Full-coverage Identification of English Light Verb Constructions

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

Abstract

The identification of light verb constructions (LVC) is an important task for several applications. Previous studies focused on some limited set of light verb constructions. Here, we address the full coverage of LVCs. We investigate the performance of different candidate extraction methods on two English full-coverage LVC annotated corpora, where we found that less severe candidate extraction methods should be applied. Then we follow a machine learning approach that makes use of an extended and rich feature set to select LVCs among extracted candidates.

1 Introduction

A multiword expression (MWE) is a lexical unit that consists of more than one orthographical word, i.e. a lexical unit that contains spaces and displays lexical, syntactic, semantic, pragmatic and/or statistical idiosyncrasy (Sag et al., 2002; Calzolari et al., 2002). Light verb constructions (LVCs) (e.g. *to take a decision, to take sg into consideration*) form a subtype of MWEs, namely, they consist of a nominal and a verbal component where the verb functions as the syntactic head (the whole construction fulfills the role of a verb in the clause), but the semantic head is the noun (i.e. the noun is used in one of its original senses). The verbal component (also called a light verb) usually loses its original sense to some extent.[1] The meaning of LVCs can only partially be computed on the basis of the meanings of their parts and the way they are related to each other (semi-compositionality). Thus, the result of translating their parts literally can hardly be considered as the proper translation of the original expression. Moreover, the same syntactic pattern may belong to a LVC (e.g. *make a mistake*), a literal verb + noun combination (e.g. *make a cake*) or an idiom (e.g. *make a meal (of something)*), which suggests that their identification cannot be based on solely syntactic patterns. Since the syntactic and the semantic head of the construction are not the same, they require special treatment when parsing. On the other hand, the same construction may function as an LVC in certain contexts while it is just a productive construction in other ones, compare *He **gave her a ring** made of gold* (non-LVC) and *He **gave her a ring** because he wanted to hear her voice* (LVC).

In several natural language processing (NLP) applications like information extraction and retrieval, terminology extraction and machine translation, it is important to identify LVCs in context. For example, in machine translation we must know that LVCs form one semantic unit, hence their parts should not be translated separately. For this, LVCs should be identified first in the text to be translated.

As we shall show in Section 2, there has been a considerable amount of previous work on LVC detection, but some authors seek to capture just verb–object pairs, while others just verbs with prepositional complements. Actually, many of them exploited only constructions formed with a limited set of light verbs and identified or extracted just a specific type of LVCs. However, we cannot see any benefit that any NLP application could get from these limitations and here, we focus on the full-coverage identification of LVCs. We train and evaluate statistical models on the Wiki50 (Vincze et al., 2011) and Szeged-ParalellFX (SZPFX) (Vincze, 2012) corpora that have recently been published with full-coverage LVC annotation.

We employ a two-stage procedure. First,

[1]*Light verbs* may also be defined as semantically empty support verbs, which share their arguments with a noun (see the NomBank project (Meyers et al., 2004)), that is, the term *support verb* is a hypernym of *light verb*.

International Joint Conference on Natural Language Processing, pages 329–337,
Nagoya, Japan, 14-18 October 2013.

we identify potential LVC candidates in running texts – we empirically compare various candidate extraction methods –, then we use a machine learning-based classifier that exploits a rich feature set to select LVCs from the candidates.

The main contributions of this paper can be summarized as follows:

- We introduce and evaluate systems for **identifying all LVCs and all individual LVC occurrences** in a running text and we do not restrict ourselves to certain specific types of LVCs.
- We systematically **compare and evaluate different candidate extraction methods** (earlier published methods and new solutions implemented by us).
- We defined and evaluated several **new feature templates** like semantic or morphological features to select LVCs in context from extracted candidates.

2 Related Work

Two approaches have been introduced for LVC detection. In the first approach, LVC candidates (usually verb-object pairs including one verb from a well-defined set of 3-10 verbs) are extracted from the corpora and these tokens – without contextual information – are then classified as LVCs or not (Stevenson et al., 2004; Tan et al., 2006; Fazly and Stevenson, 2007; Van de Cruys and Moirón, 2007; Gurrutxaga and Alegria, 2011). As a gold standard, lists collected from dictionaries or other annotated corpora are used: if the extracted candidate is classified as an LVC and can be found on the list, it is a true positive, regardless of the fact whether it was a genuine LVC in its context.

