a, citation and similar papers at core.ac.uk

brought to

provided by SZTE Publicatio Repozitórium - SZTE - Reposi

Learning to Detect English and Hungarian Light Verb Constructions

VERONIKA VINCZE, Hungarian Academy of Sciences
ISTVÁN NAGY T. and JÁNOS ZSIBRITA, University of Szeged

Light verb constructions consist of a verbal and a nominal component, where the noun preserves its original meaning while the verb has lost it (to some degree). They are syntactically flexible and their meaning can only be partially computed on the basis of the meaning of their parts, thus they require special treatment in natural language processing. For this purpose, the first step is to identify light verb constructions.

In this study, we present our conditional random fields-based tool—called FXTagger—for identifying light verb constructions. The flexibility of the tool is demonstrated on two, typologically different, languages, namely, English and Hungarian. As earlier studies labeled different linguistic phenomena as light verb constructions, we first present a linguistics-based classification of light verb constructions and then show that FXTagger is able to identify different classes of light verb constructions in both languages.

Different types of texts may contain different types of light verb constructions; moreover, the frequency of light verb constructions may differ from domain to domain. Hence we focus on the portability of models trained on different corpora, and we also investigate the effect of simple domain adaptation techniques to reduce the gap between the domains. Our results show that in spite of domain specificities, out-domain data can also contribute to the successful LVC detection in all domains.

Categories and Subject Descriptors: J.5 [**Computer Applications**]: Arts and Humanities

General Terms: Languages, Performance

Additional Key Words and Phrases: Conditional random fields, corpora, domain adaptation, English, Hungarian, light verb constructions, multiword expressions

ACM Reference Format:
Vincze, V., Nagy T., I., and Zsibrita, J. 2013. Learning to detect English and Hungarian light verb constructions. ACM Trans. Speech Lang. Process. 10, 2, Article 6 (June 2013), 25 pages.
DOI: http://dx.doi.org/10.1145/2483691.2483695

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 [Sag et al. 2002; Kim 2008; Calzolari et al. 2002]. They have recently come to the fore in the NLP research community [Rayson et al. 2010]. 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., *have lunch* or *give a try*). In several NLP applications such as information retrieval or event extraction it is important to identify LVCs in context, since they require special

This work was supported in part by the European Union and the European Social Fund through the project FuturICT.hu (grant TÁMOP-4.2.2.C-11/1/KONV-2012-2013).
Authors' addresses: V. Vincze, Research Group on Artificial Intelligence, Hungarian Academy of Sciences, 6720 Szeged, Tisza Lajos krt. 103, Hungary; I. Nagy T. and J. Zsibrita, Department of Informatics, University of Szeged, 6720 Szeged, Árpád tér 2, Hungary.
Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies show this notice on the first page or initial screen of a display along with the full citation. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, to redistribute to lists, or to use any component of this work in other works requires prior specific permission and/or a fee. Permissions may be requested from Publications Dept., ACM, Inc., 2 Penn Plaza, Suite 701, New York, NY 10121-0701 USA, fax +1 (212) 869-0481, or permissions@acm.org.
© 2013 ACM 1550-4875/2013/06-ART6 $15.00
DOI: http://dx.doi.org/10.1145/2483691.2483695

treatment, particularly because of their semantic features. Thus, LVC-detectors are needed to support these applications.

Identifying light verb constructions is not unequivocal because there are different approaches to undertake the task, for example, the linguistic phenomenon they seek to capture may differ from study to study (verb–object pairs vs. verbs with prepositional complements, for instance). As one of the main novelties of this study, we present a detailed characterization of the linguistic structures called light verb constructions in the literature and we also propose a linguistics-based classification for them, with the help of a test battery. This classification in turn makes it possible to place earlier works on LVC detection within a unified framework. In our experiments we focus on two typologically different languages. English is a subject-verb-object (SVO) language with strict word order and relatively poor morphology, whereas Hungarian is an agglutinative language with rich morphology, with subject-object-verb (SOV) as the preferred word order. Interlingual comparisons on data from these two languages may also enhance further studies on languages similar to these.

In order to investigate the domain specificity of LVCs, we made use of three corpora (SzegedParalellFX [Vincze 2012] and two newly constructed ones) for English and three subcorpora of the Szeged Treebank annotated for LVCs [Vincze and Csirik 2010] for Hungarian, which were all built by utilizing the same annotation principles. We experiment on the corpora by applying different settings in order to examine the portability of our models learnt on different corpora and also to see how the gap between the data from different domains can be reduced by domain adaptation. In these investigations, we use our newly constructed conditional random fields (CRF)-based tool, called FXTagger.

The main contributions of this study can be summarized as follows.

—We provide a *linguistics-based classification* of LVC phenomena, in which earlier approaches can be placed so as to reveal the similarities and differences between them.
—Besides using existing corpora, we created *two additional corpora* annotated for English LVCs, which are available to the community.
—In contrast to most of the previous studies, we do not just focus on verb-object pairs, but seek to identify LVCs that contain adpositional complements or nouns in an oblique case. Hence our goal is to identify a *broader range* of LVCs than previous studies did.
—We introduce our *conditional random fields*-based state-of-the-art *tool* for detecting LVCs, which makes use of contextual (shallow linguistic) features and is able to produce satisfactory results for all of the domains and languages used.
—We report our results for Hungarian and English corpora as well, which allows us to draw some conclusions on the *multilingual aspects* of LVC detection. Furthermore, to the best of our knowledge, ours is the first article to report results on *Hungarian* LVC detection.
—In our experiments we made use of three corpora for both languages. The corpora belong to different domains, namely short news, law, and newspaper texts. This selection of data makes it possible to compare the *domain-specific characteristics* of LVC detection in both languages.
—Here, we apply *domain adaptation* techniques in order to reduce the distance between domains in a setting where only limited annotated data is available for one of the domains. We report results for three domains in two languages, which enables us to make *cross-lingual comparisons* for each domain.

The remainder of the article is structured as follows. In Section 2 we examine the characteristics of LVCs from a linguistic point of view and describe our classification of

Table I. Tests for Differentiating Productive Constructions, LVCs and Idioms

		LVC		
Test	Productive	productive-like	idiom-like	Idiom
WH-word	YES	YES	NO	NO
Article	YES	YES	NO	NO
Plural	YES	YES	NO	NO
Negation	YES	YES	NO	NO
Possessor	YES	YES	NO	NO
Attributive	YES	YES	NO	NO
Coordination	YES	NO	NO	NO
Nominalization (V)	NO	NO	YES	NO
Nominalization (LVC)	YES	YES	NO	NO
Participle – 1	YES	YES	YES	YES
Participle – 2	YES	YES	NO	NO
Variativity	NO	YES	YES	NO
Changing the verb	YES	YES	NO	NO
Omitting the verb	NO	YES	YES	NO
English examples	make a cake	make a decision	make use	make a meal
Hungarian examples	kutyát tart	előadást tart	igényt tart	kordában tart
	"to have a dog"	"to have a presentation"	"to have a claim"	"to control"

LVC phenomena. Related work is presented in Section 3, then our corpora, methods, and results are elaborated in detail in Section 4. Next, results are discussed in Section 5, which is followed by an analysis of the most typical errors (Section 6). Results are finally summarized and some possible directions for future study are briefly mentioned.

2. THE CHARACTERISTICS OF LIGHT VERB CONSTRUCTIONS

Light verb constructions are verb and noun combinations where the semantic head of the construction is the noun, that is, the verb has lost its meaning to some extent and the noun is used in one of its original senses, but the verb functions as the syntactic head (the whole construction fulfills the role of a verb in the clause). They are usually distinguished from productive or literal verb + noun constructions on the one hand and idiomatic verb + noun expressions on the other (e.g., Fazly and Stevenson [2007], see also Table I), compare *make a cake* (productive), *make a decision* (LVC), and *make a meal* (idiom).

LVCs exhibit lexical and semantic idiosyncracies (to some extent). As for the former, the verbal component of the construction cannot be substituted by another verb with a similar meaning: instead of *make a decision* we cannot say **do a decision*. Still, the change of the noun for a word with a similar meaning does not yield the agrammaticality of the construction: *make a contract* and *make a treaty* are both acceptable constructions. Next, it should also be mentioned that there seem to be systematic cases where two LVCs share all of their meaning components, but their verbal components differ. Take, for instance, the following example.

