Towards the Automatic Identification of Light Verb Constructions in Bulgarian

Ivelina Stoyanova, Svetlozara Leseva, Maria Todorova
Department of Computational Linguistics
Bulgarian Academy of Sciences
{iva, zarka, maria}@dcl.bas.bg

Abstract

This paper presents work in progress focused on developing a method for automatic identification of light verb constructions (LVCs) as a subclass of Bulgarian verbal MWEs. The method is based on machine learning and is trained on a set of LVCs extracted from the Bulgarian WordNet (BulNet) and the Bulgarian National Corpus (BulNC). The machine learning uses lexical, morphosyntactic, syntactic and semantic features of LVCs.

We trained and tested two separate classifiers using the Java package Weka and two learning decision tree algorithms – J48 and RandomTree. The evaluation of the method includes 10-fold cross-validation on the training data from BulNet ($F_1 = 0.766$ obtained by the J48 decision tree algorithm and $F_1 = 0.725$ by the RandomTree algorithm), as well as evaluation of the performance on new instances from the BulNC ($F_1 = 0.802$ by J48 and $F_1 = 0.607$ by the RandomTree algorithm). Preliminary filtering of the candidates gives a slight improvement ($F_1 = 0.802$ by J48 and $F_1 = 0.737$ by RandomTree).

1. Introduction

Multiword expressions (MWEs) have been estimated to represent a substantial portion of the lexical system of a language. For example, it has been reported that 41% of the literals of the Princeton WordNet 1.7 are MWEs (Sag et al., 2002). Other scholars propose that multiword expressions are quantitatively equivalent to simple words (Jackendoff, 1997) or even that the number of MWEs is much more prevalent than the number of single words (Melčuk, 1998). This makes the systematic description of MWEs a very important task which influences largely the performance of the applications in the field of information extraction, text summarisation, machine translation and other NLP areas.

The work presented here focuses on LVCs as a subclass of Bulgarian verbal MWEs with a view to their automatic recognition and annotation. LVCs consist of a verbal component and a complement that function as a semantic whole. Like nominal predicates, the LVCs' complement carries the predicative meaning of the MWE.

After overviewing the related work in the field (Section 2.), we discuss the specific properties of LVCs (Section 3.) with a focus on Bulgarian as a morphologically rich language with free word order. Section 4. presents a method for automatic identification of LVCs which relies on machine learning and uses as features various lexical, semantic, morphosyntactic, syntactic, statistical, and derivational properties of LVCs and their components. The evaluation of the method includes 10-fold cross-validation on training data compiled from the Bulgarian WordNet (BulNet) and the Bulgarian National Corpus (BulNC), as well as evaluation of the performance on new examples from the BulNC. We conclude the paper with a discussion of the results.

2. Related Work

Recent research into MWEs focuses on verbal and other MWEs and the description of their components and structure (Villavicencio et al., 2004; Gregoire, 2010; Francopoulo, 2013; Gralinski et al., 2010). Nonetheless, the challenges to the lexicographic description of verbal MWEs posed by morphologically rich languages have not been completely addressed yet. These include: rich inventories of synthetic and analytical verb forms with a complex word order, flexible word order of the components of the verbal MWEs, structural features, such as mandatory and optional components, the possibility of having discontinuous components with intervening external elements, etc.

The description of MWEs, such as the ones proposed by Nunberg et al. (1994), Sag et al. (2002), Baldwin et al. (2003), among others, deal with the restrictions imposed on the internal structure, syntactic behaviour and semantic properties of MWEs, which affect significantly their linguistic annotation and automatic processing. LVCs require special treatment in natural language processing as part of the group of verbal MWEs since, unlike free phrases, their meaning is not fully decomposable to the meanings of their components, and they are often not translated to other languages literally. LVCs also need to be distinguished from idioms since they have greater syntactic flexibility and are more semantically predictable as compared with idioms.

Mostly, research on LVCs is focused on either a limited number of light verbs, e.g. candidates containing the verbs *make* and *take* (Stevenson et al., 2004), or on certain syntactic subtypes, e.g. verb–noun (Fazly and Stevenson, 2007) or verb–preposition–noun combinations (de Cruys and Moirón, 2007).

A variety of methods for the identification of LVCs have been reported: semantic (de Cruys and Moirón, 2007), statistical (Gurrutxaga and Alegria, 2011), rule-based (Vincze et al., 2011), or hybrid methods (Tan et al., 2006). Some methods are focused on the alignment of LVCs in parallel corpora (Samardžić and Merlo, 2010).

