Dependency Parsing for Identifying Hungarian Light Verb Constructions

Veronika Vincze[1,2], **János Zsibrita**[2] and **István Nagy T.**[2]
[1]Hungarian Academy of Sciences, Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu
[2]Department of Informatics, University of Szeged
{zsibrita,nistvan}@inf.u-szeged.hu

Abstract

Light verb constructions (LVCs) are verb and noun combinations in which the verb has lost its meaning to some degree and the noun is used in one of its original senses. They often share their syntactic pattern with other constructions (e.g. verb-object pairs) thus LVC detection can be viewed as classifying certain syntactic patterns as light verb constructions or not. In this paper, we explore a novel way to detect LVCs in texts: we apply a dependency parser to carry out the task. We present our experiments on a Hungarian treebank, which has been manually annotated for dependency relations and light verb constructions. Our results outperformed those achieved by state-of-the-art techniques for Hungarian LVC detection, especially due to the high precision and the treatment of long-distance dependencies.

1 Introduction

Multiword expressions (MWEs) are lexical items that can be decomposed into single words and display lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Kim, 2008). Light verb constructions (LVCs) form a subtype of MWEs: they are verb and noun combinations in which the verb has lost its meaning to some degree and the noun is used in one of its original senses (e.g. *make a decision* or *take a walk*). In several NLP applications like information retrieval or machine translation it is important to identify LVCs in context since they require special treatment, particularly because of their semantic features. Thus, LVCs should be identified to help these applications.

Light verb constructions (e.g. *make a mistake*) often share their syntactic pattern with literal verb + noun combinations (e.g. *make a cake*). Thus, specific syntactic constructions – e.g. verb-object pairs – can be separated into two classes: one where the noun behaves as a real object (*cake*) and one where the noun functions as the light verb object (*mistake*). Thus, LVC detection can be viewed as classifying certain syntactic patterns as LVCs or not and assigning a specific syntactic label to the argument of the light verb.

In this paper, we explore a novel way to LVC detection: we apply a dependency parser to carry out the task. Although the usability of identified multiword expressions has been investigated in the literature (see Section 4), and many MWE detection systems rely on syntactic information, we are not aware of any approach that aimed at applying a dependency parser for the dedicated task of identifying LVCs. Our approach requires a treebank annotated for syntactic and LVC information at the same time. Due to the availability of annotated resources, we focus on light verb constructions in Hungarian, a morphologically rich language. Thus, we present our experiments on the legal subcorpus of the Szeged Dependency Treebank annotated for LVCs (Vincze and Csirik, 2010) as well as dependency relations (Vincze et al., 2010). We will pay special attention to non-contiguous LVCs in our investigations as there are quite a few non-contiguous LVCs in Hungarian due to the free word order. Our results empirically prove that LVCs can be detected as a "side effect" of dependency parsing.

2 Light Verb Constructions in Hungarian

Hungarian is an agglutinative language, which means that a word can have hundreds of word forms due to inflectional or derivational morphology (É. Kiss, 2002). Hungarian word order is related to information structure, e.g. new (or emphatic) information (focus) always precedes the verb and old information (topic) precedes the

International Joint Conference on Natural Language Processing, pages 207–215,
Nagoya, Japan, 14-18 October 2013.

focus position. Thus, the position relative to the verb has no predictive force as regards the syntactic function of the given argument. In English, the noun phrase before the verb is most typically the subject whereas in Hungarian, it is the focus of the sentence, which itself can be the subject, object or any other argument.

The grammatical function of words is determined by case suffixes. Hungarian nouns can have about 20 cases, which mark the relationship between the verb and its arguments (subject, object, dative etc.) and adjuncts (mostly adverbial modifiers). Although there are postpositions in Hungarian, case suffixes can also express relations that are expressed by prepositions in English. Verbs are inflected for person and number and the definiteness of the object. There are several other linguistic phenomena that are syntactic in nature in English but they are encoded morphologically in Hungarian. For instance, causation and modality are expressed by derivational suffixes.

The canonical form of a Hungarian light verb construction is a bare noun + third person singular verb, for instance, *tanácsot ad* advice-ACC give "to give advice". Due to the above features, they may occur in non-canonical versions as well: the verb may precede the noun, or they may be not adjacent, moreover, the verb may occur in different surface forms inflected for tense, mood, person and number.