Example 2.1. make/take a decision

With regard to semantic idiosyncracy, the meaning of LVCs can, at least partially, be computed from the meanings of their parts and the way they are connected. Although it is the noun that conveys most of the meaning of the construction, the verb itself cannot be viewed as semantically bleached (see e.g., Apresjan [2004], Alonso Ramos [2004], Sanromán Vilas [2009]), since it also adds important aspects to the meaning of the construction. For instance, (2.2) and (2.3) do not mean the same, although they describe the same situation.

Example 2.2. give help

Example 2.3. receive help

Nevertheless, it is interesting to examine whether LVCs are decomposable or not (on the decomposability of MWEs, see Sag et al. [2002]). If the parts of the LVC can be interpreted as having a special sense that is unique to this construction (i.e., there can be a word-to-word mapping between the lexical and the semantic level), it is called a decomposable LVC. One example of this is given here.

Example 2.4. to make progress
progress = 'progress'
make = 'perform'

The noun occurs in its usual sense (or in one of its usual senses), whereas the verb typically has a more abstract meaning of 'doing something' or 'performing some action' rather than keeping its original sense. Because of this, the meaning of the light verb construction can be 'doing something that is encoded in the meaning of the noun' (cf., Apresjan [2004]), thus, LVCs can be viewed as decomposable.

LVCs are syntactically flexible, that is, they can manifest themselves in a variety of forms: the verb may be inflected, the noun may occur in its plural form, and the noun may be modified. The nominal and the verbal components may not be adjacent in the sentence, as in the following example.

Example 2.5. The decision he took last time proved to be fatal.

The preceding points have some consequences for the NLP treatment of LVCs. Syntactic flexibility makes the automatic identification of LVCs difficult, especially in the case of agglutinative languages such as Hungarian. Lexical and semantic idiosyncracy can also affect the machine translation of the constructions: the nominal component being the semantic center of the construction seems to be constant across languages in the case of parallel constructions, hence it can be translated literally, whereas the verb can be determined only lexically, that is, in dictionaries.

2.1. Light Verb Constructions In Hungarian

In order to understand the special features of identifying Hungarian LVCs, a brief description of the Hungarian language is required. 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, for example, new (or emphatic) information (focus) always precedes the verb and old information (topic) precedes the 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. As for verbs, they are inflected for person and number and the definiteness of the object.

The canonical form of a Hungarian light verb construction is a bare noun + third person singular verb. Due to the features above, they may occur in noncanonical versions as well: the verb may precede the noun, or they may be not adjacent, moreover, the

Table II. True Light Verbs and Vague Action Verbs in English

Test	Vague action verb	True light verb
Passivization	YES	NO
WH-movement	YES	NO
Pronominalization	YES	NO
Indefinite NP	NO	YES
NP stem is identical to a verb	NO	YES
Differences compared to verbal counterpart	NO	YES
Examples	make an inspection	give a groan

verb may occur in different surface forms inflected for tense, mood, person, and number. These issues will be considered when implementing our system for identifying Hungarian LVCs.

2.2. Types of Light Verb Constructions

Vincze [2011] presents a test battery that is able to differentiate among different types of verb + noun combinations: productive constructions, LVCs, and idioms. Two tests, namely the tests of variativity and omitting the verb play the most significant role in distinguishing LVCs from productive constructions and idioms. Variativity reflects the fact that LVCs can often be substituted by a verb derived from the same root as the nominal component within the construction: productive constructions and idioms can rarely be substituted by a single verb, even if so, there is no morphological relation between the noun and the verbal counterpart. Omitting the verb exploits the fact that it is the nominal component that mostly bears the semantic content of the LVC, hence the event denoted by the construction can be determined even without the verb in most cases. Both the noun and the verb play a key role in computing the meaning of productive constructions, while the original senses of the noun and the verb are not relevant at all as regards the meaning of an idiomatic verb + noun combination. Thus, the noun itself is not sufficient to compute the meaning of either productive or idiomatic constructions.

The other tests help us to distinguish two types of LVCs. Productive-like LVCs behave rather like productive constructions, whereas idiom-like constructions are more similar to idioms. Still, there is no sharp and distinct boundary between the groups, since belonging to a subgroup is not determined by a dichotomy of the either-or type: the place of the construction on a scale is rather a question of degree and scalability, which is true for English and Hungarian as well [Vincze 2011]. Table I states the applicability of the tests for each type, and these tests were used in annotating the corpora presented in Section 4.1.

Krenn [2008] provides some diagnostic tests for distinguishing between German idioms and LVCs. As for English, Kearns [2002] distinguishes between two subtypes of what is traditionally called light verb constructions. True light verb constructions such as *to give a wipe* or *to have a laugh* and vague action verbs such as *to make an agreement* or *to do the ironing* differ in some syntactic and semantic features and can be separated by various tests, for example, passivization, WH-movement, pronominalization, and so on, as shown in Table II. True light verb constructions roughly correspond to idiom-like LVCs in Vincze's [2011] classification, whereas vague action verbs are similar to productive-like constructions. Examples for the above types of light verb constructions can be seen in Figure 1.

From a morphological perspective, LVCs can also be divided into groups. First, perhaps the most common type is when the nominal component is the object of the verb, that is, it bears an accusative case in Hungarian. Second, the nominal component can

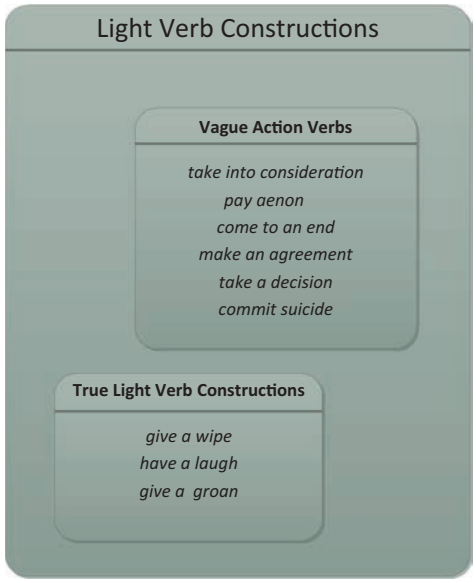

Fig. 1. Types of LVCs based on syntactic and semantic criteria.

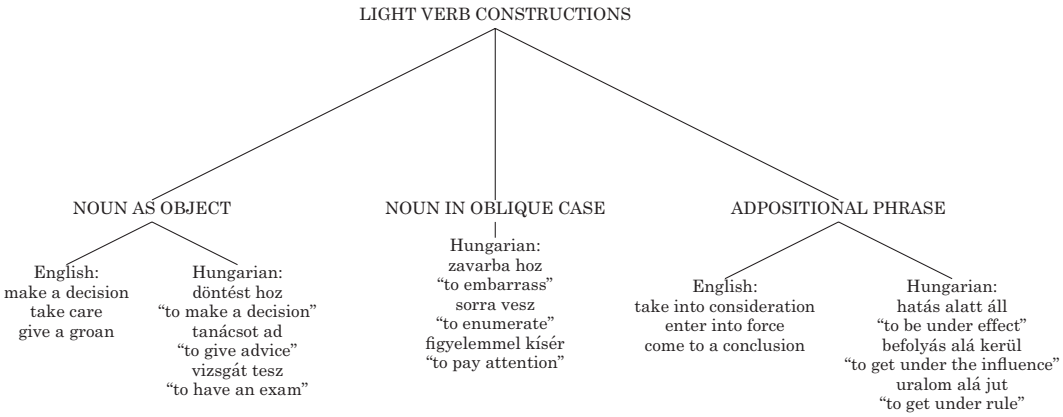

Fig. 2. Types of LVCs from a morphological point of view.

bear other (oblique) cases as well in Hungarian. (This option is not viable in English due to the lack of oblique morphological cases.) Third, a prepositional or postpositional phrase can also occur in the construction. Figure 2 presents this classification with illustrative examples.