In the second approach, the goal is to detect individual LVC token instances in a running text, taking contextual information into account (Diab and Bhutada, 2009; Tu and Roth, 2011; Nagy T. et al., 2011). While the first approach assumes that a specific candidate in all of its occurrences constitutes an LVC or not (i.e. there are no ambiguous cases), the second one may account for the fact that there are contexts where a given candidate functions as an LVC whereas in other contexts it does not, recall the example of *give a ring* in Section 1.

The authors of Stevenson et al. (2004), Fazly and Stevenson (2007), Van de Cruys and Moirón (2007) and Gurrutxaga and Alegria (2011) built LVC detection systems with statistical features. Stevenson et al. (2004) focused on classifying LVC candidates containing the verbs *make* and *take*. Fazly and Stevenson (2007) used linguistically motivated statistical measures to distinguish subtypes of verb + noun combinations. However, it is a challenging task to identify rare LVCs in corpus data with statistical-based approaches, since 87% of LVCs occur less than 3 times in the two full-coverage LVC annotated corpora used for evaluation (see Section 3).

A semantic-based method was described in Van de Cruys and Moirón (2007) for identifying verb-preposition-noun combinations in Dutch. Their method relies on selectional preferences for both the noun and the verb. Idiomatic and light verb noun + verb combinations were extracted from Basque texts by employing statistical methods (Gurrutxaga and Alegria, 2011). Diab and Bhutada (2009) and Nagy T. et al. (2011) employed ruled-based methods to detect LVCs, which are usually based on (shallow) linguistic information, while the domain specificity of the problem was highlighted in Nagy T. et al. (2011).

Both statistical and linguistic information were applied by the hybrid LVC systems (Tan et al., 2006; Tu and Roth, 2011; Samardžić and Merlo, 2010), which resulted in better recall scores. English and German LVCs were analysed in parallel corpora: the authors of Samardžić and Merlo (2010) focus on their manual and automatic alignment. They found that linguistic features (e.g. the degree of compositionality) and the frequency of the construction both have an impact on the alignment of the constructions.

Tan et al. (2006) applied machine learning techniques to extract LVCs. They combined statistical and linguistic features, and trained a random forest classifier to separate LVC candidates. Tu and Roth (2011) applied Support Vector Machines to classify verb + noun object pairs on their balanced dataset as candidates for true LVCs[2] or not. They compared the contextual and statistical features and found that local contextual features performed better on ambiguous examples.

[2]In theoretical linguistics, two types of LVCs are distinguished (Kearns, 2002). In true LVCs such as *to have a laugh* we can find a noun that is a conversive of a verb (i.e. it can be used as a verb without any morphological change), while in vague action verbs such as *to make an agreement* there is a noun derived from a verb (i.e. there is morphological change).

Some of the earlier studies aimed at identifying or extracting only a restricted set of LVCs. Most of them focus on verb-object pairs when identifying LVCs (Stevenson et al., 2004; Tan et al., 2006; Fazly and Stevenson, 2007; Cook et al., 2007; Bannard, 2007; Tu and Roth, 2011), thus they concentrate on structures like *give a decision* or *take control*. With languages other than English, authors often select verb + prepositional object pairs (instead of verb-object pairs) and categorise them as LVCs or not. See, e.g. Van de Cruys and Moirón (2007) for Dutch LVC detection or Krenn (2008) for German LVC detection. In other cases, only true LVCs were considered (Stevenson et al., 2004; Tu and Roth, 2011). In some other studies (Cook et al., 2007; Diab and Bhutada, 2009) the authors just distinguished between the literal and idiomatic uses of verb + noun combinations and LVCs were classified into these two categories as well.

In contrast to previous works, we seek to identify all LVCs in running texts and do not restrict ourselves to certain types of LVCs. For this reason, we experiment with different candidate extraction methods and we present a machine learning-based approach to select LVCs among candidates.

