applications to encode certain number of entries by hand and qualify them as training corpus for further investigation of context, syntactic features, grammatical variations etc.

Each semi-idiom entry was represented as linear sequence of POS determined words and all optional elements were separated into subclasses. A simple tagset of ten standard part of speech labels was used[2] and the elements that remain frozen were identified with "k". Over thirty five different POS sequences for verbal idioms were described and are considered for

[2] N – noun; V – verb; A –adjective; NUM – numeral; PRO – pronoun; ADV – adverb; PREP – preposition; CONJ – conjunction, S – sentence.

constructional substructures. Identfication models were used to extract candidate expressions from unclassified idiomatic data by means of regular expressions and are associated with the relevant constructional class.

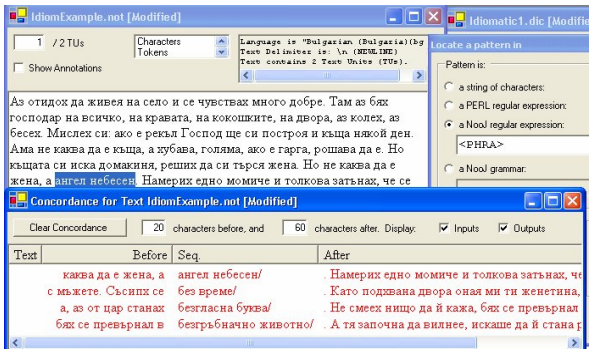

Fig.1 Examining the idiomatic subtype with constructional classes and NooJ concordance and regular expressions

Examples of some of the verbal constructional classes:

The representation VP/V+Nk/ of semi-idioms like *броя гаргите* (literally *count the crawls,* meaning '*be distracted'*) means that expression is a linear combination of verb and noun, fixed in its determinate form.

The representation VP/V+Ak+Nk/ of semi-idioms like *вадя кирливите ризи* (literally *take out the dirty shirts,* meaning '*reveal someones secrets'*) means that expression is a linear combination of verb and adjective, fixed in its determinate form and noun, fixed in its plural form.

The representation VP/Nk+V+PREP/ of semi-idioms like *боб хвърлям за* (literally *throw beans for,* meaning *'i don't care even a bit of something*') means that expression is a linear combination of fixed in its non-determined singular form noun, verb and obligatory preposition.

4.1.2. Inflective types of verbal semi-idioms

Most of the idiomatic expressions allow morphological inflection, so for the formal description it's important to define which of the elements of the idiomatic expression can be inflected. Constructional classes are subdivided additionally according to the inflection properties of their elements. Determination of relevant inflection subtypes of semi-idioms is based on their lexical head – the verb, but in some idioms the verb is in constant third person form and changes of person in idiomatic paradigm is led by personal pronoun. There are also constructions in which both the verb and the pronoun vary in form. We subclassify the verbal semi-idioms into three groups:

- Bulgarian verbal semi–idioms whose paradigm is defined by the verbal component. Their forms vary in number, person, tense and they are characterized with gender and determinateness when the participle of the verb is used. For xample: ***давам*** *ухо* literally "*give ear*" meaning '*eavesdrop'*.
- Bulgarian verbal semi–idioms whose paradigm is defined by the pronoun component. Their forms vary in number, person, case and gender. For xample: *без да* ***ми*** *мигне окото* literally "*without a wink of my eye*" meaning '*without scruples*'.
- Bulgarian verbal semi–idioms whose paradigm is defined by the verbal and pronoun component both. Usually in those constructions the verb paradigm is defective and it's fixed in third person forms, but varies in number and tense. For xample: ***лови ми*** *окото* → ***ловеше му*** *окото* literally "*it catches his eye*" meaning *'its attractive in some way'*

Every group is correspondingly divided into sybtypes according to the concrete inflection type of the head verb and the restrictions on its idiomatic paradigm, if any.