Tu and Roth (2011) propose a supervised system that applies machine learning on a manually annotated corpus of positive and negative examples of LVCs with the six most frequent English light verbs: *do*, *get*, *give*, *have*, *make* and *take*. They train two systems – one on statistical and one on contextual features. Their findings show that both show similar results in general, however the one trained with contextual features is more accurate and robust with respect to identical surface structures that may or may not be LVCs, e.g. *have a look*.

Nagy et al. (2013) aim at a full coverage of LVCs and propose a two-step method by which they first identify potential LVC candidates in running texts and then use a machine learning-based classifier to select LVCs from the candidates. The selection of LVCs is based on feature templates like semantic or morphological features of LVCs in context. Their approach distinguishes cases where the phrase does not function as an LVC from true LVCs.

Linguistic knowledge is crucial to the work on LVC recognition and researchers have made use of various linguistic features – morphological, including derivational features, lexical, syntactic, semantic features. The studies on LVCs uniformly make use of the surface syntactic structure typical of LVCs, predominantly V NP and V PP constructions. The light verbs are usually specified in advance and are restricted to a well-defined small set (Stevenson et al., 2004; Tu and Roth, 2011). The opposite approach – accounting for a broader range of light verbs, as presented in the data – is proposed by Nagy et al. (2013). This latter approach is also adopted in our study.

Syntactic information is also employed, to make sure that a potential LVC represents a syntactic unit. Depending on the adopted syntactic framework and linguistic resource, researchers use combinations that represent particular dependency relations (Nagy et al., 2013; Chen et al., 2015) or combinations of constituents (Tu and Roth, 2011).

The works on LVCs typically employ particular restrictions on the semantics of the nominal component. Tu and Roth (2011) and Nagy et al. (2013) make use of the tendency of LVCs to correspond with derivationally related verbs, e.g. *pravya poseshtenie – poseshtavam* (*pay a visit – to visit*). Nagy et al. (2013) and Chen et al. (2015) also use the semantics of the nominal component by looking at its WordNet hypernyms, such as *activity* or *event*. Stevenson et al. (2004), who look for verbs instead of for event nouns (thus capturing those LVCs in which the nominal component is a deverbal noun coinciding in form with the verb), use a selection of Levin's verb classes as the complements of light verbs. Levin's classes are also adopted by Tu and Roth (2011). Chen et al. (2015) employ diverse semantic information from different resources, including hypernyms from WordNet, the WordNet noun types (the type is the semantic primitive, such as *person, animal, artifact, change, state*, etc. that is assigned to each noun and verb synset), as well as richer semantic descriptions, such as WordNet senses, word senses from OntoNotes (Pradhan et al., 2007) and information from PropBank (Palmer et al., 2005).

Within the context of Slavic languages, work on LVCs has been reported for Russian (Mudraya et al.,), Serbian (Samardžić, 2008; Samardžić and Merlo, 2010), Croatian (Gradečak-Erdeljić and Brdar, 2012), Czech (Urešová et al., 2016) and Polish (Przybyszewski, 2015), mainly with a view to LVCs' annotation in treebanks and multilingual parallel corpora or to the theoretical description of their compositionality and semantics. With respect to Bulgarian, LVCs have been tackled within a proposal made by Koeva (2006) of a framework for morphosyntactic description of MWEs, which has subsequently been partially incorporated in the construction of a large dictionary of Bulgarian MWEs enriched with diverse morphological, syntactic and structural information (Koeva et al., 2016).

Recent work on LVCs has been undertaken for many languages, Slavic languages and Bulgarian in particular, within the PARSEME Shared Task on automatic detection of verbal Multi-Word Expressions[1]. The Shared Task aims at the automatic identification of verbal MWEs in running texts. This paper is part of the work of the Bulgarian team on the Shared Task.

3. Properties of LVCs

LVCs are usually treated as constructions involving complex predication that takes place between a verb and another predicative element (Butt, 2003; Goldberg, 2003; Jackendoff, 1974; Wittenberg et al., 2014), among others. The verb belongs to a relatively small set of verbs whose meaning is more or less abstract ('semantically bleached') and mainly express aspect, directionality or aktionsart of the predicate (Butt, 2003; Wittenberg et al., 2014). The semantic properties of light verbs correspond to the fact that they have high frequency, and, as we have observed in the data, generally exhibit high polysemy.