LVCs may occur in several forms due to their syntactic flexibility. Besides the prototypical verbal form in Hungarian, they can have a participial form (e.g. *figyelembe vevő* account-INE taking "taking into account") and they may also undergo nominalization, yielding a nominal compound (e.g. *életbe lépés* life-INE step "entering into force").[1]

From a morphological perspective, LVCs can also be divided into groups. First, the nominal component is the object of the verb, i.e. it bears an accusative case in Hungarian (e.g. *döntést hoz* decision-ACC bring "to make a decision" or *tanácsot ad* advice-ACC give "to give advice"). Second, the nominal component can bear other (oblique) cases as well (e.g. *zavarba hoz* embarrassment-ILL bring "to embarrass" or *figyelemmel kísér* attention-INS follow "to pay attention"). Third, – although rarely – a postpositional phrase can also occur in the construction (e.g. *uralom alá jut* rule under get "to get under rule" or *hatás alatt áll* effect under stand "to be under effect").

3 Light Verb Constructions as Complex Predicates

Although light verb constructions are made of two parts, namely, the nominal component and the verb, thus, they show phrasal properties, it can be argued that from a semantic point of view they form one unit. First, many light verb constructions have a verbal counterpart with the same meaning (e.g. *döntést hoz* decision-ACC bring "to make a decision" – *dönt* "to decide"). Second, there are meanings that can only be expressed through a light verb construction (e.g. *házkutatást tart* (search.of.premises-ACC hold) 'to conduct search of premises' in Hungarian). Third, there are languages that abound in verb + noun constructions or multiword verbs (such as Estonian (Muischnek and Kaalep, 2010) or Persian (Mansoory and Bijankhan, 2008)): verbal concepts are mostly expressed by combining a noun with a light verb (Mansoory and Bijankhan, 2008).

On the other hand, there are views that the relationship between the verbal and the nominal component is not that of a normal argument. For instance, Meyers et al. (2004) assume that support verbs (a term related to light verbs) share their arguments with a noun. Chomsky (1981, p.37) calls *advantage* a quasi-argument of *take* in the idiom *take advantage of*.[2] Alonso Ramos (1998) proposes the role of quasi-object: this relationship holds between parts of idiomatic constructions, which is in accordance with Chomsky's usage of the term *idiom*. In this spirit, the term *quasi-argument* might be extended to signal the relationship between the verbal and the nominal components of light verb constructions as well since they behave as a semantic unit, forming one complex predicate.

Higher-level NLP applications can also profit from this solution because the identification of light verb constructions can be enhanced in this way, which has impact on e.g. information extrac-

[1] Due to some orthographical rules, certain nominal or participial occurrences of LVCs should be spelt as one word in Hungarian (such as *tanácsadó* advice.giver "consultant"). These latter cases are not identifiable with syntax-based methods, only with morphological methods, thus we omit them from our investigations.

[2] In our view, *take advantage of* is a light verb construction rather than an idiom.

tion (IE). For instance, in event extraction the parser should recognize the special status of the quasi-argument and treat it in a specific way as in the following sentence:

Pete **made a decision** on his future.

Thus, the following data can be yielded by the IE algorithm:

EVENT: decision-making
ARGUMENT$_1$: Pete
ARGUMENT$_2$: his future

Instead of:

*EVENT: making
ARGUMENT$_1$: Pete
ARGUMENT$_2$: decision
ARGUMENT$_3$: his future

Thus, there is an event of **decision-making**, **Pete** is its subject and it is about **his future** (and not an event of **making** with the arguments **decision**, **Pete** and **his future** as it would be assumed if *decision* was not marked as a quasi-argument of the verb).

In order to reach this way of representation, there are two possibilities. First, we employ linguistic preprocessing of the data (including dependency parsing), then an LVC detector is used and in a post-processing step after syntactic parsing, the special relation of the nominal and the verbal component should be marked, i.e. certain syntactic labels are overwritten. Second, we execute parsing in a way that the training dataset already contains LVC-specific syntactic labels, that is, it is the dependency parser that carries out LVC detection. In this paper, we experiment with both ways and present and evaluate our results.