4.1.3 Fixed paradigm properties and morphological anomalies

Some properties of idioms violate the general criteria of regularity in language. Considering Sailer (2003) and Soehn (2006) criteria for regularity: a) every lexical item is morphologically of a regularly built shape; b) every word belongs to a regular inflectional paradigm, we outline the violations of these criteria in the paradigm of Bulgarian idiomatic expressions. The idiosyncratic properties on the morphological level, represented as anomalies and frozen properties show that idioms don't follow the usual lexical inventory of a language. For the classification of subtypes except for the examined entries we also followed the conclusions and examples in literature on Bulgarian idioms [Cholakova, 1968; Nicheva, 1987].

A) Paradigmatic constraints of a form of the head word – a particular form in the regular paradigm of the head word excludes the idiomatic meaning.

- Limitation in usage of tenses. For example the use of the verb in aorist form in the expression *бия на очи,* literally "*beat eyes*", meaning '*be too obvious*', causes loss not only of the idiomacity of the expression, but of its meaning at all **?бих на очи*.
- Limitation in usage of number. For example the use in singular form of both the verb and noun in the expression *броим се на пръсти* literally "*we count on fingers*", meaning '*to be too few for something*' causes loss not only of the idiomacity of the expression, but of its meaning at all **?броя се на пръсти; *?броим се на пръст; *?броя се на пръст*.
- Limitation in usage of verb aspect pairs. For example the use of finite correspondence of the head verb in the expression *блъскам си главата* literally "*hit one's head*", meaning '*think hardly on something*', causes loss of the idiomacity of the expression *?блъсвам си главата*.

B) Defectiveness of paradigm formation – a particular form in the paradigm of the head word is constructed in irregular way.

- Idioms with defective formation of one of their forms - For example in the expression ***глътвам*** *си езика* literally "*swallow one's tongue*", meaning '*loose my ability to talk from great emotion*' the future tense of expression with its head verb in infinite aspect is constructed with its finite verb correspondence → *ще си* ***глътна*** *езика*
- Idioms which have only one paradigm form. For example in the expression ***пиши*** *го на челото си* literally "*write it*

on your forehead", meaning '*demonstrate the exaggerated value of something done*' the verb "*пиши*" has only imperative form in its idiomatic reading.

C) Fixedness of a non-head element in a concrete morphological form. For example the expressions *вдигам на крак* literally "*pick up on foots*", meaning *`mobilize`* and *вдигам на крака* literally "*pick up on my feet*, meaning *`cure`,* differ only in the definiteness of the idiomatically fixed element *"крак" / "крака"*. Both expressions are almost homonymous, the change of the form of one of the components of the expression results in different meaning.

Paradigmatic Constraints were represented in NooJ with Dictionary and Negation functions of Grammar. The intersection between constructional class and the negation local grammars representations doesn't represent relevant grouping. At present only a separate entries are described.

4.1.4. Variations in the idiomatic construction

With a view to the flexibility of verbal semi-idioms different kinds of variations in their construction are possible. We investigate here the insertion of modifier or of an element from context, elision of element of the idiomatic construction and word order variations, which form the so called discontinuous idiomatic structures, represented in NooJ by graphs and local grammars (See fig 2).

4.1.4.1. Insertions of lexical components in the idiomatic constructions

It's considered that the degree of insertion of modifications in the idiom construction depends on their semantic cohesion [Cruse 1997, 40]. Idiomatic expressions typically resist interposition of external elements from context and reordering of their components. Ernst (1981) defines external, and internal for the idiomatic meaning modifications. Melchuk (1995) claims that the modifications of idioms with external for the construction elements are allowed only for the head word or to the whole meaning. As the subclass of semi-idioms is partly fixed it usually consists of components with different degree of variation and frozeness. Comparing the defined constructunal classes and the types of modifications possible for semi-idioms we can determine the following dependencies in the investigated data.

A) Insertion of an adjective in the idiomatic construction is investigated only in construction classes consisting noun or noun phrase. The modification with adjective incorporated in the structure of expression functions as intensifier of the head of two or more component NP.