The structure of LVCs varies according to the light verb's formal complement. It is usually a noun phrase that corresponds to the direct object position of the verb – *vzemam reshenie* (*make a decision*) or a PP corresponding to an indirect object – *vlizam v kontakt* (*come into contact*). The light verb's complement may also be an adjective – *pravya lud* (*make crazy*) or an adverb *vzemam predvid* (*take into consideration*).

The LVC's complement is in fact the semantic predicate and contributes the major part of the complex meaning of the expression. As a rule, it is an abstract entity with eventive or similar semantics and is frequently expressed by a deverbal predicative noun, although as noted by Cinkovà and Kolářová (2005), deadjectival nouns such as *vazmozhnost* (*possibility*) also occur. It has been proposed that the complements' sense is non-figurative (Vincze et al., 2016).

As frequently noted in the literature, an important trait of LVCs, which follows from the eventive nature of the complement, is that the construction often has a corresponding synonymous single verb derivationally related to the eventive noun. The verb–noun relation is usually through suffixation *resha/V – reshenie/N* (*decide – decision*) or through zero derivation *dokladvam/V – doklad/N* (*a report – to report*). This trait has often been employed as an additional diagnostic for LVCs although its large coverage over the data is usually taken for granted. At least judging from our data for Bulgarian, we may conclude that it is not very large. We have found out that only 265 of the 621 LVCs found in the Bulgarian WordNet have a single verb counterpart. For instance, the substitution may not be possible due to lexical gaps, e.g. *slagam kray (na)* (*put an end to*) does not have a corresponding verb that is derivationally related to the noun *kray* (*end*). On the other hand, it does not automatically follow from the substitutability with a single verb that an MWE is an LVC. The 265 LVCs in BulNet that have a corresponding single verb make up half of the 541 BulNet MWEs with a single verb correspondence. Nevertheless, as derivation is relatively easy to identify, this feature may be successfully employed in conjunction with more decisive ones.

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

Another characteristic of LVCs is the possibility to refer to the same event by using the nominal alone, e.g. *He had a walk in the park* vs. *His walk in the park* (Vincze et al., 2016). This trait was used as a diagnostic in the manual categorisation of MWEs in the Bulgarian WordNet as either being LVCs or non-LVCs, as well as in the inspection of LVC candidates extracted from the Bulgarian National Corpus which took place in the process of compiling the training and the test data.

The nominal complement in V NP LVCs tends to be able to take a plural and/or a definite form, e.g. *vzemam reshenie* (*make a decision*) may be found as *vzemam reshenieto* (sg. def.), *vzemam resheniya* (pl. indef.) *vzemam resheniyata* (pl. def.) although this is not always the case – e.g. *vzemam uchastie take part* does not allow free variation of the noun. Still, this tendency of LVCs may serve in addition to other diagnostics to distinguish LVCs from idioms with the same surface structure.

Another specific feature of the LVC complement is that it may be modified, e.g. *vzemam vazhno reshenie* (*make an important decision*), *vzemam deyno uchastie* (*take an active part*). This is another linguistic trait that distinguishes LVCs from some other types of MWEs, idioms in particular, which may allow only very limited modification (practically a lexical variant of the idiom), e.g. *vdigam letvata* (*raise the bar*) and its variant *vdigam letvata visoko* (*raise the bar high*).

The components of an LVC may also be separated by other elements, such as adjuncts of the entire LVC, e.g. *vzemam* ***barzo*** *reshenieto* (*make* ***quickly*** *the decision*), or elements that are external with respect to the LVC. Among the latter are the question particle *li* and pronominal clitics, e.g. *Vzeha* ***li*** *reshenieto?* (*Did they make the decision?*) and *Napraviha* ***mu*** *operatsiya.* (*They made* ***him*** *an operation*). More than one external element may be found *Napraviha* ***li mu*** *operatsiya?* (*Did they make* ***him*** *an operation?*) as well as longer sequences. This trait is shared with free phrases and many idioms, but we point it out as it needs to be taken into account when determining the search scope for an LVC in the corpus.

4. A Method for Automatic Identification of LVCs

We developed a method for automatic recognition of LVCs in running text based on observations made on the properties of light verbs and LVCs discussed by various authors (see Section 3.), as well as some specific features that we consider relevant for Bulgarian and other morphologically rich languages. The method is implemented in Java, using the Weka library for data mining (Hall et al., 2009).