LVCs may occur in several forms due to their syntactic flexibility. Besides the prototypical verb + noun combination in English and the noun + verb combination in Hungarian, they can have a participial form (e.g., *photos taken*) and may also undergo nominalization, yielding a nominal compound (e.g., *decision maker*). In split LVCs (e.g., *a decision which has been recently made*) the noun and the verb may be situated far from each other in the sentence, so their identification requires going beyond clause boundaries.

3. RELATED WORK

In this section we present related work on detecting LVCs, and then describe corpora annotated for LVCs.

3.1. Approaches to Identifying Light Verb Constructions

There are two basic approaches to identifying LVCs. In the first approach, several studies attempted to classify LVC candidates, which means that they extracted LVC candidates (usually verb-object pairs including one verb from a well-defined set of 3–10 verbs) from texts and then they applied different methods to decide whether they are 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]. In the second approach, other studies identified LVCs in running texts, having taken contextual information into account [Diab and Bhutada 2009; Tu and Roth 2011; Vincze et al. 2011a; Nagy T. et al. 2011]. While the first approach assumes that a specific candidate is an LVC or not, 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, due to structural or morphological homonymies.[1] "Compare *the government will make decisions on foreign policy issues* vs. *they will make decisions taken by the government publicly available* or *számba vettem a lehetőségeket* (consideration-ILL take-PAST-1SGOBJ the possibility-PL-ACC) "I considered the possibilities" vs. *számba vettem a nyalókát* (mouth-1SGPOSS-ILL take-PAST-1SGOBJ the lollipop-ACC) "I put the lollipop into my mouth," where the first occurrences of *make decisions* and *számba vettem* are LVCs, whereas the second ones are not. In this article we identify LVCs in running text, that is, we follow the second approach and carry out a token-based identification of LVCs instead of a type-based one. In other words, we decide whether the given sequence of words is an LVC within its context or not.

3.2. Methods for Identifying Light Verb Constructions

There are several applications developed for identifying MWEs and LVCs, which can be classified according to the methods they apply [Piao et al. 2003; Dias 2003]. First, statistical models rely on word frequencies, co-occurrence data and contextual information to decide whether a bigram or trigram (or even an n-gram, that is, a sequence of words) can be considered a multiword expression or not, see for example, Bouma [2010], Villavicencio et al. [2007]. Statistical systems can be easily adapted to other languages and other types of multiword expressions, but they are not able to identify rare multiword expressions, which is the main drawback of these methods, as about 70% of multiword expressions occur only once or twice in a large corpus [Piao et al. 2003; Vincze 2011]. As for LVC detection, Stevenson et al. [2004], Fazly and Stevenson [2007], Van de Cruys and Moirón [2007], and Gurrutxaga and Alegria [2011] built their system on statistical features, among others. Stevenson et al. [2004] focused on deciding whether true LVC candidates[2] containing the verbs *make*, *take,* or *give* are acceptable or not. Fazly and Stevenson [2007] used linguistically motivated statistical measures to distinguish subtypes of verb + noun combinations. Van de Cruys and Moirón [2007] described a semantic-based method for identifying verb-preposition-noun combinations

[1] An intermediate solution is that of `mwetoolkit` [Ramisch et al. 2010b, 2010a], which provides a list of MWEs extracted from texts. Hence MWE candidates that occur at least once as an MWE within the text are treated as MWEs, however, non-MWE uses of the same unit are ignored.

[2] Although some of the authors of papers cited here may not explicitly use the term *true LVC*, we decided to apply this term wherever it was unequivocal on the basis of the examples and the corpora that their research was restricted to true LVCs. In this way, we would like to underline the subtle differences in the approaches taken to identifying LVCs.

in Dutch. Their method relied on selectional preferences for both the noun and the verb, and they also utilized automatic noun clustering when considering the selection of semantic classes of nouns for each verb. Gurrutxaga and Alegria [2011] extracted idiomatic and light verb noun + verb combinations from Basque texts by employing statistical methods. Since Basque is a free word-order language, they hypothesized that a wider window would yield more significant cooccurrence statistics, but their initial experiments did not confirm this.

Other studies employ rule-based systems in LVC detection [Diab and Bhutada 2009; Nagy T. et al. 2011; Vincze et al. 2011a; Sinha 2011], which are usually constructed on the basis of (shallow) linguistic information. Diab and Bhutada [2009] used a supervised system for classifying verb-noun combinations as literal or idiomatic in context. Vincze et al. [2011a] exploited shallow morphological features for identifying LVCs in English texts, while the domain specificity of the problem was highlighted in Nagy T. et al. [2011]. Sinha [2011] found that linguistic-based information can help when identifying Hindi multiword expressions in an English–Hindi parallel corpus.

Some hybrid systems make use of both statistical and linguistic information as well [Dias 2003; Tan et al. 2006; Bannard 2007; Cook et al. 2007; Tu and Roth 2011; Samardžić and Merlo 2010], which results in better recall scores. Dias [2003] presents a system which is based on word statistics and information from POS-tagging and syntactic parsing. Tan et al. [2006] tried to identify true light verb constructions by applying machine learning techniques. They found that in this task it is especially the random forest classifier that can efficiently combine statistical and linguistic features. Bannard [2007] sought to identify verb and noun constructions in English on the basis of syntactic fixedness. He examined whether the noun can have a determiner or not, whether the noun can be modified, and whether the construction can have a passive form, which features were exploited in the identification of the constructions. Cook et al. [2007] differentiated between literal and idiomatic uses of verb and noun constructions in English. Their basic hypothesis was that the canonical form of each construction occurs mostly in idioms, since they show syntactic variation to a lesser degree than constructions in literal usage. Samardžić and Merlo [2010] analyzed English and German LVCs in parallel corpora: they paid special attention to their manual and automatic alignment. They found that linguistic features (i.e., the degree of compositionality) and the frequency of the construction both have an impact on aligning the constructions. Tu and Roth [2011] classified verb + noun object pairs as being LVCs or not by using a Support Vector Machine. They employed both contextual and statistical features and concluded that on ambiguous examples, local contextual features perform better.

Linguistics-based or hybrid methods may be highly language-dependent because of the amount of encoded linguistic rules, so it is costly to adapt them to different languages or even to different types of multiword expressions. Still, the combination of different methods may improve the performance of systems for LVC detection [Pecina 2010].

As for Hungarian, we are aware of one system that identifies multiword verbs (LVCs and idioms), however, it does not make a distinction between the two classes. Sass [2010] developed a method for extracting multiword verbs from parallel corpora. By aligning the verbs in parallel clauses, a complex verb is produced and their arguments are marked with tags denoting the language which they come from. From these representations the original algorithm is able to detect the multiword verbs for each language of the parallel corpus, along with cases where a multiword verb corresponding to a single word verb in the other language can also be extracted.

Although most of the previous studies focus only on LVCs where the noun functions as the object of the verb [Stevenson et al. 2004; Tan et al. 2006; Fazly and Stevenson

2007; Cook et al. 2007; Bannard 2007; Tu and Roth 2011], as the prepositional object [Van de Cruys and Moirón 2007; Krenn 2008] or only true light verb constructions are considered [Stevenson et al. 2004; Tan et al. 2006; Tu and Roth 2011], we—in line with Vincze et al. [2011a] and Nagy T. et al. [2011]—seek to identify all types of LVCs in our study and do not restrict ourselves to certain types of LVCs. On the other hand, some earlier work [Cook et al. 2007; Diab and Bhutada 2009; Sass 2010] just distinguished between the literal and idiomatic uses of verb + noun combinations. Here we argue that it is important to separate LVCs and idioms because LVCs are semiproductive and semicompositional—which may be exploited in applications such as machine translation or information extraction (see, e.g., Apresjan and Tsinman [2002])—in contrast to idioms, which have neither feature.