- In constructions of transitive verb and noun VP/V+Nk/ or of intransitive verb, preposition and noun VP/V+PREP+Nk/ like ***завирам / забивам*** *нос* literally *put nose in something,* meaning '*show disappointment'* and like ***отивам / вървя*** *по дяволите* literally *go to hell,* meaning '*ruin'* usually adjectival modifications are not accepted.
- In constructions where the NP consists adjective like VP/V+Ak+Nk/ or VP/V+PREP+Ak+Nk/ - verb, (preposition) adjective and noun, a modification of the idiomatic component NP is usual. For example in the expressions like *навличам си (голяма) беля на главата,* literally *„bring a big problem to my head",* meaning *'inflict (great) trouble on oneself'*. When the idiomatic component is prepositional phrase the adjective is inserted after the preposition *Скривам се* ***в миша дупка*** - *Скривам се в най-дълбоката миша дупка* literally *„hide in a (the deepest) mouse hall",* meaning '*hide in a place where it's hard to find'*.
- In constructions with more than one NP like VP/V+Nk+CONJ+Nk/ - *загубвам ума и дума* literally "*loose brain and words*" meaning *'be under a very strong excitement'* a modifications are not investigated.

As pointed in section 4.2.1. ajectival modifications are typical mainly for verb - complement construction, which has turned into internal relation in the idiomatic expression, but still keeps the regular verb – object / adjunkt relation transparent. The dependency among the internal argument structure of the verb in the borders of the expression and the acceptability of modifications is also represented from the modification with possessive pronoun. It is possible for the constructions in which the selectional properties of the verb include prepositional object "to someone" like in ***играе / върви / ходи*** *по свирката / гайдата / тъпана на някой,* literally *'plays on someone's pipe '* meaning '*follow someone's will and desires*'.

B) The insertion of an adverb in the idiomatic construction is less restricted in comparison with the adjectival modification, because of the considerably unrestricted function of the head verb in idiomatic expressions without defectiveness of the paradigm. For example *давам (понякога) ухо* literally "*give ear (sometimes)*", meaning '*eavesdrop*'.

C) Insertions of particles from context is also defined of the level of fixedness of the element or elements. The interrogative particle "ли", which can be placed after every word in Bulgarian in regular constructions, is not used inside a fixed idiomatic construction. For example in the expression *даваш* ***ли*** *си (точна, ясна) сметка,* literally "*give a clear account"* meaning '*evaluate the situation properly*', the interrogative particle "ли" cannot separate the idiomatic combination *?даваш си ясна* ***ли*** *сметка?*.

4.1.4.2. Replacement of a component with pronoun or adverb usually is not allowed in semi-idiomatic constructions, because the idiomatic NPs have nonreferential character. For example: *Скривам се* ***в миша дупка*** literally *„hide in a mouse hall"* the idiomacity of the expression is lost in *?Скривам се там. ?hide there* or*? Скривам се в такава дупка." ?hide in such a hole.*

4.1.4.3. Elision of component usually is not allowed in semi-idiomatic constructions. Look at the example in **4.1.4.2. A)** *Скривам се* ***в миша дупка*** *?Скривам се* ***в дупка*** *?hide in a hall"*.

4.1.4.4. Word order variations of elements within the semi-idiom structure

Verbal semi-idioms have a considerably free word order. Because of the relative autonomy of verb head, in those constructions almost all typical for free phrases variations are possible. The verbal component can be placed on the left or on the right of the idiomatic component. *Търся/ диря* ***игла в купа сено*** *(търся/ диря)*. Movement reordering transformations are allowed in fixed

expressions in which the paradigmatic variation depends on the pronoun. *Мед **ми** капе на сърцето → мед капе на сърцето **ми** → на сърцето **ми** мед капе → капе **ми** мед на сърцето.*

Discontinuous Structures of Verbal Semi-idioms are represented in Nooj by graphs and local grammars. We use Nooj local grammars to describe some productive patterns of idiomatic lexical items and some combinatorial constraints. The grammars capture certain idiom types using lexical and syntactic criteria of regularity. The semi-decomposable idioms are formally described by means of syntactic graphs in combination with local grammars. A detailed description of the paradigmatic and syntagmatic features of semi-idioms by means of grammars and dictionaries can be used in future for the implementation of convenient and effective means of treating newly emerging idiomatic expressions, as well as in the investigation of new structural patterns and types.