4 Related Work

There have been a considerable number of studies on LVC detection for several languages. They have been automatically identified in several languages such as English (Cook et al., 2007; Tu and Roth, 2011), Dutch (Van de Cruys and Moirón, 2007), Basque (Gurrutxaga and Alegria, 2011) and German (Evert and Kermes, 2003) just to mention a few.

We are aware of one machine learning system that identifies Hungarian LVCs in texts: the system described in Vincze et al. (2013) selects LVC candidates from texts on the basis of syntactic information, then in a second step it classifies them as genuine LVCs or not, using morphological, lexical, syntactic and semantic features.

Regarding the methods they use, Fazly and Stevenson (2007), Van de Cruys and Moirón (2007) and Gurrutxaga and Alegria (2011) used statistical features for identifying LVCs. Others employed rule-based systems (Diab and Bhutada, 2009; Nagy T. et al., 2011), which usually make use of (shallow) linguistic information. Some hybrid systems integrated both statistical and linguistic information as well (Tan et al., 2006; Tu and Roth, 2011).

As we aim at identifying LVCs by applying a dependency parser, next we concentrate on studies that are based on syntactic information and are related to MWE extraction. Seretan (2011) developed a method for collocation extraction based on syntactic constraints. Wehrli et al. (2010) argued that collocations can highly contribute to the performance of the parser since many parsing ambiguities can be excluded if collocations are known and treated as one syntactic unit. Nivre and Nilsson (2004) analyzed the influence of (previous) MWE recognition on dependency parsing and showed that known MWEs have a beneficial effect on parsing results. Korkontzelos and Manandhar (2010) investigated whether known MWEs improve the performance of statistical shallow parsers and found that they can significantly contribute to the efficiency of parsing. Eryiğit et al. (2011) analysed the impact of extracting MWEs on improving the accuracy of a dependency parser in Turkish. They found that the integration of compound verb and noun formations (which concept is similar to the one of light verb constructions applied here) has a detrimental effect on parsing accuracy since it increases lexical sparsity.

As can be seen, many previous studies examined the effects of already identified MWEs on the efficiency of parsing. On the other hand, there have been some current studies that aim at experimenting in the other direction, namely, using parsers for identifying MWEs: constituency parsing models are employed in identifying contiguous MWEs in French and Arabic (Green et al., 2013). Their method relied on a syntactic treebank, an MWE list and a morphological analyzer.

In this paper, we also experiment in this area: we employ a dependency parser for identifying

LVCs in Hungarian texts as a "side effect" of parsing sentences. Our dependency parser based method for identifying Hungarian LVCs is novel since to the best of our knowledge, dependency parsers have not been directly applied to identify LVCs. Moreover, it requires only a syntactic treebank enhanced with LVC annotation, in other words, there is no need to implement a separate LVC detector from scratch. In the following, we present our experiments and discuss our results.

5 Experiments

In this section, we will present our corpus, our methodology for detecting light verb constructions and we will show our results.

5.1 The Corpus

The Szeged Constituency Treebank has been manually annotated for light verb constructions (Vincze and Csirik, 2010). This treebank exists in another manually annotated version, namely, with dependency annotation (Vincze et al., 2010). Thus, manual annotations for LVCs and dependency structures are available for the same bunch of texts, which made it possible to map the two manual annotations. Thus, dependency relations were enhanced with LVC-specific relations that can be found between the two members of the constructions. For instance, instead of the traditional OBJ (object) relation, which occurred in the original version of the Szeged Dependency Treebank, the relation OBJ-LVC can be found between the words *döntést* (decision-ACC) and *hoz* "bring", members of the LVC *döntést hoz* "to make a decision" in the version used in this experiment. Here we provide a list of LVC-specific relations that occurred in our data (neglecting a handful of cases which were mislabeled due to some annotation errors in the dependency treebank):

- ATT-LVC – relation between a noun and a participial occurrence of a light verb:

  (a tegnap) adott tanács

  (the yesterday) given advice

  "(the) advice that was given (yesterday)"

- OBJ-LVC – relation between a light verb and its object:

  bejelentést tesz

  announcement-ACC makes

  "to make an announcement"

- OBL-LVC – relation between a light verb and its nominal argument (which is not the subject or object or dative):

  életbe lép

  life-ILL step

  "to take effect"

- SUBJ-LVC – relation between a light verb and its subject:

  sor kerül (vmire)

  turn get sg-SUB

  "the time has come for sg"

When mapping the LVC annotations and the dependency structures, we paid attention to the fact that it is only LVCs spelt as two tokens that could be identified with our methodology since no internal structure of compound words are marked in the Hungarian treebank and thus no dependency relation can be found among the members of the compound. So, we neglect LVCs spelt as one word and focus only on verbal and participial LVCs that consist of two members (cf. Footnote 1).