3.3. Related Corpora and Databases

In order to identify LVCs in texts, well-designed and tagged corpora are invaluable for training and testing algorithms. An Estonian database and a corpus of multiword verbs was constructed [Kaalep and Muischnek 2006, 2008, Muischnek and Kaalep 2010]), and Krenn [2008] developed a database of German PP-verb combinations. The Prague Dependency Treebank was also annotated for multiword expressions [Bejcek and Stranák 2010], thus for LVCs, too [Cinková and Kolářová 2005]. For Portuguese, Hendrickx et al. [2010] created an annotated corpus of complex predicates (i.e., multiword verbs), and Sanches Duran et al. [2011] analyzed complex predicate candidates extracted from a Brazilian Portuguese corpus using the mwetoolkit [Ramisch et al. 2010b]. NomBank [Meyers et al. 2004] contains the argument structure of common nouns, paying attention to those occurring in LVCs as well. Literal and idiomatic uses of English verb + noun combinations are annotated in the VNC-Tokens dataset [Cook et al. 2008]. In the Wiki50 corpus, several types of multiword expressions (including LVCs) are marked [Vincze et al. 2011b]. The corpus used in the experiments of Tu and Roth [2011] is also publicly available, which contains true light verb constructions. As for Hungarian, an annotated corpus and a database containing LVCs are described in Vincze and Csirik [2010] and an English–Hungarian annotated parallel corpus of LVCs was recently published [Vincze 2012]. Nevertheless, as already presented in Section 2.2, these corpora may treat the notion of LVC differently, so their annotation principles may differ from each other.

4. EXPERIMENTS

In this section we present our corpora, our methodology for detecting LVCs, and we show our results.

4.1. Corpora

In our experiments we made use of three corpora for both English and Hungarian, which are described below. When choosing the texts, we kept in mind the fact that the same domains would be employed for both languages: we selected texts from the domains called newspaper, short news, and law, so interlingual comparisons across domains could be made as well.

The SzegedParalellFX corpus contains parallel texts in English and Hungarian taken from various domains [Vincze 2012]. For our purposes, we selected the English versions of texts from bilingual magazines.

The JRC-Acquis Multilingual Parallel Corpus consists of legislative texts for a range of languages used in the European Union [Steinberger et al. 2006]. For this study, we randomly selected 60 documents from the English version of the corpus and annotated LVCs in them.

Table III. Statistical Data on the Corpora. VERB: Verbal Occurrences. PART: Participial LVCs. NOM: Nominal LVCs. SPLIT: Split LVCs

			VERB		PART		NOM		SPLIT		
Corpus	Sentences	Tokens	#	%	#	%	#	%	#	%	Total
SzegedParalellFX	5,760	115,621	354	67.7	55	10.5	31	5.9	83	15.9	523
JRC-Acquis	5,619	103,963	204	41.9	157	32.2	24	4.9	102	21.0	487
CoNLL-2003	8,467	107,620	235	59.2	83	20.9	16	4.0	63	15.9	381
SzT newspaper	10,210	223,286	453	58.8	198	25.7	55	7.1	65	8.4	771
SzT law	9,278	258,722	629	27.9	672	29.9	714	31.8	234	10.4	2249
SzT short news	9,574	227,239	563	40.3	700	50.1	92	6.6	43	3.0	1398

In addition, we annotated LVCs in 500 randomly selected pieces of short news from the CoNLL-2003 dataset originally developed for named entity recognition [Tjong Kim Sang and De Meulder 2003].

As for the Hungarian corpora, they form part of the Szeged Treebank annotated for LVCs [Vincze and Csirik 2010]. Among the subcorpora, the domains of law, short business news, and newspaper texts were selected for the purpose of this study.

The newspaper texts contain one- or two-page-long articles from newspapers and magazines, involving various topics and types of texts, for example, interviews, reviews, or analyses. On the other hand, the short news domain contains short pieces of news (each consisting of only a couple of sentences) and the topic of the news is also more restricted than the case for newspaper texts: It is mostly news of politics, finance, or sport that occur in the English short news corpus, and it is mostly finance or economy that can be found in the Hungarian one. Thus, newspaper texts and short news have different stylistic characteristics, and we treat them as separate domains in our investigations.

The corpora were annotated by three independent linguists, who are native speakers of Hungarian and could speak English at an advanced level. They were instructed to annotate all occurrences of LVCs, and they marked LVCs according to their grammatical category (i.e., verbal, participial, or nominal occurrences). About 1000 sentences containing 174 LVCs were annotated by all the annotators, hence the interagreement rates could be calculated, which were 0.8381/0.7356/0.7815 and 0.7867/0.7117/0.7423 on average in terms of precision, recall, and F-score, and 0.7172 and 0.6778 in κ-measure for English and Hungarian, respectively.

Statistical data on the corpora can be seen in Table III. All the corpora were annotated on the basis of the test battery described in Section 2.2, but no subtypes of LVCs are distinguished (i.e., vague action verbs and true light verb constructions are annotated in the same way), as we are not aware of any higher-level application that can profit from the latter distinction. Thus, we followed Vincze's [2011] tests, but we neglected those of Kearns' [2002]. In this way, like Nagy T. et al. [2011] and Vincze et al. [2011a], we seek to identify all types of LVCs, as opposed to earlier studies that focused only on true light verb constructions [Tu and Roth 2011], verb-object pairs [Fazly and Stevenson 2007; Bannard 2007; Gurrutxaga and Alegria 2011], or verb-preposition-noun triplets [Van de Cruys and Moirón 2007; Krenn 2008].

As seen in Table III, each corpus contains annotations for participial, nominal, and split occurrences of LVCs as well. However, we focus only on verbal occurrences of LVCs due to the sparsity of data on nominal, participial, and split LVCs in some of the corpora. On the other hand, due to some orthographical rules of Hungarian, in some cases the participial and nominal occurrences of LVCs are spelt as one compound word, for example, *tanácsadó* (*tanács*+*adó* advice+giver "someone who gives advice"). The identification of such cases would require a significantly different approach, that is, deep

Table IV. Length of LVCs and LVC Lemmas Including or Excluding Prepositions and Articles

	English LVCs						English LVC lemmas					
Token	SzPFX		JRC-Acquis		CoNLL-2003		SzPFX		JRC-Acquis		CoNLL-2003	
length	#	%	#	%	#	%	#	%	#	%	#	%
2	99	27.97	42	20.59	67	28.51	76	35.19	26	30.59	53	30.64
3	151	42.65	110	53.92	97	41.28	130	60.19	49	57.65	113	65.32
4≤	104	29.38	52	25.49	71	31.21	10	4.63	10	11.76	7	4.05
sum	354	100.00	204	100.00	235	100.00	216	100.00	85	100.00	173	100.00
	English LVCs filtered						English LVC lemmas filtered					
Token	SzPFX		JRC-Acquis		CoNLL-2003		SzPFX		JRC-Acquis		CoNLL-2003	
length	#	%	#	%	#	%	#	%	#	%	#	%
2	203	57.34	139	68.14	104	44.26	213	98.61	84	98.82	167	96.53
3	115	32.49	46	22.55	103	43.83	3	1.39	1	1.18	6	3.47
4≤	36	10.17	19	9.31	28	11.91	0	0.00	0	0.00	0	0.00
sum	354	100.00	204	100.00	235	100.00	216	100.00	85	100.00	173	100.00
	Hungarian LVCs						Hungarian LVC lemmas					
Token	SzT newspaper		SzT law		SzT short news		SzT newspaper		SzT law		SzT short news	
length	#	%	#	%	#	%	#	%	#	%	#	%
2	412	90.95	588	93.48	502	89.17	236	98.74	165	98.80	221	93.64
3	27	5.96	23	3.66	49	8.70	2	0.84	2	1.20	15	6.36
4≤	14	3.09	18	2.86	12	2.13	1	0.42	0	0.00	0	0.00
sum	453	100.00	629	100.00	563	100.00	239	100.00	167	100.00	236	100.00

Table V. Statistical Data on LVCs in the Corpora

Corpus	Verbal LVCs	Lemmas	Occ. of lemmas
SzegedParalellFX	354	216	1.64
JRC-Acquis	204	85	2.40
CoNLL-2003	235	173	1.36
SzT newspaper	453	238	1.90
SzT law	629	167	3.77
SzT short news	563	236	2.29

morphological analysis of Hungarian compounds; in order to compare our system's performance on English and Hungarian data, we neglected such cases. Furthermore, the detection of split LVCs would require a more refined syntactic/semantic analysis such as coreference resolution and treating long-distance movements, which we plan to tackle in a future study.