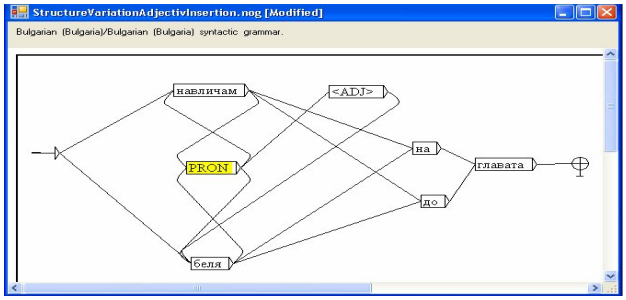

*Fig. 2. Representation of Discontinuous Structure of Verbal Semi-idiom **"навличам си беля на главата "***

4.1.5. Variability / Optionallity of one or more constituents

The variability of constituents in the idiomatic structure concerns both their structure and the problem of the definition of their borders. It's also related with the question of synonymy among idioms. According to the variations in structure two types can be determined:

- Verbal semi – idioms with optional components. They are represented as short and long variants of discontinuous structures where the optional elements are indicated. The uses of obligatory elements in combination with the optional ones give rise to different word order constructions. For example the expression *вадя му думите с ченгел (от устата)* literally "*take out his words with hook (from his mouth)*" meaning '*hardly make somebody to speak*' in its long variant has several optional word order structures. The optional component causes different movement reordering transformations – it can be placed on the left or on the right of the idiomatic component - *(от устата) вадя му думите с ченгел (от устата);* but also can be inserted within the structure of the short variant - *вадя (от устата) му думите с ченгел* **or** *вадя му (от устата) думите с ченгел* or *вадя му думите (от устата) с ченгел.*
- Verbal semi – idioms with components which vary in a closed synonym set. In comparison with the previous group, the structure of this type semi-idioms is independent of element variations. According to the component which vary they are:

a) Verbal semi – idioms with variative nominal components. For example: *трия сол (лук) на главата* literally "*smash onion (salt) on someone's head*", meaning *'mumble someone'.*

b) Verbal semi – idioms with variative verbal components. For example: *нещо ми **мърда / хлопа / шава*** literally "*something is moving / rambling in my head",* meaning '*one is not psyhically stable'.*

4.2. External Structure

The degree of the morpho-syntactic flexibility of verb is defined from its argument realizations and from restrictions on syntactic transformations like passivization and pronominalization. We outline two relevant criteria for the description of external structure of idioms.

- The syntactic role of idiomatic verbal arguments and the availability of a non-idiomatic counterpart, belonging to the same part of speech.
- The regular or irregular semantic transformations of the components.

4.2.1. Number and type of idiomatic arguments

The syntactic features of the verbs are represented with their argument structure. A typical suggestion for the syntactic representation of idioms is the distinction between internal and external arguments for the idiomatic construction (Keil (1997) and Burger (2003) (following citation of Soehn (2006)), O'Grady (1998), Ifill (2002)). An internal argument is an integral part of the idiom chain [O'Grady, 1998] that cannot be elided, altered or changed without the loss of the idiomatic meaning. External for the idiomatic construction arguments are subcategorized by the verb and can vary according to the context. A great part of the selected verbal idioms can be organized in classes with similar patterns.