Figure 1 shows an example of a sentence with and without LVC-specific dependency labels. As can be seen, we have the light verb construction *döntést hoz* decision-ACC bring "to make a decision" in the sentence. However, it is parsed as a "normal" object of the verb in the first case (OBJ) and as a light verb object (OBJ-LVC) in the second case. Moreover, it is also seen that the two components of the LVC are not adjacent hence there are crossing branches in the dependency graph.

Although the entire Szeged Corpus contains manual LVC and dependency annotation, for the purpose of our study, we just selected texts from the law domain since they contain the biggest number of LVCs. Sentences in the law subcorpus were further filtered due to the fact that state-of-the-art dependency parsers cannot adequately treat verbless sentences, hence verbless sentences were ignored (see Farkas et al. (2012) for a detailed discussion of the problem). After this filtering step, we experimented with 6173 sentences, which consist of 156,744 tokens and contain 1101 LVCs. We present statistical data on the frequency of the LVC-specific relations in Table 1.

As Hungarian is a free word order language, the two components of LVCs, namely, the noun and the light verb, may not be adjacent in all cases,

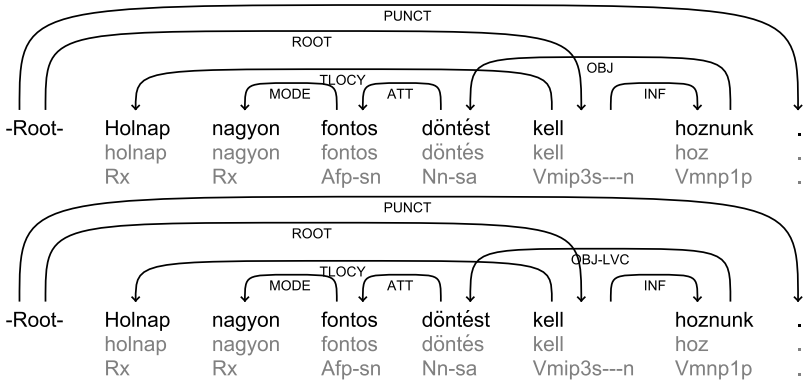

Figure 1: Dependency graph of the sentence *Holnap nagyon fontos döntést kell hoznunk* "Tomorrow we will have to make a very important decision" with or without LVC-specific dependency relations.

Relation	#	Non-contiguous	%
ATT-LVC	142	60	42.3
OBJ-LVC	587	231	39.4
OBL-LVC	266	50	18.8
SUBJ-LVC	102	4	3.9
Other-LVC	4	2	50
Total	1101	347	31.5

Table 1: Distribution of relations in the gold standard data and the frequency of non-contiguous LVCs.

which has a potentially detrimental effect on their identification in texts. Thus, we investigated the frequency of such cases in the data. Table 1 reveals that it is a quite frequent phenomenon in the corpus: almost one third of LVCs are non-contiguous. The largest distance between the noun and the verb is 21 tokens and the average distance between the two non-adjacent components is 4.28 tokens. All this suggests that sequence labeling approaches for LVC detection may not be as effective on the data as expected, however, a dependency parser that is able to identify long-distance dependencies may deal with the problem of non-adjacent but grammatically dependent elements in a more accurate way, which we will test below.

5.2 Dependency Parsing for LVC Detection

Farkas et al. (2012) carried out the first experiments on Hungarian dependency parsing. They empirically showed that state-of-the-art dependency parsers achieve similar results – in terms of attachment scores – on Hungarian and English. Although the results are not directly comparable due to domain differences and annotation schema divergences, they concluded that the difficulty of parsing Hungarian is very similar to parsing English and statistical dependency parsing is a viable way of parsing Hungarian, a morphologically rich language with free word order.