Table IV includes some statistics on the length of LVCs. A typical example of a two-token LVC is *take care*, one for a three-token-long is *take a decision* and a four-token LVC is *come to a conclusion*. In order to minimize the typological differences between the two languages, we also calculated the length of LVCs and LVC lemmas for English with prepositions and articles omitted, and it was shown that similar to Hungarian, most of the LVC lemmas contain only two words.

4.2. Light Verb Constructions in the Corpora

In order to confirm the domain-specificity of detecting LVCs, we carried out a detailed data analysis on the LVCs occurring in the corpora. First, LVCs were gathered from the corpora and lemmatized and the frequency of each lemma was calculated. Data is presented in Table V and Tables VI and VII list the most frequent LVCs in each corpus. As can be seen, the distribution of LVCs in the corpora varies somewhat: the top 10 LVCs are responsible for only 17.6% and 25.7% of the LVC occurrences in the

Table VI. The Most Frequent English LVCs

	Paralell		JRC		CoNLL	
1.	take place	25	enter into force	27	take place	7
2.	play a role	17	take into account	18	give detail	6
3.	give a concert	9	take account	12	play a game	6
4.	take a look	7	meet the requirements	11	catch fire	4
5.	take part	7	take place	9	fall short	3
6.	spend time	6	take measure	7	have an impact	3
7.	have an effect	5	carry out an activity	5	make a debut	3
8.	make a debut	5	play a role	5	play cricket	3
9.	pay attention	5	deliver an opinion	4	take a step	3
10.	take care	5	give a judgment	4	take part	3

Table VII. The Most Frequent Hungarian LVCs

	SzT newspaper		SzT law		SzT short news	
1.	részt vesz "to take part"	31	sor kerül "the time has come"	109	nyilvánosságra hoz "to publish"	40
2.	sor kerül "the time has come"	14	lehetőséget ad "to offer a possibility"	37	hírül ad "to make a report"	38
3.	őrizetbe vesz "to take into custody"	11	szerződést köt "to make a contract"	31	ajánlatot tesz "to make an offer"	28
4.	szerződést köt "to make a contract"	10	sor kerül "the time has come"	29	tárgyalást folytat "to conduct a negotiation"	18
5.	szert tesz "to get access"	8	eleget tesz "to fulfill"	23	szerződést köt "to make a contract"	13
6.	lehetőséget ad "to offer a possibility"	7	forgalomba hoz "to put into circulation"	19	megállapodást köt "to make an agreement"	11
7.	támogatást kap "to receive support"	7	határozatot hoz "to make a verdict"	17	megbízást ad "to give an assignment"	8
8.	döntést hoz "to take a decision"	6	nyilvánosságra hoz "to publish"	17	döntést hoz "to take a decision"	7
9.	helyet kap "to get space"	6	igényt tart "to have a claim"	15	eleget tesz "to fulfill"	7
10.	igénybe vesz "to take up"	6	részt vesz "to take part"	15	feljelentést tesz "to make an accusation"	7

CoNLL-2003 and SzegedParalellFX corpora, respectively, while this value is 50% in the JRC-Acquis corpus. As for the Hungarian case, the situation is similar: the 10 most frequent LVCs represent 49.5% of the LVCs in the law subcorpus, whereas it is only 31.4% and 23.4% in the short news and newspaper subcorpora, respectively.

We also investigated the extent to which the corpora overlap, that is, how many LVCs occur in each corpus or in at least two of the corpora. The Dice and Jaccard distances between the corpora were also calculated on the basis of the union and intersection of the LVCs found in the corpora. Table VIII shows these values. We only found 11 LVCs that occur in each of the English corpora and 28 that occur in each of the Hungarian corpora, which aptly underlines the domain-specificity of the problem, namely, different corpora contain different LVCs.

With the corpora at hand, we were able to examine the proportion of LVC and non-LVC uses of some specific LVC candidates. For instance, the phrase *tárgyalást folytat* (negotiation-ACC continues) usually means "to conduct a negotiation," which is an LVC, but in certain contexts, it can mean "to continue a(n ongoing) negotiation," which is not an LVC. In the corpora, there are 13 LVC uses and 1 non-LVC use. However, the

Table VIII. Distance Between the Corpora

Corpora	Intersection	Dice	Jaccard
JRC-CoNLL	18	0.1395	0.9250
JRC-Paralell	17	0.1130	0.9400
Paralell-CoNLL	27	0.1388	0.9254
SzT law-SzT news	41	0.2035	0.8867
SzT law-SzT paper	52	0.2568	0.8180
SzT paper-SzT news	73	0.3080	0.8527

sequence *megbeszélést tart* (meeting-ACC holds) "to have a meeting"—which can also be considered an LVC (out of context)—occurs only once in the corpus, and in a non-LVC use: *megbeszélést tart célszerűnek* (meeting-ACC holds necessary-DAT) "he thinks that a meeting is required". Thus, non-LVC usage of LVC-candidates is not very frequent, but the corpora contain some examples.

4.3. Methodology

For the automatic identification of LVCs in corpora, we implemented a machine learning approach, which we elaborate upon below.

The MALLET implementations [McCallum 2002] of the first-order linear chain Conditional Random Fields (CRF) classifier [Lafferty et al. 2001] were utilized for this. To apply other popular machine learning methods (like SVM or decision tree) to identify LVCs in running text, positive and negative annotated examples are required, as we found in the Tu and Roth dataset. However, the corpora were only annotated for the LVCs in the running texts so negative examples were not available, which excluded the use of the other methods. As we only focus on identifying verbal LVCs in running text, we consider this problem as a sequence labeling problem. As Table IV shows, in the case of English, most of the verbal LVCs are bigrams or trigrams. The CRF approach is able to handle this kind of problem.

Our tool called FXTagger[3] is based on a general Named Entity feature set [Szarvas et al. 2006], with the following categories: *orthographical features:* capitalization, word length, bit information about the word form (contains a digit or not, has an uppercase character inside the word, etc.), character-level bi/trigrams, suffixes; *dictionaries* of first names, company types, denominators of locations; *frequency information:* frequency of the token, the ratio of the token's capitalized and lowercase occurrences, the ratio of capitalized and sentence beginning frequencies of the token, what was derived from the Gigaword dataset;[4] *shallow linguistic information:* part of speech; *contextual information:* sentence position, trigger words (the most frequent and unambiguous tokens in a window around the word) from the training database and the word between quotes.

The basic feature set was implemented for named entity recognition. Since LVCs never contain named entities, these features may also contribute to performance; however, we extended this basic feature set with LVC-specific features. Some features were language dependent, for instance, we added language-specific *lists of light verbs* to the dictionaries. The light verb lists exploit the fact that the most common verbs are typically light verbs. Hence the 15 most frequent verbs were collected from the English corpora and 25 from the Hungarian ones. In the case of English, the *LVC list* contains the lemmatized LVCs of the Wiki50 corpus (287 items). In the case of Hungarian, LVCs from the subcorpora of the SzegedCorpus that were not used in our experiments were

[3]The tool and the annotated corpora are available at http://www.inf.u-szeged.hu/rgai/lvc.
[4]Linguistic Data Consortium (LDC), catalogId: LDC2003T05.

included in the LVC list (578 items). We used it as a binary feature whether or not the LVC candidate occurred in the lists.

We also extended the English feature list with a *Prediction List*. We trained a CRF classifier with our LVC specific features on the Wiki50 corpus [Vincze et al. 2011b] and extracted potential LVCs from 10,000 Wikipedia pages. We created the Prediction List from the most frequent LVCs. This list contains 424 different potential LVCs and it was investigated whether the LVC candidate occurred on the list (binary feature).

The shallow linguistic features were extended with the *POS-pattern*, *SubPOS*, *VerbalStem* and *Syntax* features. If the POS-tag sequence in the text matched one pattern typical of LVCs (e.g., `VB NN`), the sequence tags were marked as *true*, otherwise as *false*. For English POS-tagging, we applied the Stanford POS-tagger [Toutanova and Manning 2000], and `magyarlanc` was used for Hungarian [Zsibrita et al. 2010]. As Hungarian is a morphologically rich language, we selected those morphological features that seemed to play an important role in determining whether an LVC candidate is a genuine LVC in context or not, and unnecessary features were deleted from the representation. For instance, the number and person features of a verb are irrelevant for LVC detection and thus were neglected (*SubPOS*).