3 Datasets

In our experiments, three freely available corpora were used. Two of them had fully-covered LVC sets manually annotated by professional linguists. The annotation guidelines did not contain any restrictions on the inner syntactic structure of the construction and both true LVCs and vague action verbs were annotated. The Wiki50 (Vincze et al., 2011) contains 50 English Wikipedia articles that were annotated for different types of MWEs (including LVCs) and Named Entities. SZPFX (Vincze, 2012) is an English–Hungarian parallel corpus, in which LVCs are annotated in both languages. It contains texts taken from several domains like fiction, language books and magazines. Here, the English part of the corpus was used.

In order to compare the performance of our system with others, we also used the dataset of Tu and Roth (2011), which contains 2,162 sentences taken from different parts of the British National Corpus. They only focused on true LVCs in this dataset, and only the verb-object pairs (1,039 positive and 1,123 negative examples) formed with the verbs *do*, *get*, *give*, *have*, *make*, *take* were marked. Statistical data on the three corpora are listed in Table 1.

Corpus	Sent.	Tokens	LVCs	LVC lemma
Wiki50	4,350	114,570	368	287
SZPFX	14,262	298,948	1,371	706
Tu&Roth	2,162	65,060	1,039	430

Table 1: Statistical data on LVCs in the Wiki50 and SZPFX corpora and the Tu&Roth dataset.

Despite the fact that English verb + prepositional constructions were mostly neglected in previous research, both corpora contain several examples of such structures, e.g. *take into consideration* or *come into contact*, the ratio of such LVC lemmas being 11.8% and 9.6% in the Wiki50 and SZPFX corpora, respectively. In addition to the verb + object or verb + prepositional object constructions, there are several other syntactic constructions in which LVCs can occur due to their syntactic flexibility. For instance, the nominal component can become the subject in a passive sentence (*the **photo** has been **taken***), or it can be extended by a relative clause (*the **photo** that has been **taken***). These cases are responsible for 7.6% and 19.4% of the LVC occurrences in the Wiki50 and SZPFX corpora, respectively. These types cannot be identified when only verb + object pairs are used for LVC candidate selection.

Some researchers filtered LVC candidates by selecting only certain verbs that may be part of the construction, e.g. Tu and Roth (2011). As the full-coverage annotated corpora were available, we were able to check what percentage of LVCs could be covered with this selection. The six verbs used by Tu and Roth (2011) are responsible for about 49% and 63% of all LVCs in the Wiki50 and the SZPFX corpora, respectively. Furthermore, 62 different light verbs occurred in the Wiki50 and 102 in the SZPFX corpora, respectively. All this indicates that focusing on a reduced set of light verbs will lead to the exclusion of a considerable number of LVCs in free texts.

Some papers focus only on the identification of true LVCs, neglecting vague action verbs (Stevenson et al., 2004; Tu and Roth, 2011). However, we cannot see any NLP application that can benefit if such a distinction is made since vague action verbs and true LVCs share those properties that are relevant for natural language processing (e.g. they must be treated as one complex predicate (Vincze,

2012)). We also argue that it is important to separate LVCs and idioms because LVCs are semi-productive and semi-compositional – which may be exploited in applications like machine translation or information extraction – in contrast to idioms, which have neither feature. All in all, we seek to identify all verbal LVCs (not including idioms) in our study and do not restrict ourselves to certain specific types of LVCs.

4 LVC Detection

Our goal is to identify each LVC occurrence in running texts, i.e. to take input sentences such as *'We often have lunch in this restaurant'* and mark each LVC in it. Our basic approach is to syntactically parse each sentence and extract potential LVCs with different candidate extraction methods. Afterwards, a binary classification can be used to automatically classify potential LVCs as LVCs or not. For the automatic classification of candidate LVCs, we implemented a machine learning approach, which is based on a rich feature set.