As their results indicated, the Bohnet dependency parser (Bohnet, 2010) proved to be the most effective on Hungarian data (Farkas et al., 2012), thus we applied it in our experiments too. It is an efficient second order dependency parser that models the interaction between siblings as well as grandchildren. Its decoder works on labeled edges, i.e. it uses a single-step approach for obtaining labeled dependency trees. It uses a rich and well-engineered feature set and it is enhanced by a Hash Kernel, which leads to higher accuracy.

Due to the free word order, there are quite many long-distance dependencies in Hungarian sentences, where a word and its parent are not adjacent (see also Figure 1). However, these linguistic phenomena are reasonably well-treated by dependency parsers. Furthermore, there seem to be quite a lot of non-contiguous LVCs in Hungarian. Hence, we think that these facts justify our experiments on applying a dependency parser for identifying LVCs.

5.3 Methodology

We trained and evaluated the Bohnet parser on the data in a ten-fold cross validation manner. To evaluate the quality of the dependency parsing, we applied the Labeled Attachment Score (LAS) and Unlabeled Attachment Score (ULA) metrics, taking into account punctuation as well. On the other hand, we also employed $F_{\beta=1}$ scores inter-

Method	Precision	Recall	F-score
Dictionary matching	0.7849	0.1229	0.2125
Classification	0.8284	0.6760	0.7445
Dependency parser	0.8660	0.6712	0.7563

Table 2: Results on LVC detection.

Method	Precision	Recall	F-score
Contiguous LVCs			
Classification	0.8746	0.7854	0.8276
Dependency parser	0.9008	0.7357	0.8099
Non-contiguous LVCs			
Classification	0.7103	0.5188	0.6000
Dependency parser	0.7940	0.5362	0.6401

Table 3: Results on detecting contiguous and non-contiguous LVCs.

preted on the LVC-specific relations to evaluate the performance of detecting LVCs in the corpus and we evaluated our system on contiguous and non-contiguous LVCs as well.

As baselines, we made use of the methods described in Vincze et al. (2013). They first employed dictionary matching, where LVCs collected from a parallel corpus annotated for Hungarian LVCs (Vincze, 2012) were mapped to the lemmatized texts. We also applied dictionary matching as one of our baselines. The main method of Vincze et al. (2013) first parsed each sentence and extracted potential LVCs on the basis of the dependency relations found between verb-object, verb-subject, verb-prepositional object, verb-other argument and noun-modifier pairs. The dependency labels were provided by `magyarlanc` (Zsibrita et al., 2013). Later, C4.5 decision trees were applied to classify candidate LVCs, which exploits a rich feature set. For instance, morphological features exploited the fact that the nominal component of LVCs is typically derived from a verbal stem or coincides with a verb, on the other hand, the POS tags of the words and surrounding words were also used as features. As for semantic features, the `activity` or `event` semantic senses were looked for among the upper level hyperonyms of the head of the noun phrase in the Hungarian WordNet[3]. As lexical features, fifteen typical light verbs were selected from the list of the most frequent verbs taken from the Szeged ParalellFX corpus (Vincze, 2012) and it was checked whether the lemmatised verbal component of the candidate was one of these fifteen verbs. The lemma of the noun was also applied as a lexical feature.

We evaluated our database with this system too in a ten-fold cross validation manner (using the same data splits as previously) and as evaluation metrics, we employed $F_{\beta=1}$ scores. The results of our experiments are shown in Tables 2 and 3.

6 Results

As Table 2 shows, the dependency parser with the LVC-specific relations achieved an F-score of 0.7563 (recall: 0.6712, precision: 0.8660) interpreted on the LVC-specific relations. This result exceeds the ones obtained by the baselines: it outperforms the dictionary matching method by 54.38% in terms of F-score with a considerably better recall value, and, on the other hand, it also performs better than the classification method with a 1.18% gain in F-score – the results are significant (ANOVA, p = 0.012). In the latter case, the improvement is due to the higher precision value.

The identification of non-contiguous LVCs proved to be more difficult for both methods than that of contiguous LVCs. The classification approach significantly outperforms the dependency parser on the contiguous LVC class (ANOVA, p = 0.0455) but on the non-contiguous class the dependency parser performs significantly better with an F-score of 0.6401 (ANOVA, p = 0.0343).