The *VerbalStem* binary feature focuses on the stem of the noun. In the case of 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, that is, it coincided with a stem of a verb.

Syntactic information—provided by the Stanford parser [Klein and Manning 2003]—can also be exploited in identifying LVCs: the dependency label between the noun and the verb was added as a feature. In Hungarian, we made use of the dependency labels found in the Szeged Dependency Treebank [Vincze et al. 2010]. When applying the Stanford POS Tagger [Toutanova and Manning 2000], the *stems* and *lemmas* of the words were also used as a feature.

We extended the orthographical features with the *Suffix* feature, that is, it was checked whether the lemma of the noun ends in a given character bi- or trigram. It exploited the fact that many nominal components in LVCs are derived from verbs.

The *productDeriv* feature was used to detect nonproductive derivations in Hungarian in the case of those nouns that were derived historically from a verb but the derivational suffix is no longer considered productive.

In addition, we also specified the *other entities in the sentence*, like named entities (NEs) and noun compounds, which were also used as features. We employed the Stanford Named Entity Recognition tool [Finkel et al. 2005] and detected noun compounds, following the methods of Nagy T. et al. [2011].

We trained the first-order linear chain CRF classifier with the feature set above, and evaluated it on the English and Hungarian corpora in a 10-fold cross-validation setting at the document level. We trained the CRF models with the default settings in Mallet for 200 iterations or until convergence was reached.

To compare the performance of our system with others, we evaluated it—with the necessary modifications (e.g., detecting only true light verb constructions)—on the Tu and Roth dataset [Tu and Roth 2011] also. This dataset contains 2,162 sentences with verb-object pairs formed with the verbs *do*, *get*, *give*, *have*, *make*, and *take* (1,039 positive and 1,123 negative examples). Our methods can achieve an accuracy score of 73.93%, which is 5.41% higher than the one achieved by the Tu and Roth method [Tu and Roth 2011] (68.52%). We also evaluated our modified feature set with a Support Vector Machine (SVM) learner on this dataset which gave an accuracy score of 73.11%, which is 4.59% higher than the Tu and Roth's result, but 0.82% lower than our CRF-based method.

Table IX. Utility of Individual Features in Hungarian for Recall, Precision, and F-Score

Feature	Recall	Precision	F-score	Diff
Dictionary labeling	20.81	45.56	28.57	–
Base features	42.73	73.25	53.98	–
All features	60.82	79.58	68.94	–
LVC Lists	55.32	76.10	66.46	−2.48
POS-pattern	58.33	79.66	67.35	−1.59
SubPos	57.98	78.42	66.67	−2.27
Syntax	59.04	78.35	67.34	−1.60
Stem	60.28	77.63	67.86	−1.08
Suffix	57.62	78.50	66.46	−2.48
VerbalStem	60.11	79.85	68.48	−0.46
productDeriv	60.06	80.05	68.62	−0.32
RB Prediction	59.40	79.76	68.09	−0.85

Table X. Utility of Individual Features in English for Recall, Precision, and F-Score

Feature	Recall	Precision	F-score	Diff
Dictionary labeling	7.65	69.23	13.79	–
Base features	28.39	47.86	35.64	–
All features	50.85	71.43	9.41	–
LVC Lists	48.73	70.12	57.50	−1.91
Prediction List	47.03	68.94	55.92	−3.49
POS-pattern	46.19	70.78	55.90	−3.51
VerbalStem	41.95	68.75	52.11	−7.30
Syntax	40.25	64.63	49.61	−9.8
Stem	42.37	62.89	50.63	−8.78
Suffix	49.58	72.22	58.79	−0.62
Other entities	46.61	68.75	55.56	−3.85

In order to examine the effectiveness of each individual feature, we carried out an ablation analysis. Tables IX and X tell us how useful the individual features are for both languages. The performance scores of the features were compared with that obtained by applying all features described in our article. In the case of Hungarian, the CRF classifier was trained on the Szeged Treebank short news corpus. In the case of English, we performed another ablation study on the CoNLL-2003 corpus. That is, for each LVC specific feature, we trained a CRF classifier with all of the features except that one. We then compared the performance to that obtained with all the features.

In the case of Hungarian, *LVC lists* and the *Suffix* feature were the most useful: the lack of these features led to the lowest result. Part-of-speech-related features were also important, especially the detailed morphological information (*SubPos*). The other features seemed to have a lower impact on the results, but were still effective. In the case of English, the *Syntax*, *Stem*, and *VerbalStem* features were the most useful. However, the features *Suffix*, and *LVC list* were less effective, but still contributed to the overall performance.

As evaluation metrics, we employed $F_{\beta=1}$ scores. As we only identify verbal LVCs in running texts, we applied a phrase-based evaluation of LVCs. The training dataset was in IOB format, where B-VERBFX labels the first word of an LVC, I-VERBFX labels all other subsequent words which are part of the LVC, and O labels nonentities. In our case, the labeling of LVCs was only accepted if all of its members were labeled correctly and no other neighboring words were marked (*true positive*, *TP*). We consider it a *false*

negative (FN) example when there was an LVC entity in the running text, but the system could not correctly recognize it. In other words, the system could notice that there was an LVC but got its boundaries wrong or there was an entity but the system missed it. In the case of *false positives (FP)*, there was no LVC in the text but the system hypothesized one. To calculate F_1-scores we define precision and recall as follows:

$$Precision = \frac{TP}{TP + FP} \quad \text{and} \quad Recall = \frac{TP}{TP + FN}.$$

And the F_1 score is the harmonic mean of the precision and recall:

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall}.$$

In the pure in-domain setting, we performed a 10-fold cross-validation at the document level on each corpus (TARGET).

To compare the different domains, we used a pure cross-domain (CROSS) setting where our model was trained on the source domain and evaluated on the target (i.e., no labeled target domain datasets were used for training); for example, we trained the model on SzegedParalellFX and tested it on JRC-Acquis.

As three different domains were available for both languages, we examined how domain adaptation could enhance the results if we only have a limited amount of annotated target data. Domain adaptation is especially useful when there is only a limited amount of annotated data available for one domain, but there is plenty of data for another domain. Using the domain with a lot of annotated data as the source domain and a domain with limited data as the target domain, domain adaptation techniques can successfully contribute to the learning of a model for the target domain (see, e.g., Daumé III [2007]). A very simple approach was used for domain adaptation: the training dataset was extended with sentences from the target. First, we extended the training dataset with 500 target sentences, then kept adding 500 sentences until we reached 3000. To evaluate the domain adaptation, we performed a 10-fold cross-validation at the document level by training on the union of the source data and the sentences selected from the target domain (DA). For each fold, 10% of target data was used for testing, and additional sentences for training were randomly selected from the sentences not used for testing. We also investigated what could be achieved if the system was trained only on the added target sentences without using the source domain in the training process (ID). This model was also evaluated in a 10-fold cross-validation setting.

As a baseline, we applied simple dictionary-based labeling (DL). Texts were lemmatized and if an item from the lists used by the LVC list feature occurred in the text, it was marked as an LVC. We also compared our results with those of a rule-based LVC recognition method (RB) [Vincze et al. 2011a], which basically depends on POS-rules. It means each n-gram that matched the predefined patterns was accepted as an LVC, just like our POS-pattern feature. As this method provides a big pool of potential LVCs, they are filtered by some further criteria: the same Suffix, Stem, and Syntax features were applied as we presented. The results of our experiments can be seen in Tables XI, XII, XIII, and XIV.

4.4. Results

Tables XI and XII give the results for the English corpora, while Tables XIII and XIV show those for the Hungarian corpora. The domain adaptation results were obtained by extending the source domain with 3000 sentences from the target domain.