4.1 Candidate Extraction

As we had two full-coverage LVC annotated corpora where each type and individual occurrence of a LVC was marked in running texts, we were able to examine the characteristics of LVCs in a running text, and evaluate and compare the different candidate extraction methods. When we examined the previously used methods, which just treated the verb-object pairs as potential LVCs, it was revealed that only 73.91% of annotated LVCs on the Wiki50 and 70.61% on the SZPFX had a verb-object syntactic relation. Table 2 shows the distribution of dependency label types provided by the Bohnet parser (Bohnet, 2010) for the Wiki50 and Stanford (Klein and Manning, 2003) and the Bohnet parsers for the SZPFX corpora. In order to compare the efficiency of the parsers, both were applied using the same dependency representation. In this phase, we found that the Bohnet parser was more successful on the SZPFX corpora, i.e. it could cover more LVCs, hence we applied the Bohnet parser in our further experiments.

We define the extended **syntax-based candidate extraction** method, where besides the *verb-direct object* dependency relation, the *verb-prepositional*, *verb-relative clause*, *noun-participial modifier* and *verb-subject of a passive construction* syntactic relations were also investigated among verbs and nouns. Here, 90.76% of LVCs in the Wiki50 and 87.75% in the SZPFX corpus could be identified with the extended syntax-based candidate extraction method.

It should be added that some rare examples of split LVCs where the nominal component is part of the object, preceded by a quantifying expression like *he **gained** much of his **fame*** can hardly be identified by syntax-based methods since there is no direct link between the verb and the noun. In other cases, the omission of LVCs from candidates is due to the rare and atypical syntactic relation between the noun and the verb (e.g. `dep` in *reach conform*). Despite this, such cases are also included in the training and evaluation datasets as positive examples.

Edge type	Wiki50		SZPFX			
			Stanford		Bohnet	
dobj	272	73.91	901	65.71	968	70.6
pobj	43	11.69	93	6.78	93	6.78
nsubjpass	6	1.63	61	4.45	73	5.32
rcmod	6	1.63	30	2.19	38	2.77
partmod	7	1.9	21	1.53	31	2.26
sum	334	90.76	1,106	80.67	1,203	87.75
other	15	4.07	8	0.58	31	2.26
none	19	5.17	257	18.75	137	9.99
sum	368	100.0	1,371	100.0	1,371	100.0

Table 2: Edge types in the Wiki50 and SZPFX corpora. dobj: object. pobj: preposition. nsubjpass: subject of a passive construction. rcmod: relative clause. partmod: participial modifier. other: other dependency labels. none: no direct syntactic connection between the verb and noun.

Our second candidate extractor is the **morphology-based candidate extraction** method (Nagy T. et al., 2011), which was also applied for extracting potential LVCs. In this case, a token sequence was treated as a potential LVC if the POS-tag sequence matched one pattern typical of LVCs (e.g. `VERB-NOUN`). Although this method was less effective than the extended syntax-based approach, when we **merged the extended syntax-based and morphology-based methods**, we were able to identify most of the LVCs in the two corpora.

The authors of Stevenson et al. (2004) and Tu and Roth (2011) filtered LVC candidates by selecting only certain verbs that could be part of the construction, so we checked what percentage of LVCs could be covered with this selection when we treated just the verb-object pairs as LVC candidates. We found that even the least stringent selec-

tion covered only 41.88% of the LVCs in Wiki50 and 47.84% in SZPFX. Hence, we decided to drop any such constraint.

Table 3 shows the results we obtained by applying the different candidate extraction methods on the Wiki50 and SZPFX corpora.

Method	Wiki50		SZPFX	
	#	%	#	%
Stevenson et al. (2004)	107	29.07	372	27.13
Tu&Roth (2011)	154	41.84	656	47.84
dobj	272	73.91	968	70.6
POS	293	79.61	907	66.15
Syntactic	334	90.76	1,203	87.75
POS ∪ Syntactic	339	92.11	1,223	89.2

Table 3: The recall of candidate extraction approaches. dobj: verb-object pairs. POS: morphology-based method. Syntactic: extended syntax-based method. POS ∪ Syntactic: union of the morphology- and extanded syntax-based candidate extraction methods.

4.2 Machine Learning Based Candidate Classification

For the automatic classification of the candidate LVCs we implemented a machine learning approach, which we will elaborate upon below. Our method is based on a rich feature set with the following categories: statistical, lexical, morphological, syntactic, orthographic and semantic.