4.1. Resources

For the purposes of automatic identification of LVCs we compiled a subcorpus of the Bulgarian National Corpus (BulNC)[2] (Koeva, 2014a), containing news (35,758 texts, amounting to 10,655,068 words) and fiction texts (443 texts, a total of 6,237,024 words). The corpus was annotated using the Bulgarian Language Processing Chain (Koeva and Genov, 2011), which is available as a web service using a RESTful API. The annotation includes sentence splitting, tokenisation, POS tagging and lemmatisation.

We also used another language resource, the Bulgarian Wordnet (BulNet) (Koeva, 2014b), from which we extracted a list of 2,239 verbal MWEs (MWE synonyms in verb synsets) containing at least a verb and a noun. We determined the internal syntactic structure of each MWE by analysing its components as a sequence of POS tags and obtained the following structural types: verb – direct object (V–NP), e.g. *vzemam dush* (*take a shower*) or verb – indirect object (V–PP), e.g. *vzemam pod vnimanie* (*take into consideration*). MWEs of other syntactic types, e.g. V–AdvP, V–AP, were not taken into account. The set of MWEs selected in this way constitutes the main part of the training data for the machine learning, after being manually divided into LVCs and non-LVCs.

Further, we used BulNet to extract words that can occur as part of LVCs. First, we extracted 74 highly ambiguous verbs (verbs with 15 or more senses in BulNet). These verbs were subsequently examined and non-light verbs were filtered out. The remaining 46 verbs were merged with a list of 81 verbs that were found as the heads of those MWEs in BulNet that were manually validated as LVCs. After the duplicate entries were removed, the compiled list totaled 105 verbs. Table 1 presents the distribution of light verbs with respect to the number of senses in BulNet and their frequency in the BulNC. Only a

[2] http://search.dcl.bas.bg/

# senses	# verbs	Frequency	# verbs
<5	13	<50	4
≥ 5	68	≥ 50	77
≥ 10	42	≥ 100	70
≥ 20	23	≥ 500	43
≥ 50	4	≥ 1000	31

Table 1: Distribution of light verbs according to: (a) number of senses in BulNet; (b) frequency in the BulNC.

small number of verbs have less than 5 senses (13 verbs) or low frequency of less than 50 occurrences (4 verbs), and no verb has both low frequency and a small number of senses.

Next, BulNet served us to extract semantic information about the components of the LVCs. All the verb and noun synsets in the Princeton WordNet (and respectively in BulNet) are each assigned a single semantic primitive out of a list of language-independent primitives that represent the unique beginners of the separate hierarchies in WordNet (Miller, 1998) (initially organised in separate lexicographer files). We consider 10 of the noun semantic primitives, such as *noun.act*, *noun.state*, *noun.cognition*, etc., as potentially expressing predicative meaning, while excluding the remaining 15 noun primitives, such as *noun.artefact*, *noun.person*, etc.[3] The set of potential semantic primitives of all the possible senses of a given noun were used as features in the machine learning.

4.2. Machine Learning Features

For the purposes of machine learning we defined a number of features capturing the essential linguistic traits of MWEs and LVCs in particular.

1. Lexical features

   We use the verb's lemma as a feature in the machine learning, relying on the fact that certain light verbs can potentially combine with certain (classes of) nouns, e.g. *poemam* {*risk, otgovornost*} (*assume* {*risk, responsibility*}), while other combinations are limited or impossible, e.g. **vzemam* {*risk, otgovornost*} (*take* {*risk, responsibility*}).

2. Semantic features

   The semantic features include the semantic primitives of the nouns which are extracted from BulNet. As noted above, we selected 10 (of the overall 25) of the noun semantic primitives which are relevant for predicative nouns: *noun.act, noun.cognition, noun.communication, noun.event, noun.feeling, noun.motive, noun.phenomenon, noun.process, noun.relation, noun.state*. For a given ambiguous noun, all the possible labels were extracted and represented as a set. In the cases where the different senses of a noun correspond to different labels, additional procedures were performed. If a noun is associated with a semantic primitive that is not typical for predicative nouns, the primitive (and the respective sense) was excluded from the noun's description. For instance, the noun *vapros* (*question, issue*) was found in BulNet with the following primitives: {noun.act, noun.communication, noun.cognition, noun.attribute, noun.event} and the sense having the primitive {noun.attribute} was excluded. However, a noun which predominantly appears in BulNet in non-predicative senses (more than half of the senses), is taken to be non-predicative and is consequently ignored as a possible nominal component within an LVC.