Table XI shows the results on the English corpora. Based on the 10-fold cross validation results, our system was the most effective in the case of the legal domain

Table XI. Experimental Results on Different Target and Source English Domain Pairs for F-Score. (TARGET: in-domain setting. CROSS: cross-domain setting. RB: rule-based methods. DL: dictionary labeling. Diff_{CROSS}: differences between the TARGET and CROSS results. Diff_{RB}: differences between the TARGET and RB results. Diff_{DL}: differences between the TARGET and DL results)

Corpus Source	Corpus Target	TARGET	CROSS	RB	DL	Diff_{CROSS}	Diff_{RB}	Diff_{DL}
SzegedParalellFX	JRC-Acquis	64.09	59.05	39.93	25.78	−5.04	−24.16	−38.31
SzegedParalellFX	CoNLL-2003	59.41	47.35	47.63	13.79	−12.06	−11.78	−45.62
JRC-Acquis	SzegedParalellFX	62.50	50.83	44.06	20.50	−11.67	−18.44	−42.00
JRC-Acquis	CoNLL-2003	59.41	44.38	47.63	13.79	−15.03	−11.78	−45.62
CoNLL-2003	JRC-Acquis	64.09	57.59	39.93	25.78	−6.50	−24.16	−38.31
CoNLL-2003	SzegedParalellFX	62.50	51.84	44.06	20.50	−10.66	−18.44	−42.00
Avg.	-	62.00	51.84	43.87	20.02	−10.16	−18.13	−41.98

Table XII. Domain Adaptation Results on English Corpora for F-score. (DA: domain adaptation setting. ID: training on a limited set of target data. Diff_{DA}: differences between the CROSS and DA results. $\text{Diff}_{DA/ID}$: differences between the DA and ID results)

Corpus Source	Corpus Target	CROSS	DA	ID	Diff_{DA}	$\text{Diff}_{DA/ID}$
SzegedParalellFX	JRC-Acquis	59.05	67.04	59.88	7.99	7.16
SzegedParalellFX	CoNLL-2003	47.35	51.61	43.37	4.26	8.24
JRC-Acquis	SzegedParalellFX	50.83	61.92	60.99	11.09	0.93
JRC-Acquis	CoNLL-2003	44.38	52.05	43.37	7.67	8.68
CoNLL-2003	JRC-Acquis	57.59	68.04	59.88	10.45	8.16
CoNLL-2003	SzegedParalellFX	51.84	62.14	60.99	10.30	1.15
Avg.	-	51.84	60.47	54.74	8.63	5.73

Table XIII. Experimental Results on Different Target and Source Hungarian Domain Pairs for F-Score. (TARGET: in-domain setting. CROSS: cross-domain setting. RB: rule-based methods. DL: dictionary labeling. Diff_{CROSS}: differences between the TARGET and CROSS results. Diff_{RB}: differences between the TARGET and RB results. Diff_{DL}: differences between the TARGET and DL results)

Corpus Source	Corpus Target	TARGET	CROSS	RB	DL	Diff_{CROSS}	Diff_{RB}	Diff_{DL}
SzT news	SzT paper	53.51	52.07	39.80	32.72	−1.44	−13.71	−20.79
SzT news	SzT law	78.97	67.85	58.56	33.50	−11.12	−20.41	−45.47
SzT paper	SzT news	68.94	51.93	36.70	28.57	−17.01	−32.24	−40.37
SzT paper	SzT law	78.97	68.74	58.56	33.50	−10.23	−20.41	−45.47
SzT law	SzT news	68.94	43.61	36.70	28.57	−25.33	−32.24	−40.37
SzT law	SzT paper	53.51	37.85	39.80	32.72	−15.66	−13.71	−20.79
Avg.	-	67.14	53.67	45.02	31.60	−13.46	−22.12	−35.54

Table XIV. Domain Adaptation Results on Hungarian Corpora for F-Score. (DA: domain adaptation setting. ID: training on a limited set of target data. Diff_{DA}: differences between the CROSS and DA results. $\text{Diff}_{DA/ID}$: differences between the DA and ID results)

Corpus Source	Corpus Target	CROSS	DA	ID	Diff_{AD}	$\text{Diff}_{DA/ID}$
SzT news	SzT paper	52.07	55.08	46.89	3.01	8.19
SzT news	SzT law	67.85	74.00	74.18	6.15	−0.18
SzT paper	SzT news	51.93	62.21	52.57	10.28	9.64
SzT paper	SzT law	68.74	71.96	74.18	3.22	−2.22
SzT law	SzT news	43.61	59.58	52.57	15.97	7.01
SzT law	SzT paper	37.85	51.76	46.89	13.91	4.87
Avg.	-	53.67	59.97	54.20	8.76	4.55

JRC-Acquis (F-score = 64.09%). At the same time the CoNLL-2003 short news domain proved to be the most difficult corpus, where the F-score was only 59.41%. In the case of cross-experiments, the best results were obtained for the JRC-Acquis legal domain. The average results of the CROSS experiments of the three different corpora were 10.16% less than the corresponding TARGET results. The rule-based (RB) approach proved to be the best on CoNLL-2003 with an F-score of 47.63%. The difference between the average results of the TARGET and the RB experiments was 18.13%. The results of the baseline Dictionary labeling method were considerably exceeded by the TARGET results.

The DA column of Table XII lists the results obtained for the English domain adaptation task. Domain adaptation was the most effective when SzegedParalellFX was the target corpus. The domain adaptation results exceeded the cross-experiments by 7.44%. The average difference between in-domain and domain adaptation experiments was 5.73%.

Table XIII shows the baseline and 10-fold cross-validation target results for the Hungarian corpora. Our system proved to be the most effective for the legal domain (an F-score of 78.97%). The average CROSS F-score was 13.46% less than the TARGET scores. The rule-based approach proved to be the best on the SzT law corpus with 58.56%. Dictionary labeling achieved 31.6% on the three corpora, which was exceeded by the TARGET results by 35.54%.

Table XIV lists the results for Hungarian domain adaptation. Based on these values, domain adaptation proved to be the best (better by 15.97%) when SzT law was the source and SzT news was the target. The average domain adaptation results were 8.76% higher than the CROSS results. The average difference between in-domain and domain adaptation results was 4.55%.

The size of the target data added to the source datasets greately influenced the results, as shown by two typical settings in Figure 3. The first part of the diagram shows the JRC-Acquis target results, cross experiments, baselines, and domain adaptation results obtained when the source was CoNLL-2003. This model can already outperform the JRC-Acquis TARGET result when we add only 1500 target sentences to the training data, and the F-score was 3.95% better when 3000 target sentences were added. The gap between the in-domain and domain adaptation results progressively decreases with the size of the dataset. But the gap between the CROSS and domain adaptation progressively increases with the amount of the data added. The second diagram shows the results obtained when the Szeged Treebank newspaper domain was the target and news was the source. The results got from this model also exceeded the TARGET results when we added over 2500 target sentences to the training dataset.

5. DISCUSSION

Machine learning methods extensively outperformed our baseline models, that is, the rule-based model and dictionary labeling, which demonstrates that our CRF-based approach can be suitably applied to LVC detection. This is also supported by the fact that our model outperformed that of Tu and Roth [2011], using the same test set. As illustrated by ablation, the most useful features of the model were morphological, but the effect of syntactic information was more noticeable in English than in Hungarian. Since Hungarian morphology encodes a lot of (morpho)syntactic information, it is not surprising that syntax contributes to LVC detection to a lesser extent in a morphologically rich language, although the quality of tagging may also influence the results. Furthermore, the Suffix feature proved more useful for Hungarian than for English. This may be due to the fact that, in English, conversion is also a possible linguistic means to derive a verb from a noun (such as *change*), while nominal derivation is usually executed by adding derivational suffixes to the verb (such as *ajánl*