- Verb-subject semi-idioms. The subject argument of the verbal component is part of the idiomatic chain. For example the expression - *кракът ми не е стъпвал* literally "*my food didn't step here*" meaning *'I have never been there'*
- Verb - object semi-idioms. The semi-idiom consists of a transitive or intransitive verb or of a closed set of verbs that take a specific idiomatic object such as *давам ухо,* literally *"give ear,* meaning *'eavesdrop'.* As pointed above in 4.1.4.1. A), when the object of this type of semi-idioms is represented only from a noun, the construction is characterized with a considerable fixedness with a view to adjective modifications and word order. When the internal for the construction verbal object is represented from a noun phrase, most of the semi-idioms in this group present a large degree of variability, especially in terms of their syntax. They allow variable elements (*хвърлям **някого** на лъвовете - throw **someone** to the lions*), and optional ones (*вкарвам в **правия** път - put in the **right** way*). Syntactically, such idioms correspond to verb phrases, with a fixed direct object argument and an open indirect object argument *(давам нещо на някого (give something to someone)* → *хвърлям боб за... (*literally *"throw beans for...")*. This verb phrases are completely regular in their syntactic behavior. In particular, they undergo syntactic operations such as adverbial

modification, passive, etc.

- Verb – adjunct semi-idioms. The fixed idiomatic component is prepositional phrase functioning as adjunct in the non - idiomatic reading of the expression, but obligatory for the idiomatic realization of the verb. The semi-idioms in this group allow modifications in some cases as in 4.1.4.2. A), but are restricted for some syntactic transformations as passivization. For example for the expression *тръгва по мед и масло* literally "*it goes over milk and honey*" meaning *'things take the right direction'* the passivization is meaningless - ? *по мед и масло е тръгнато.*

An increase or decrease in the number of arguments in the semi-idiomatic expressions is encountered in comparison with the corresponding non-idiomatic verb. Often an idiomatic verb has the same number of arguments as its non-idiomatic counterpart.

4.2.2. Syntactic transformations

According to Cruse (1997) and Melchuk (1995) some of the restrictions on the syntactic variations of idioms are semantically motivated. The morpho-syntactic and the semantic behavior of semi-idioms are relevant for determining whether an idiom is syntactically mobile. The verbal part of the idiom must be able to undergo passivization under non-idiomatic circumstances. As we saw in 4.2.1 the modifications and transformations of semi-idioms are defined from the inner argument structure of the whole idiomatic construction. Syntactic modifications are possible when the internal argument position of the verb is represented in the composition. The syntax of the non-idiomatic correspondence expression maps the syntactic of the idiomatic phrase.

Except for internal argument structure, another criterion for defining the possibility of syntactic transformations is the meaning. The most popular example with the idiom *kick the bucket* illustrates this phenomena, its synonym idioms also cannot passivize and moreover its correspondence in Bulgarian *ритвам камбаната* and the synonymous *гушвам букет, хвърлям топа* etc. also cannot passivize, because they express the same meaning – die, that cannot take direct object.

4.3. External. Semantic substitution of idiom components by other words or phrases

The represented dependencies on the form and meaning of semi-idioms pose the main problematic fields in the incorporation of a semi-idiom in wordnet structure, the synonymy of idioms, the synonymy of idioms and non-idiomatic literals, the definition of idiomatic hyperonyms etc. Considering the proposed formal properties of idiomatic expressions we outline the following subtypes of idiomatic literals in the BulNet structure:

- Idiomatic literals whose hyperonym corresponds to the head verb of the semi-idiom. For example the semi-idiom *изваждам кирливите ризи* literally "*take out the dirty shirts*", meaning '*reveal unpleasant secrets*' is hyponim of the verb *разкривам* (*reveal*), which is synonymous of the head element - *take out.*
- Idiomatic literals with non-transparent internal argument structure, whose hyperonym corresponds to the meaning of the whole semi-idiom. For example the semi-idiom *дим да ме няма* literally "*as smoke I'm not here*", meaning '*disappear*' is hyponym of the verb *изчезвам* (disappear).
- Idiomatic literals, incorporated in one synset and considered for synonyms. There are only few examples in wordnet, and in language at all, of instances like synonymous *ритвам камбаната, гушвам букета, хвърлям топа,* corresponding to *kick the bucket,* that has nothing common in structure, but express the same meaning. We incorporate as synonymous also the described above in 4.1.5. semi-idiomatic constructions with varying and optional components.
- Idiomatic literal, incorporated as synonym of non-idiomatic literal in the same synset.