3. Statistical features

   The statistical features contain information about the frequency of potential LVCs and their components in the corpus, i.e. the log-frequency (logarithm of the observed absolute frequency to the base of 2) of: (a) the verb, (b) the noun, and (c) the LVC candidate. The logarithmic transformation linearises the distribution of frequencies and allows for simpler correlation analysis with other

[3] The list of primitives is available at `https://wordnet.princeton.edu/man/lexnames.5WN.html`

features. Based on the observed frequencies we also calculated the association measure (using Mutual Information, MI) of the LVC candidate in order to determine whether it is a collocation and, potentially, an MWE.

4. Morphosyntactic features

   The morphosyntactic features account for the fact that the nominal complement of many LVCs does not occur in a single fixed form, but may take both singular and plural and/or indefinite and definite forms. Of course, there are cases in which there are restrictions on the form of the nominal complement, e.g. *pravya vpechatlenie* (*make an impression*), in which the noun is used as part of the LVC only in the singular indefinite form. Moreover, in rare cases the noun may occur with two different senses in different LVCs with the same verb, where the only difference is the form of the noun, e.g. *vzemam myarka* (sg. indef.) (*take measures*, to measure dimensions) as opposed to *vzemam merki* (pl. indef.) (*take measures, actions*). We leave the detailed analysis and handling of these cases for the future.

   Variability in components is more likely for LVCs than for idioms, that is why we introduce a binary feature which takes **true** if the noun is found in more than one form in the corpus (singular and/or plural, indefinite and/or definite, count (for masculine nouns)) and **false** if the noun is invariable.

5. Syntactic features

   The syntactic features included in the machine learning account for the following properties of LVCs:

   (a) **LVCs allow different word order.** As the relatively free word order in Bulgarian makes it possible for the complement to precede the verb in various contexts, we took into account both word order variants. The feature takes the value *true* if more than one word order is registered in the corpus and *false* otherwise.

   (b) **Components may take modifiers.** As mentioned above, the LVC components may be separated either by modifiers and adjuncts of the LVC or by external elements. For the purposes of the current study, we limited the distance between the light verb and its noun complement (or the noun complement of the PP in V PP LVCs) to be up to two tokens. Possible modifiers of the noun were limited to adjectives preceding the noun. The feature takes the value *true* if an example with a modifier is found in the corpus and *false* otherwise.

   (c) **LVCs allow external elements to occur between their components**. External elements were identified by their POS in order to generalise the cases. The feature takes the value **true** when the POS tags of the elements (found at distance of at most two tokens) are other than 'adjective' or 'preposition', and the value **false** otherwise. Adjectives are considered as possible modifiers to the noun (see (b) above), while prepositions are likely to introduce a PP component of the vMWE. Another restriction currently adopted is that the tokens that may separate the components of an LVC cannot be punctuation marks or conjunctions since these usually mark phrase or clause borders.

6. Derivational features

   We defined a derivational feature that takes into consideration the strong tendency for predicative nouns to be of deverbal stems and therefore – to be derivationally related to a verb. The feature takes the value **true** if a derivational relation is found, and the value **false** otherwise.

   In order to establish a derivational relation, we looked for a common stem between a (potential) nominal component of an LVC and any verb. The common stem was estimated empirically using the output of a stemmer implemented for this and other related tasks. The stemmer matches words which share a substring whose length is at least 70% of each word's length and longer than 4 characters. For instance, in the LVC *nanasyam vreda* (*cause damage*) the noun *vreda* (*damage*) is marked as derivationally related to the verb *vredya* (*to damage*) by matching the stem *vred-*.

4.3. Compilation of the training and the test dataset

The main part of the training dataset consists of the 2,239 V–NP and V–PP MWEs which were classified into two categories – 'LVC' and 'non-LVC' (see Section 4.1.) using automatic procedures and manual post-editing. As a result, a total of 461 MWEs were identified as LVCs and the remaining – as other types of verbal MWEs. To overcome the low number of the LVCs in the training data and the lack of non-MWE instances, we extended the training set with additional data from the BulNC. To this end we extracted verb–noun pairs with frequency of at least 10 in the corpus, which were then manually categorised into 'LVC' (true) and 'non-LVC' (false) by two independent annotators. We took into account the instances in which the annotators agreed.

In determining whether an MWE from BulNet or a candidate extracted from the BulNC is in fact an LVC, the annotators took into consideration several linguistic factors: (a) whether the verb qualifies as a light verb (i.e. is on the list of light verbs we identified); (b) whether the noun denotes an event or a similar semantic type of entity (state, property, etc.); (c) whether the noun is used in a non-figurative meaning; (d) whether the noun alone may be used to denote the same event.