Statistical features: Potential LVCs were collected from 10,000 Wikipedia pages by the union of the morphology-based candidate extraction and the extended syntax-based candidate extraction methods. The number of their **occurrences** was used as a feature in case the candidate was one of the syntactic phrases collected.

Lexical features: We exploit the fact that the **most common verbs** are typically light verbs, so we selected fifteen typical light verbs from the list of the most frequent verbs taken from the corpora. In this case, we investigated whether the lemmatised verbal component of the candidate was one of these fifteen verbs. The **lemma of the head of the noun** was also applied as a lexical feature. The nouns found in LVCs were collected from the corpora, and for each corpus the noun list got from the union of the other two corpora was used. Moreover, we constructed **lists of lemmatised LVCs** from the corpora and for each corpus, the list got from the union of the other two corpora was utilised. In the case of the Tu&Roth dataset, the list got from Wiki50 and SZPFX was filtered for the six light verbs and true LVCs they contained.

Morphological features: The POS candidate extraction method was used as a feature, so when the POS-tag sequence in the text matched one typical **'POS-pattern'** of LVCs, the candidate was marked as *true*; otherwise as *false*. The **'Verbal-Stem'** binary feature focuses on the stem of the noun. For LVCs, the nominal component is typically one that is derived from a verbal stem (*make a decision*) or coincides with a verb (*have a walk*). In this case, the phrases were marked as *true* if the stem of the nominal component had a verbal nature, i.e. it coincided with a stem of a verb. *Do* and *have* are often light verbs, but these verbs may occur as auxiliary verbs too. Hence we defined a feature for the two verbs to denote whether or not they were **auxiliary verbs** in a given sentence.

Syntactic features: The **dependency label** between the noun and the verb can also be exploited in identifying LVCs. As we typically found in the candidate extraction, the syntactic relation between the verb and the nominal component in an LVC is `dobj`, `pobj`, `rcmod`, `partmod` or `nsubjpass` – using the Bohnet parser (Bohnet, 2010), hence these relations were defined as features. The **determiner** within all candidate LVCs was also encoded as another syntactic feature.

Orthographic features: in the case of the **'suffix'** feature, it was checked whether the lemma of the noun ended in a given character bi- or trigram. It exploits the fact that many nominal components in LVCs are derived from verbs. The **'number of words'** of the candidate LVC was also noted and applied as a feature.

Semantic features: In this case we also exploited the fact that the nominal component is derived from verbs. `Activity` or `event` semantic senses were looked for among the hypernyms of the noun in WordNet (Fellbaum, 1998).

We experimented with several learning algorithms and our preliminary results showed that decision trees performed the best. This is probably due to the fact that our feature set consists of a few compact – i.e. high-level – features. We trained the J48 classifier of the WEKA package (Hall et al., 2009), which implements the decision trees algorithm C4.5 (Quinlan, 1993) with the above-mentioned feature set. We report results with Support Vector Machines (SVM) (Cortes and Vapnik, 1995) as well, to compare our methods with Tu &

Method	Wiki50						SZPFX					
	J48			SVM			J48			SVM		
	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score	Prec.	Rec.	F-score
DM	56.11	36.26	44.05	56.11	36.26	44.05	72.65	27.83	40.24	72.65	27.83	40.24
POS	60.65	46.2	52.45	54.1	48.64	51.23	66.12	43.02	52.12	54.88	42.42	47.85
Syntax	61.29	47.55	53.55	50.99	51.63	51.31	63.25	56.17	59.5	54.38	54.03	54.2
POS∪Syntax	58.99	51.09	54.76	49.72	51.36	50.52	63.29	56.91	59.93	55.84	55.14	55.49

Table 4: Results obtained in terms of precision, recall and F-score. DM: dictionary matching. POS: morphology-based candidate extraction. Syntax: extended syntax-based candidate extraction. POS ∪ Syntax: the merged set of the morphology-based and syntax-based candidate extraction methods.

Roth.