The place of the idiom in the structure of semantic database also could be used as criterion for description the formal structure of semi-idiom. The precise definition of the place of an idiom in the WN structure requires subcategorization analysis of verbal idioms involving the study of the argument structure and syntactic constraints of semantically related idiomatic items (hyperonyms and hyponyms), as well as of idioms and their non-idiomatic synonyms, the examination of the different syntactic behavior of structurally and semantically similar idioms, etc.

5. Conclusions:

In this paper we described a possible architecture for the formal description of a particular type of MWEs. The formal description of idiomatic verb forms in Bulgarian and the level of their regularity in context as proposed here will allow for the automatic recognition of such forms. This makes it possible for a computer system to search for and identify the idiomatic form as a whole. A general characterization of the idiosyncratic morpho-syntactic features of idiomatic expressions is necessary in order to identify which idiomatic properties should be captured by the identification models for the purposes of WSD and MT. The combination of formal description of properties of idiomatic expressions and their incorporation of in the WordNet structure could resolve some of the problems, coming from the structural mismatches between languages in a MT system.

6. Future directions

In order to elaborate a module for recognition and analysis of idiomatic expressions for the purposes of WSD and MT a wide coverage dictionary of Bulgarian idioms is needed. The dictionary of semi-idioms and their corresponding inflection types should be enlarged and applied to large-sized corpora. The syntactic analysis of semi – idioms should be improved and the descriptions of their paradigmatic and syntagmatic features, structural patterns and types should be applied in the treating newly emerging idiomatic expressions and in order to investigate new entries. The examination of the syntactic behavior of structurally and semantically similar semi-idioms and subcategorization analysis of verbal idioms involves the description of the argument structure, the external selection restrictions and combinatorial possibilities in respect to both semantic roles and the syntactic form and function.

References:

1. [Almind et. all, 2006] Almind R., Bergenholtz H., Vrang V. *Theoretical and Computational Solutions for Phaseological Lexicography* – Linguistik online 27, 2/06

2. [Arnold et. all,] Doug Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys, Louisa Sadler *MACHINE TRANSLATION: An Introductory Guide* - http://www.essex.ac.uk/linguistics/clmt/MTbook/HTML/book.html

3. [Baldwin 2006] Baldwin T. *Compositionality* and Multiword Expressions: Six of One, Half a Dozen of the Other? ACL workshop on MWE 2006. - www.inf.ufrgs.br/~avillavicencio/mwe-papers/author-slides/baldwin.pdf

4. [Bauer et all. 2004.] Bauer L, Grant L. *Criteria for Re-defining Idioms:* Are we Barking up the Wrong Tree? Applied Linguistics 25/1: 38-61.

5. [Burger, 2003] Burger, Harald. *Phraseologie. Eine Einfuhrung am Beispiel des Deutschen*. ed. Berlin. (Grundlagen der Germanistik 36).

6. [Cholakova, 1968] Kristalina Cholakova. *Za leksiko-gramatichnata ekvivalentnost ne fazeologichnata edinica s dumata.* Slavistichen sbornik. Sofia

7. [Cruse, 1997] Cruse D. A. *Lexical Semantics*. Cambridge University Press, Cambridge.