For instance, using these diagnostics we conclude that the candidate *nanasyam shteti* (*cause damage*), which complies with (a)–(d), is an LVC: the verb has an abstract causative meaning with which it combines with a variety of nouns; the noun denotes an event or a result of an event; it is used in its primary literal sense; the noun can be used alone to refer to the event, as in: *Shtetite ni ne byaha kompensirani.* (*Our damages were not recompensed.*) In contrast, consider the idiom *podavam raka* (*lend a hand*) where: the verb is not semantically bleached; the noun may denote an act, but only in a figurative sense, whereas its literal sense denotes a body part. Besides, the noun cannot be used alone to refer to the event.

As a result, the training dataset compiled from BulNet and the BulNC consists of 2,623 instances, 897 of which are LVCs and the remaining are either non-MWEs or other categories of MWEs (e.g., idioms).

The test dataset comprises 200 unique candidates with frequency of at least 10 extracted from the BulNC in the same way as the additional training instances and annotated by the annotators into LVCs and non-LVCs, with equal number of both categories.

4.4. Method outline

We trained and tested two classifiers on the feature set (Section 4.2.) and the training set (Section 4.3.) using two different learning algorithms based on decision trees – J48 and RandomTree (Hall et al., 2009). The method for LVC identification is performed in the following steps:

(1) Identify LVC candidates in the corpus – the occurrences of a verb and a noun in the corpus which have at most two tokens between them (except punctuation and conjunctions), taking into account the possibility for a free word order.

(2) Filter the LVC candidates – remove candidates with low frequency in the corpus as their statistical measures are unreliable.

(3) Analyse the LVC candidates based on the occurrences of the verb–noun pairs in the corpus in order to determine the variations in their form and word order, the possible modifiers and external elements separating the LVCs components.

(4) Apply the trained classifier to classify the LVC candidates – distinguish LVCs from other categories of phrases: (a) other types of decomposable MWEs – where the verb is a content verb, or non-decomposable MWEs – idioms; and (b) collocations which are not MWEs.

4.5. Evaluation

We performed two-step evaluation: cross-validation on the training set and evaluation on new test data. Table 2 shows the results from the 10-fold cross-validation on the training set.

Algorithm	Precision	Recall	F_1
J48	0.739	0.794	0.766
RandomTree	0.710	0.741	0.725

Table 2: Comparison of the 10-fold cross-validation using different algorithms (J48 and RandomTree).

	Main method			Main method & Filtering		
Algorithm	Precision	Recall	F_1	Precision	Recall	F_1
J48	0.776	0.830	0.802	0.794	0.810	0.802
RandomTree	0.482	0.820	0.607	0.684	0.800	0.737

Table 3: Results from the application of the method on the test dataset of LVC candidates.

Table 3 presents the results from the application of the method on the test dataset of 200 unique LVC candidates extracted from the BulNC. The evaluation is lemma-based and each candidate is counted once (and not with its frequency in the corpus). The table provides a comparison between the main method with two different decision tree algorithms (J48 and RandomTree) and the main method supplemented with filtering of LVC candidates. The filtering included: excluding candidates with low association measure below the threshold of 2.0 (which are unlikely to be MWEs); and excluding candidates with verbs that are not light verbs (which are not in the list of 105 verbs, see Section 4.1.) and/or nouns that do not belong to the predicative categories (as defined by the semantic primitives). Performing filtering prior to machine learning ensured that a large number of improbable LVC candidates were excluded before the application of the ML method which does not perform well for low frequency candidates due to their unreliable statistical measures.

5. Discussion

The results reported in this paper are comparable to the performance of similar methods for other languages, such as the one developed by Nagy et al. (2013), while outperforming others which do not take into account semantic features, such as the method reported by Vincze et al. (2011). This emphasises the importance of semantic features such as the semantic primitive of the noun. Experiments with reducing the group of predicative noun primitives to only *noun.act* and *noun.event* show that these are the most significant primitives and although the recall falls (0.790 with J48), the precision improves (0.782 with J48).

As a large proportion of the training data were extracted from BulNet (a lexical database), they do not cover all types of MWEs, and LVCs in particular, in terms of usage variety. One of the most important results at this stage is that we obtained a reliable set of Bulgarian LVCs extracted (semi-)automatically from different language resources, using linguistic heuristics. The list of light verbs we compiled is more comprehensive than the usually adopted lists and reflects the diversity and productivity of LVCs. Moreover, the training set was extended to include LVCs from the BulNC (from unrestricted texts), which significantly improved the results (compared to $F_1 = 0.494$ trained purely on instances from BulNet and using J48). This is expected since the data from the corpus reflect the usage of LVCs while BulNet also includes rare and untypical LVCs which have low frequency in the corpus and hence – yield unreliable statistical measures. The inclusion of more real-life examples is expected to improve further the performance of the method.