As the investigated corpora were not sufficiently big for splitting them into training and test sets of appropriate size, besides, the different annotation principles ruled out the possibility of enlarging the training sets with another corpus, we evaluated our models in 10-fold cross validation manner on the Wiki50, SZPFX and Tu&Roth datasets. But, in the case of Wiki50 and SZPFX, where only the positive LVCs were annotated, we employed $F_{\beta=1}$ scores interpreted on the positive class as an evaluation metric. Moreover, we treated all potential LVCs as negative which were extracted by different extraction methods but were not marked as positive in the gold standard. The resulting datasets were not balanced and the number of negative examples basically depended on the candidate extraction method applied.

However, some positive elements in the corpora were not covered in the candidate classification step, since the candidate extraction methods applied could not detect all LVCs in the corpus data. Hence, we treated the omitted LVCs as false negatives in our evaluation.

5 Experiments and Results

As a baseline, we applied a context-free dictionary matching method. First, we gathered the gold-standard LVC lemmas from the two other corpora. Then we marked candidates of the union of the extended syntax-based and morphology-based methods as LVC if the candidate light verb and one of its syntactic dependents was found on the list.

Table 4 lists the results got on the Wiki50 and SZPFX corpora by using the baseline dictionary matching and our machine learning approach with different machine learning algorithm and different candidate extraction methods.The dictionary matching approach got the highest precision on SZPFX, namely 72.65%. Our machine learning-based approach with different candidate extraction methods demonstrated a consistent performance (i.e. an F-score over 50) on the Wiki50 and SZPFX corpora. It is also seen that our machine learning approach with the union of the morphology- and extended syntax-based candidate extraction methods is the most successful method in the case of Wiki50 and SZPFX. On both corpora, it achieved an F-score that was higher than that of the dictionary matching approach (the difference being 10 and 19 percentage points in the case of Wiki50 and SZPFX, respectively).

In order to compare the performance of our system with others, we evaluated it on the Tu&Roth dataset (Tu and Roth, 2011) too. Table 5 shows the results got using dictionary matching, applying our machine learning-based approach with a rich feature set, and the results published in Tu and Roth (2011) on the Tu&Roth dataset. In this case, the dictionary matching method performed the worst and achieved an accuracy score of 61.25. The results published in Tu and Roth (2011) are good on the positive class with an F-score of 75.36 but the worst with an F-score of 56.41 on the negative class. Therefore this approach achieved an accuracy score that was 7.27 higher than that of the dictionary matching method. Our approach demonstrates a consistent performance (with an F-score over 70) on the positive and negative classes. It is also seen that our approach is the most successful in the case of the Tu&Roth dataset: it achieved an accuracy score of 72.51%, which is 3.99% higher that got by the Tu&Roth method (Tu and Roth, 2011) (68.52%).

Method	Accuracy	F1+	F1-
DM	61.25	56.96	64.76
Tu&Roth Original	68.52	**75.36**	56.41
J48	**72.51**	74.73	**70.5**

Table 5: Results of applying different methods on the Tu&Roth dataset. DM: dictionary matching. Tu&Roth Original: the results of Tu and Roth (2011). J48: our model.

6 Discussion

The applied machine learning-based method extensively outperformed our dictionary matching baseline model, which underlines the fact that our approach can be suitably applied to LVC detection. As Table 4 shows, our presented method proved to be the most robust as it could obtain roughly the same recall, precision and F-score on the Wiki50 and SZPFX corpora. Our system's performance primarily depends on the applied candidate extraction method. In the case of dictionary matching, a higher recall score was primarily limited by the size of the dictionary, but this method managed to achieve a fairly good precision score.

As Table 5 indicates, the dictionary matching method was less effective on the Tu&Roth dataset. Since the corpus was created by collecting sentences that contain verb-object pairs with specific verbs, this dataset contains a lot of negative and ambiguous examples besides annotated LVCs, hence the distribution of LVCs in the Tu&Roth dataset is not comparable to those in Wiki50 or SZPFX. In this dataset, only one positive or negative example was annotated in each sentence, and they examined just the verb-object pairs formed with the six verbs as a potential LVC. However, the corpus probably contains other LVCs which were not annotated. For example, in the sentence *it have been held that a gift to a charity of shares in a close company* ***gave rise*** *to a charge to capital transfer tax where the company* ***had an interest*** *in possession in a trust*, the phrase *give rise* was listed as a negative example in the Tu&Roth dataset, but *have an interest*, which is another LVC, was not marked either positive or negative. This is problematic if we would like to evaluate our candidate extractor on this dataset since it would identify this phrase, even if it is restricted to verb-object pairs containing one of the six verbs mentioned above, thus yielding false positives already in the candidate extraction phase.