In order to analyze the performance in more detail, we compared the precision, recall and F-scores for each LVC-specific label. Data in Table 4 reveal that SUBJ-LVCs are the easiest ones to predict (with both high precision and recall values) and participial uses of LVCs are the most difficult to identify (ATT-LVC) relation between the noun and the participle, mostly due to the low recall value. Although the precision value is rather low in the case of objects (OBJ-LVC), objects and other arguments (OBL-LVC) can be detected reasonably well. Table 5 shows results for (non-)contiguous LVC classes. It is revealed that for OBL-LVCs, there is no substantial difference between contiguous and non-contiguous LVCs but for objects and participial LVCs, the dis-

[3] `http://www.inf.u-szeged.hu/rgai/HuWN`

Relation	#	Precision	Recall	F-score
ATT-LVC	142	0.8267	0.4366	0.5714
OBJ-LVC	587	0.8365	0.6712	0.7448
OBL-LVC	266	0.9175	0.7105	0.8008
SUBJ-LVC	102	0.9592	0.9216	0.9400
Other-LVC	4	–	–	–

Table 4: Distribution of relations in the gold standard data and results in terms of precision, recall and F-score as predicted by the dependency parser.

Relation & type	Precision	Recall	F-score
ATT-LVC C	0.9524	0.4878	0.6452
ATT-LVC NC	0.6667	0.3667	0.4731
OBJ-LVC C	0.8535	0.7507	0.7988
OBJ-LVC NC	0.8025	0.5478	0.6512
OBL-LVC C	0.9226	0.7176	0.8073
OBL-LVC NC	0.8947	0.6800	0.7727
SUBJ-LVC C	0.9785	0.9286	0.9529
SUBJ-LVC NC	0.6000	0.7500	0.6667

Table 5: Results in terms of precision, recall and F-score as predicted by the dependency parser for (non-)contiguous (NC/C) LVC classes.

tance between the two components of the LVC has an essential effect on the efficiency.[4]

As for the performance on dependency parsing, we got 90.38 (LAS) and 92.12 (ULA) when training with LVC-specific relations. If these results are compared to those achieved with traditional (i.e. non-LVC-specific) relations, then it is revealed that in the latter case LAS is 90.63, i.e. 0.25 percentage point higher, which can be considered negligible.

7 Discussion

As the results show, the dependency parsing approach achieved the best results on LVC detection, especially due to the high precision score. This is probably due to the rich feature set applied by the Bohnet parser. Furthermore, our approach to solve the problem of LVC detection as a classification of syntactic constructions by using a dependency parser is also justified by these results.

A comparison with previous parser-based approach to MWE detection might also prove useful. Green et al. (2013) employed constituency parsers to identify contiguous MWEs in French and Arabic. As a main difference between our approach and theirs, we applied a dependency parser for the task of LVC detection, which proved especially effective since we worked with a free word order language, thus we had to deal with non-contiguous LVCs as well. Our dependency parser approach could adequately identify them as well, however, experimenting with a constituency parser will be a possible way to continue our work.

In Hungarian, it sometimes happens that a sequence that looks like an LVC is actually not an LVC in the specific context as in *A dékán újabb előadást tartott szükségesnek* the dean new-COMP presentation-ACC hold-PAST-3SG necessary-DAT "The dean thought that another presentation was necessary". In other contexts, *előadást tart* presentation-ACC hold "to have a presentation" would most probably function as an LVC. However, in this case we encounter with another fixed grammatical construction of Hungarian, namely, *valamilyennek tart valamit* somewhat-DAT hold something-ACC "to regard something as something", e.g. *szépnek tartja a lányt* beautiful-DAT hold-3SG-OBJ the girl-ACC "he thinks that the girl is beautiful". Thus, there is no LVC in the above example, but approaches that heavily build on MWE lexicons may falsely identify this verb-object pair as a light verb object-light verb pair since they hardly consider contextual information. In contrast, dependency parsers have access to information about other dependents of the verb hence they may learn that in such cases the presence of a dative dependent argues against the identification of the verb-object pair as an LVC.