Although LVCs fall into a small and clear-cut set of syntactic structures, they also are syntactically flexible as they allow intervening elements, as well as various transformations such as passivisation, nominalisation, etc., which makes their discovery in unrestricted text much more challenging. The results reported in existing literature and in this paper show that although LVCs seem to be a relatively well-defined class, their semantic traits are not specific enough to distinguish them with high precision from free phrases, collocations and idioms. These facts point to the necessity to include more contextual and semantic features and to use the LVCs' traits in a more productive way in engineering the ML features.

References

Baldwin, T., Bannard, C., Tanaka, T., and Widdows, D. (2003). An Empirical Model of Multiword Expression Decomposability. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. ACL.

Butt, M. (2003). The Light Verb Jungle. In *Harvard Working Paper in Linguistics*, volume 9, pages 1–49. John Benjamins.

Chen, W., Bonial, C., and Palmer, M. (2015). English Light Verb Construction Identification Using Lexical Knowledge. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA*, pages 2368–2374. AAAI Press.

Cinkovà, S. and Kolářová, V. (2005). Nouns as Components of Support Verb Constructions in the Prague Dependency Treebank. In Šimkovà, M., Ed., *Insight into Slovak and Czech Corpus Linguistics*, pages 113–139. Veda.

de Cruys, T. V. and Moirón, B. V. (2007). Semantics-based Multiword Expression Extraction. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions, Prague, June 2007*, pages 25–32. ACL.

Fazly, A. and Stevenson, S. (2007). Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In *Proceedings of the Workshop on MWEs, Prague, Czech Republic, June, 2007*, pages 9–16. ACL.

Francopoulo, G. (2013). *Lexical Markup Framework*. John Wiley and Sons.

Goldberg, A. (2003). Words by Default: The Persian Complex Predicate Construction. In Francis, E. J. and Michaelis, L. A., Eds., *Mismatch: Form-Function Incongruity and the Architecture of Grammar*, volume 22, pages 117–146. Stanford: CSLI Publications.

Gradečak-Erdeljić, T. and Brdar, M. (2012). Constructional Meaning of Verbo–nominal Constructions in English and Croatian. *Suvremena lingvistika*, 38.

Gralinski, F., Savary, A., Czerepowicka, M., and Makowiecki, F. (2010). Computational Lexicography of Multi-Word Units. How Efficient Can It Be? In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications. Coling, August 2010*, pages 2–10.

Gregoire, N. (2010). DuELME: A Dutch Electronic Lexicon of Multiword Expressions. *Language Resources and Evaluation*, 44:23–39.

Gurrutxaga, A. and Alegria, I. (2011). Automatic Extraction of NV Expressions in Basque: Basic Issues on Cooccurrence Techniques. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland, Oregon, USA*, pages 2–7. ACL.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA Data Mining Software: An Update. In *SIGKDD Explorations*, volume 11.

Jackendoff, R. (1974). A Deep Structure Projection Rule. *Linguistic Inquiry*, 5(4):481–505.

Jackendoff, R. (1997). The Architecture of the Language Faculty. *Computational Linguistics*, 24.

Koeva, S. and Genov, A. (2011). Bulgarian Language Processing Chain. In *Proceeding to The Integration of multilingual resources and tools in Web applications Workshop in conjunction with GSCL 2011*. University of Hamburg.

Koeva, S., Stoyanova, I., Todorova, M., and Leseva, S. (2016). Semi-automatic compilation of the dictionary of bulgarian multiword expressions. In *Proceedings of the Workshop on Lexicographic Resources for Human Language Technology (GLOBALEX 2016), Portorož, Slovenia, 24 May 2016*, pages 86–95.

Koeva, S. (2006). Inflection Morphology of Bulgarian Multiword Expressions. In *Computer Applications in Slavic Studies*, pages 201–216. Boyan Penev Publishing House.

Koeva, S. (2014a). The Bulgarian National Corpus in the context of World Theory and Practice (Balgarskiyat natsionalen korpus v konteksta na svetovnata teoriya i praktika). In Koeva, S., Ed., *Language Resources and Technologies for Bulgarian (Ezikovi resursi i tehnologii za balgarski)*, pages 29–52. Marin Drinov Academic Publishing House.