Moreover, the results got with our machine learning approach overperformed those reported in Tu and Roth (2011). This may be attributed to the inclusion of a rich feature set with new features like semantic or morphological features that was used in our system, which demonstrated a consistent performance on the positive and negative classes too.

To examine the effectiveness of each individual feature of the machine learning based candidate classification, we carried out an ablation analysis. Table 6 shows the usefulness of each individual feature type on the SZPFX corpus.

Feature	Precision	Recall	F-score	Diff
Statistical	60.55	55.88	58.12	-1.81
Lexical	71.28	28.6	40.82	-19.11
Morphological	62.3	54.77	58.29	-1.64
Syntactic	59.87	55.8	57.77	-2.16
Semantic	60.81	54.77	57.63	-2.3
Orthographic	63.3	56.25	59.56	-0.37
All	**63.29**	**56.91**	**59.93**	-

Table 6: The usefulness of individual features in terms of precision, recall and F-score using the SZPFX corpus.

For each feature type, we trained a J48 classifier with all of the features except that one. We then compared the performance to that got with all the features. As our ablation analysis shows, each type of feature contributed to the overall performance. The most important feature is the list of the most frequent light verbs. The most common verbs in a language are used very frequently in different contexts, with several argument structures and this may lead to the bleaching (or at least generalization) of its semantic content (Altmann, 2005). From this perspective, it is linguistically plausible that the most frequent verbs in a language largely coincide with the most typical light verbs since light verbs lose their original meaning to some extent (see e.g. Sanromán Vilas (2009)).

Besides the ablation analysis we also investigated the decision tree model yielded by our experiments. Similar to the results of our ablation analysis we found that the lexical features were the most powerful, the semantic, syntactic and orthographical features were also useful while statistical and morphological features were less effective but were still exploited by the model.

Comparing the results on the three corpora, it is salient that the F-score got from applying the methods on the Tu&Roth dataset was considerably better than those got on the other two corpora. This can be explained if we recall that this dataset applies a restricted definition of LVCs, works with only verb-object pairs and, furthermore, it contains constructions with only six light verbs. However, Wiki50 and SZPFX contain all LVCs, they include verb + preposition + noun combinations as well, and they are not restricted to six verbs. All these characteristics demonstrate that identifying LVCs in the latter two corpora is a more realistic

and challenging task than identifying them in the artificial Tu&Roth dataset. For example, the very frequent and important LVCs like *make a decision*, which was one of the most frequent LVCs in the two full-coverage LVC annotated corpora, are ignored if we only focus on identifying true LVCs. It could be detrimental when a higher level NLP application exploits the LVC detector.

We also carried out a manual error analysis on the data. We found that in the candidate extraction step, it is primarily POS-tagging or parsing errors that result in the omission of certain LVC candidates. In other cases, the dependency relation between the nominal and verbal component is missing (recall the example of objects with quantifiers) or it is an atypical one (e.g. `dep`) not included in our list. The lower recall in the case of SZPFX can be attributed to the fact that this corpus contains more instances of nominal occurrences of LVCs (e.g. *decision-making* or *record holder*) than Wiki50, which were annotated in the corpora but our morphology-based and extended syntax-based methods were not specifically trained for them since adding POS-patterns like `NOUN-NOUN` or the corresponding syntactic relations would have resulted in the unnecessary inclusion of many nominal compounds.