As for the specific LVC-relations, our approach was most successful on LVCs where the noun fulfilled the role of the subject (i.e. it had the relation SUBJ-LVC). This may be attributed to the fact that these LVCs are the least diverse in the corpus: there are only a handful of such types, and each LVC type has several occurrences in the data thus they can be easily identified. On the other hand, participial uses of LVCs (ATT-LVC) were the hardest to detect, which is partly due to their lexical divergence and partly due to the fact that currently adjectives and participles are not distinguished in Hungarian morphological parsing, i.e. they have the same morphological codes. Thus, the parser, which heavily builds on morpho-

[4] As there were hardly any non-contiguous SUBJ-LVCs in the dataset, we cannot draw any conclusions on the difficulty level of identifying non-contiguous light verb subjects.

logical information, has no chance to learn that it is only participles that tend to occur as parts of LVCs but adjectives do not. A distinction of participles and adjectives in the Hungarian computational morphology would most probably have beneficial effects on identifying LVCs.

Our results empirically prove that a dependency parser may be effectively applied to identify LVCs in free texts, provided that we have a dependency model trained on LVC-specific relations, which itself requires a treebank manually annotated for dependency relations and LVCs. Although the LAS scores are somewhat lower than in the case of LVC-less dependency relations, the task of LVC detection can be also performed by the parser. On the other hand, the classification approach needs a trained dependency model since it classifies LVC candidates selected on the basis of syntactic information. It also uses LVC lists gathered from annotated corpora and in order to denote LVC-specific relations (i.e. quasi-arguments) in the case of complex predicates, an extra post-processing step is needed in the workflow. Thus, the resources needed by the two approaches are the same but with the dependency parsing approach, the implementation of a new LVC-detector from scratch might be saved and complex predicates are provided immediately by the parser. Moreover, another advantage of the dependency parser is that it performs better on non-contiguous LVCs, which are frequent in Hungarian.

We also carried out an error analysis in order to compare the two methods. It was difficult for both the dependency parser and the classifier to recognize rare LVCs or those that included a non-frequent light verb. A typical source of error for the dependency parser was that sometimes an LVC-specific relation was proposed for non-nouns (e.g. adverbs or conjunctions) as well, like in *akár írnia* (either write-INF.3SG) "either he should write", where *akár* was labeled as an LVC-object of the verb instead of a conjunction. Furthermore, the classifier often made an error in cases where the sentence included an LVC but another argument of the verb was labeled as part of the LVC, e.g. *filmet forgalomba hoz* (film-ACC circulation-INE bring) "to put a film into circulation", where the gold standard LVC is *forgalomba hoz* "to put something into circulation" but *filmet hoz* "to bring a film" was labeled as a false positive LVC. Since different phenomena proved to be difficult for the two systems, a possible direction for future work may be to combine the two approaches in order to minimize prediction errors.

Here we experimented with Hungarian, a morphologically rich language. Nevertheless, we believe that the method of applying a dependency parser for LVC detection is not specific to this typological class of languages and it can be employed for any language that has a dependency treebank which contains annotation for LVCs.

8 Conclusions

In this paper, we empirically showed that a dependency parser can be employed to detect LVCs in free texts. For this, we used a Hungarian treebank, which has been manually annotated for dependency relations and light verb constructions. Our results outperformed those achieved by state-of-the-art techniques for Hungarian LVC detection and the main advantages of our system is its high precision on the one hand and the adequate treatment of non-contiguous LVCs on the other hand.

The error analysis of the systems applied suggests that since the two systems make errors in different cases, combining them may lead to more precise results. Another possible way of improving the system is to explore methods for the treatment of participial LVCs. Furthermore, as future work we aim at experimenting with the dependency parser in other scenarios (e.g. the newspaper subcorpus of the Szeged Dependency Treebank) in order to make further generalizations on the role of dependency parsing in LVC detection.

Acknowledgments

This work was supported in part by the European Union and the European Social Fund through the project FuturICT.hu (grant no.: TÁMOP-4.2.2.C-11/1/KONV-2012-0013) and by the COST Action PARSEME (IC1207).

References

Margarita Alonso Ramos. 1998. *Etude sémantico-syntaxique des constructions à verbe support*. Ph.D. thesis, Université de Montréal, Montreal, Canada.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of Coling 2010*, pages 89–97.

Noam Chomsky. 1981. *Lectures on Government and Binding*. Foris, Dordrecht.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of MWE 2007*, pages 41–48, Morristown, NJ, USA. ACL.