Koeva, S. (2014b). WordNet and BulNet (Wordnet i BulNet). In Koeva, S., Ed., *Language Resources and Technologies for Bulgarian (Ezikovi resursi i tehnologii za balgarski)*, pages 154–173. Marin Drinov Academic Publishing House.

Melčuk, I. (1998). Collocations and Lexical Functions. In Cowie, P., Ed., *Phraseology. Theory, Analysis, and Applications*, pages 23–53. Oxford: Clarendon Press.

Miller, G. (1998). Nouns in WordNet. In *WordNet: An Electronic Lexical Database*, pages 24–45. MIT Press.

Mudraya, O., Piao, S. S., Rayson, P., Sharoff, S., Babych, B., and L L.).

Nagy, I., Vincze, V., and Farkas, R. (2013). Full-coverage Identification of English Light Verb Constructions. In *Proceedings of the International Joint Conference on Natural Language Processing, Nagoya, Japan, 14-18 October 2013*, pages 329–337. University of Hamburg.

Nunberg, G., Sag, I., and Wasow, T. (1994). Idioms. In Everson, S., Ed., *Language*, pages 491–538. Cambridge University Press.

Palmer, M., Guildea, D., and Kingsbury, P. (2005). The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–105.

Pradhan, S. S., Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2007). Ontonotes: a unified relational semantic representation. In *First IEEE International Conference on Semantic Computing (ICSC-07), Irvine, CA*, pages 517–526. IEEE.

Przybyszewski, S. (2015). Some Problems with the Description of Paradigms of Polish Verbal Multiword Units. In E. Gutierrez Rubio, M. Falkowska, E. K. M. S. W., Ed., *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)*, volume 18, pages 213–223.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. 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*, CICLing '02, pages 1–15. Springer-Verlag.

Samardžić, T. and Merlo, P. (2010). Cross-lingual Variation of Light Verb Constructions: Using Parallel Corpora and Automatic Alignment for Linguistic Research. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground, Uppsala, Sweden*, pages 52–60. ACL.

Samardžić, T. (2008). Light verb constructions In English Language and Literature Studies. In *Structures across Cultures*, pages 59–73. Faculty of Philology, Belgrade.

Stevenson, S., Fazly, A., and North, R. (2004). Statistical Measures of the Semi-Productivity of Light Verb Constructions. In *Proceedings of the Workshop on MWEs, Barcelona, Spain, July, 2004*, pages 1–8. ACL.

Tan, Y. F., Kan, M.-Y., and Cui, H. (2006). Extending Corpus-based Identification of Light Verb Constructions Using a Supervised Learning Framework. In *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts, Trento, Italy*, pages 49–56. ACL.

Tu, Y. and Roth, D. (2011). Learning English Light Verb Constructions: Contextual or Statistical. In *Proceedings of MWE 2011, Portland, Oregon, USA*, pages 31–39. ACL.

Urešová, Z., Bejček, E., and Hajič, J. (2016). Inherently Pronominal Verbs in Czech: Description and Conversion Based on Treebank. pages 78–83.

Villavicencio, A., Copestake, A., Waldron, B., and Lambeau, F. (2004). The Lexical Encoding of MWEs. In Tanaka, T., A. Villavicencio, F. B., and Korhonen, A., Eds., *Proceedings of the ACL 2004 workshop on multiword expressions: Integrating processing. Barcelona, Spain*, pages 80–87.

Vincze, V., Nagy, I., and Berend, G. (2011). Detecting Noun Compounds and Light Verb Constructions: A Contrastive Study. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World (MWE 2011), Portland, Oregon, USA*, pages 116–121. ACL.

Vincze, V., Savary, A., Candito, M., and Ramisch, C. (2016). *Annotation Guidelines for the PARSEME Shared Task on Automatic Detection of Verbal Multiword Expressions. Version 5.0.* `http://typo.uni-konstanz.de/parseme/images/shared-task/guidelines/PARSEME-ST-annotation-guidelines-v5.pdf`.

Wittenberg, E., Jackendoff, R., Kuperberg, G., Paczynski, M., Snedeker, J., and Wiese, H. (2014). The Mental Representation and Processing of Light Verbs. In Bachrach, A., Roy, I., and Stockall, L., Eds., *Structuring the Argument. Multidisciplinary Research on Verb Argument Structure*, pages 61–80. John Benjamins.