8. [Copestake 2002] Copestake A., Lambeau F., Villavicencio A., Bond F, Baldwin T., Sag I., and Flickinger D. *Multiword expressions: Linguistic precision and reusability*. Proceedings of the 3-rd International Conferenceon Language Resources and Evaluation (LREC 2002), 1941–7

9. [Erbach, 1992] Gregor Erbach. *Head-Driven Lexical Representation of Idioms in HPSG*. Proceedings of International Conference of Idioms, Tilburg (NL),

10. [Ifill, 2002] Tim Ifill. *Seeking the Nature of Idioms*. Cambridge: Harvard College, 2002.

11. [Keil, 1997] Keil, Martina. *Wort fur Wort.* Representation und Verarbeitung verbaler Phraseologismen. Tubingen. (Sprache und Information; 35).

12. [Koeva et. all, 2006a] Koeva Sv., Sv. Leseva, M. Todorova*. Bulgarian Sense Tagged Corpus*. Proceedings of Strategies for Developing Machine Translation for Minority Languages, SALTMIL2006, 23 May 2006, Genoa Italy.

13. [Koeva et. all, 2006b] Koeva, Sv., Sv. Leseva, I. Stoyanova, E. Tarpomanova, M. Todorova. *Bulgarian Tagged Corpora*. Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages*,* , Sofia 2006, pp. 78-86

14. [Mel'cuk 1995] Mel'cuk I. 1995. *Phrasemes in language and phraseology in linguistics.* Idioms: Structural and Psychological Perspectives, chapter 8. Lawrence Erlbaum Associates

15. [Moore R. 2003], Richard Moore. *Learning Translations of Named-Entity Phrases from Parallel Corpora*: EACL 2003: 259-266

16. Nicheva K. 1987. *Bulgarian Phraseology.* Sofia

17. [Nunberg et all., 1983] Wasow, T., Sag, I., Nunberg, G. 1983. *Idioms: an interim report.* Proceedings of the XIIIth International Congress of Linguistics, pages 102–115, 1983.

18. Poulsen S. 2005. Collocations as a language resource. A functional and cognitive study in English phraseology. PHD dissertation, Institute of Language and Communication University of Southern Denmark

19. Polguere A. 2000. A "Natural" Lexicalization Model for Language Generation. Proceedings of the Fourth Symposium on Natural Language Processing 2000 (SNLP'2000). Chiangmai, Thailand, 37-50.

20. [Sailer 2003] Manfred Sailer. Combinatorial Semantics and Idiomatic Expressions in Head-Driven Phrase Structure Grammar. Philosophische Dissertation angenommen von der Neuphilologischen Fakultät der Universität Tübingen.

21. Silberzstein Max, 2007 *Nooj Manual*

22. [Soehn, 2006] Jan-Philipp Soehn (Tübingen) On Idiom Parts and their Contexts, Linguistik online 27, 2/06

23. [Stathi 2006] Katerina Stathi. Corpus Linguistics meets Cognitive Linguistics: A framework for the analysis of idioms In: Cognitive-linguistic approaches: What can we gain by computational treatment of data? A theme session at DGKL-06/GCLA-06 (Meeting of the German Cognitive Linguistics Association), Munich, Germany.

24. [Todorova, 2006] Todorova M. *On The Classification of Bulgarian Non-Free Phrases.* Proceedings of the Fifth International Conference Formal Approaches to South Slavic and Balkan Languages, 18-20 October 2006, Sofia, Bulgaria, pp. 251-256.

25. [O'Grady, 1998] William O'Grady. *The Syntax of Idioms.* In Natural Language and LinguisticTheory 16: 279-312.

26. [Van der Linden, 1991] Erik-Jan van der Linden. *Idioms, Non-literal Language and Knowledge Representation*. Proceedings of the IJCAI Workshop on Computational Approaches to Non-Literal Language: Metaphor, Metonymy, Idiom, Speech Acts, Implicature. Sydney.

27. [Villavicencio et all] Aline Villavicencio, Ann Copestake, Benjamin Waldron, Fabre Lambeau. *Lexical Encoding of MWEs*