As for the errors made during classification, it seems that it was hard for the classifier to label longer constructions properly. It was especially true when the LVC occurred in a non-canonical form, as in a relative clause (*counterargument that can be made*). Constructions with atypical light verbs (e.g. *cast a glance*) were also somewhat more difficult to find. Nevertheless, some false positives were due to annotation errors in the corpora. A further source of errors was that some literal and productive structures like *to give a book (to someone)* – which contains one of the most typical light verbs and the noun is homonymous with the verb *book* "to reserve" – are very difficult to distinguish from LVCs and were in turn marked as LVCs. Moreover, the classification of idioms with a syntactic or morphological structure similar to typical LVCs – *to have a crush on someone* "to be fond of someone", which consists of a typical light verb and a deverbal noun – was also not straightforward. In other cases, verb-particle combinations followed by a noun were labeled as LVCs such as *make up his mind* or *give in his notice*. Since Wiki50 contains annotated examples for both types of MWEs, the classification of verb + particle/preposition + noun combinations as verb-particle combinations, LVCs or simple verb + prepositional phrase combinations could be a possible direction for future work.

7 Conclusions

In this paper, we introduced a system that enables the full coverage identification of English LVCs in running texts. Our method detected a broader range of LVCs than previous studies which focused only on certain subtypes of LVCs. We solved the problem in a two-step approach. In the first step, we extracted potential LVCs from a running text and we applied a machine learning-based approach that made use of a rich feature set to classify extracted syntactic phrases in the second step. Moreover, we investigated the performance of different candidate extraction methods in the first step on the two available full-coverage LVC annotated corpora, and we found that owing to the overly strict candidate extraction methods applied, the majority of the LVCs were overlooked. Our results show that a full-coverage identification of LVCs is challenging, but our approach can achieve promising results. The tool can be used in preprocessing steps for e.g. information extraction applications or machine translation systems, where it is necessary to locate lexical items that require special treatment.

In the future, we would like to improve our system by conducting a detailed analysis of the effect of the features included. Later, we also plan to investigate how our LVC identification system helps higher level NLP applications. Moreover, we would like to adapt our system to identify other types of MWE and experiment with LVC detection in other languages as well.

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).

References

Gabriel Altmann. 2005. Diversification processes. In *Handbook of Quantitative Linguistics*, pages 646–659, Berlin. de Gruyter.

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of MWE 2007*, pages 1–8, Morristown, NJ, USA. ACL.

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

Nicoletta Calzolari, Charles Fillmore, Ralph Grishman, Nancy Ide, Alessandro Lenci, Catherine MacLeod, and Antonio Zampolli. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of LREC 2002*, pages 1934–1940, Las Palmas.

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.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning*, 20(3):273–297.

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

Christiane Fellbaum, editor. 1998. *WordNet An Electronic Lexical Database*. The MIT Press, Cambridge, MA ; London, May.

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.

Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations*, 11(1):10–18.

Kate Kearns. 2002. *Light verbs in English*. Manuscript.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Annual Meeting of the ACL*, volume 41, pages 423–430.

Brigitte Krenn. 2008. Description of Evaluation Resource – German PP-verb data. In *Proceedings of MWE 2008*, pages 7–10, Marrakech, Morocco, June.

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.

István Nagy T., Veronika Vincze, and Gábor Berend. 2011. Domain-Dependent Identification of Multiword Expressions. In *Proceedings of the RANLP 2011*, pages 622–627, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Ross Quinlan. 1993. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of CICLing 2002*, pages 1–15, Mexico City, Mexico.

Tanja Samardžić and Paola Merlo. 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*, pages 52–60, Uppsala, Sweden, July. ACL.

Begoña Sanromán Vilas. 2009. Towards a semantically oriented selection of the values of Oper_1. The case of *golpe* 'blow' in Spanish. In *Proceedings of MTT 2009*, pages 327–337, Montreal, Canada. Université de Montréal.

Suzanne Stevenson, Afsaneh Fazly, and Ryan North. 2004. Statistical Measures of the Semi-Productivity of Light Verb Constructions. In *MWE 2004*, pages 1–8, Barcelona, Spain, July. ACL.

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, István Nagy T., and Gábor Berend. 2011. Multiword Expressions and Named Entities in the Wiki50 Corpus. In *Proceedings of RANLP 2011*, pages 289–295, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

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