Mona Diab and Pravin Bhutada. 2009. Verb Noun Construction MWE Token Classification. In *Proceedings of MWE 2009*, pages 17–22, Singapore, August. ACL.

Katalin É. Kiss. 2002. *The Syntax of Hungarian*. Cambridge University Press, Cambridge.

Gülşen Eryiğit, Tugay Ilbay, and Ozan Arkan Can. 2011. Multiword expressions in statistical dependency parsing. In *Proceedings of SPMRL 2011*, pages 45–55, Dublin, Ireland. ACL.

Stefan Evert and Hannah Kermes. 2003. Experiments on candidate data for collocation extraction. In *Proceedings of EACL 2003*, pages 83–86.

Richárd Farkas, Veronika Vincze, and Helmut Schmid. 2012. Dependency Parsing of Hungarian: Baseline Results and Challenges. In *Proceedings of EACL 2012*, pages 55–65, Avignon, France, April. ACL.

Afsaneh Fazly and Suzanne Stevenson. 2007. Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In *Proceedings of MWE 2007*, pages 9–16, Prague, Czech Republic, June. ACL.

Spence Green, Marie-Catherine de Marneffe, and Christopher D. Manning. 2013. Parsing models for identifying multiword expressions. *Computational Linguistics*, 39(1):195–227.

Antton Gurrutxaga and Iñaki Alegria. 2011. Automatic Extraction of NV Expressions in Basque: Basic Issues on Cooccurrence Techniques. In *Proceedings of MWE 2011*, pages 2–7, Portland, Oregon, USA, June. ACL.

Su Nam Kim. 2008. *Statistical Modeling of Multiword Expressions*. Ph.D. thesis, University of Melbourne, Melbourne.

Ioannis Korkontzelos and Suresh Manandhar. 2010. Can recognising multiword expressions improve shallow parsing? In *NAACL HLT 2010*, pages 636–644, Los Angeles, California, June. ACL.

Niloofar Mansoory and Mahmood Bijankhan. 2008. The possible effects of Persian light verb constructions on Persian WordNet. In *Proceedings of GWC 2008*, pages 297–303, Szeged, Hungary, January. University of Szeged.

Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. The NomBank Project: An Interim Report. In *HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 24–31, Boston, Massachusetts, USA. ACL.

Kadri Muischnek and Heiki Jaan Kaalep. 2010. The variability of multi-word verbal expressions in Estonian. *Language Resources and Evaluation*, 44(1-2):115–135.

István Nagy T., Veronika Vincze, and Gábor Berend. 2011. Domain-dependent identification of multiword expressions. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

Joakim Nivre and Jens Nilsson. 2004. Multiword units in syntactic parsing. In *Proceedings of MEMURA 2004*.

Violeta Seretan. 2011. *Syntax-Based Collocation Extraction*. TSD. Springer, Dordrecht.

Yee Fan Tan, Min-Yen Kan, and Hang Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of MWE 2006*, pages 49–56, Trento, Italy, April. ACL.

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

Tim Van de Cruys and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of MWE 2007*, pages 25–32, Morristown, NJ, USA. ACL.

Veronika Vincze and János Csirik. 2010. Hungarian corpus of light verb constructions. In *Proceedings of Coling 2010*, pages 1110–1118, Beijing, China, August. Coling 2010 Organizing Committee.

Veronika Vincze, Dóra Szauter, Attila Almási, György Móra, Zoltán Alexin, and János Csirik. 2010. Hungarian Dependency Treebank. In *Proceedings of LREC 2010*, Valletta, Malta, May. ELRA.

Veronika Vincze, István Nagy T., and Richárd Farkas. 2013. Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach. In *Proceedings of ACL-2013: Short Papers*, Sofia. ACL.

Veronika Vincze. 2012. Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus. In *Proceedings of LREC 2012*, Istanbul, Turkey.

Eric Wehrli, Violeta Seretan, and Luka Nerima. 2010. Sentence analysis and collocation identification. In *Proceedings of MWE 2010*, pages 28–36, Beijing, China, August. Coling 2010 Organizing Committee.

János Zsibrita, Veronika Vincze, and Richárd Farkas. 2013. magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian. In *Proceedings of RANLP-2013*, Hissar, Bulgaria.