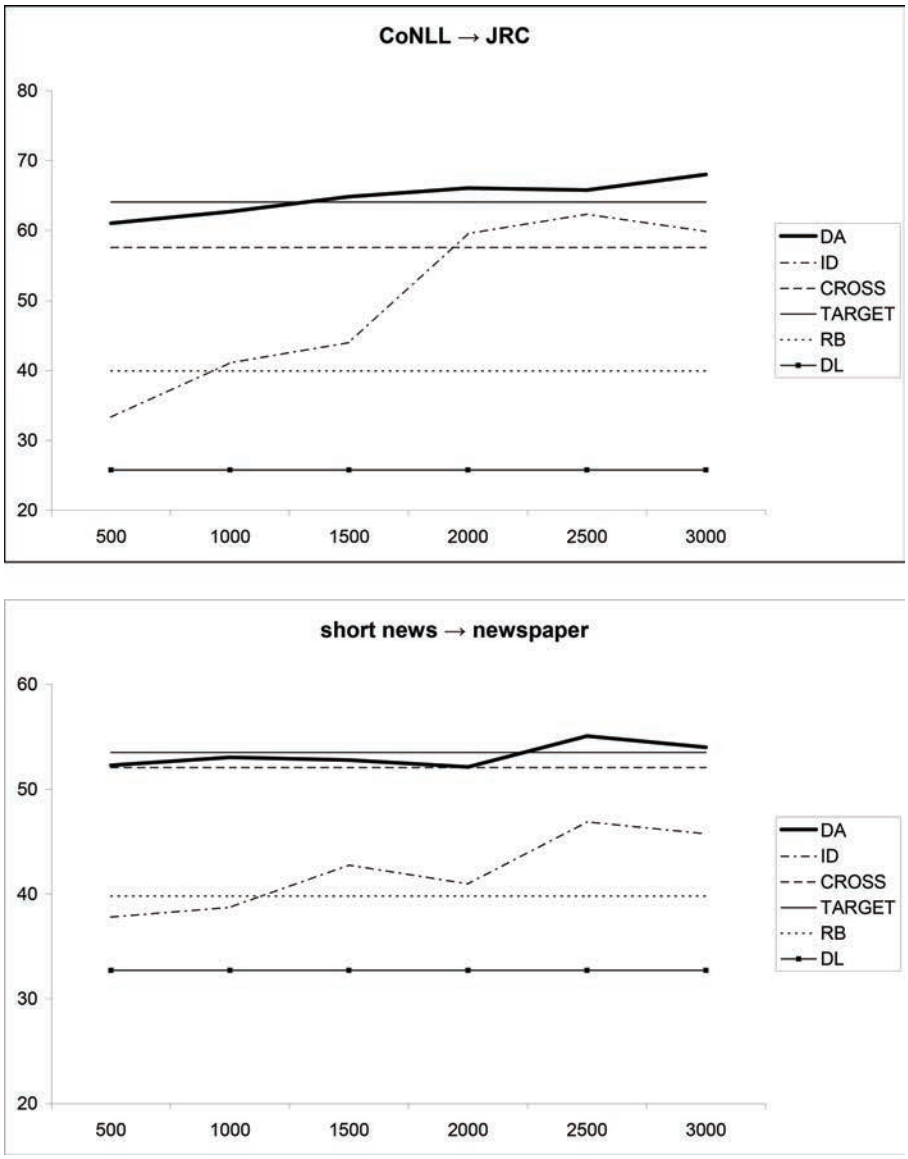

Fig. 3. The effect of the size of the target data on detecting LVCs. DA: domain adaptation setting. ID: training on a limited set of target data. CROSS: cross-domain setting. TARGET: in-domain setting. RB: rule-based methods. DL: dictionary labeling.

"to offer"—*ajánlat* "offer") in Hungarian, and conversion is almost never applied. Hence, many Hungarian nouns in LVCs end in a derivational suffix, while in English this is only true for vague action verbs, which means that this feature may play a significant role in distinguishing between vague action verbs and true light verb constructions. We would like to explore this issue later.

Our cross-domain experiments highlighted the domain dependency of detecting LVCs, since the cross-domain results were always worse than the corresponding in-domain (TARGET) results. However, when there is only a limited amount of target

data available, domain adaptation is more effective because the outdomain dataset also contributes to the training process, and training only on the amount of annotated target data (500, 1000, etc., sentences) cannot achieve such outstanding results. There is only one notable exception: the law domain in Hungarian does not seem to profit from outdomain data: it just confuses learning, and even with a small amount of annotated target data (around 1500 sentences), it is possible to beat the results of cross-training and domain adaptation. This may be explained by the fact that the legal domain apparently has a specific language different from the other domains. The distance between the domains also justifies this fact: the newspaper and short news domains are more similar to each other than any of them and the legal domain (see Table VIII). The special nature of the legal domain is also evident from the baseline results: compared to the other domains, here the rule-based system is able to achieve a fairly good result (58.56%). This suggests that the morphological and syntactic patterns of LVCs in the Hungarian law corpus typically follow the canonical form of Hungarian LVCs, and thus can be identified by rules.

In English, the effect of using outdomain data is especially fruitful in the case of the short news domain, which may be attributed to the fact that, in this domain, the frequency of LVCs is lower than those in the other domains: 4.5% of the sentences contain an LVC, in contrast with the newspaper and law domains (8.67% and 9.08%, respectively). Thus, the same number of target sentences contain fewer LVCs on average and outdomain data can add some more training examples. Nevertheless, cross-domain results can substantially be improved by adding target data in the newspaper domain, which suggests that this domain has some special characteristics which can only be learned from the target data. In the case of the legal domain, domain adaptation even outperformed results achieved by training exclusively on the target dataset in a 10-fold cross-validation setting, which is due to the fact that the legal domain contains the fewest LVCs, and also that there is not such a big difference among the domains in English as in Hungarian, where adding outdomain data to the legal domain just confused learning.

Cross-training by itself did not prove sufficient in many cases, so to reduce the gap between domains, the inclusion of annotated target data into the training dataset was necessary. The domain adaptation settings showed that by adding some outdomain data to the training dataset, it was possible to achieve results similar to—or in some cases, even better than—the target results. It was also found (see Figure 3) that similar results could be achieved on, for example, the JRC-Acquis corpus if we have (1) 2500 annotated target sentences and a substantial amount of annotated outdomain data or (2) at least 5000 annotated target sentences. These values are comparable to those reported in Szarvas et al. [2012], where the domain specificity of uncertainty cue detection is analyzed in detail.

As regards the different domains, the legal domain apparently differs from the other two in both languages. The best TARGET results could be achieved on this domain, which may be because this is the most homogeneous domain: the law corpora contain the fewest LVC lemmas, but the average frequency of LVC lemmas was the highest here. Furthermore, the number of hapax legomena (i.e., LVCs occurring only once in the corpus) is low compared with the other corpora. This also explains why it is easy to adapt a model to the law domain, whereas it is difficult to adapt a model from it to other domains: the limited legal LVC vocabulary can be effectively learned from a small amount of target data, whereas the more extensive vocabulary of the newspaper and short news domains cannot be easily acquired if the training dataset contains a lot of texts from the source domain (i.e., law) and only a few sentences from the target domain.

Table XV. Results for LVCs with Different Lengths on English Corpora

Token length	SzegedParalellFX	JRC-Acquis	CoNLL-2003
2	73.74/86.90/79.78	68.29/84.85/75.68	60.29/77.36/67.77
3	54.67/70.69/61.65	56.19/77.63/65.19	51.55/71.43/59.88
4≤	33.98/58.33/42.94	41.18/63.64/50.00	40.28/64.44/49.57
All	54.29/73.64/62.5	55.38/76.06/64.09	50.85/71.43/59.41

Table XVI. Results for LVCs with Different Lengths on Hungarian Corpora

Token length	SzT newspaper	SzT law	SzT short news
2	48.42/68.62/56.78	76.92/86.71/81.52	66.00/80.39/72.49
3	11.11/33.33/16.67	29.17/53.85/37.84	20.41/62.50/30.77
4≤	0.00/0.00/0.00	18.18/100.00/30.77	16.67/100.00/28.57
All	44.69/66.67/53.51	73.02/85.98/78.97	60.82/79.58/68.94

The Hungarian newspaper domain turned out to be the hardest for LVC detection among all corpora, with a TARGET F-score of only 53.51%. This corpus seemed to contain the most heterogeneous LVCs and their distribution is rather balanced, in other words, there are no very frequent LVCs, which may be responsible for a big percentage of LVC occurrences. What is more, LVCs with nontypical verbal components are also frequent in this corpus, which makes their identification harder (see Section 6). Lastly, certain errors in LVC detection were due to erroneous annotation.

Comparing the results obtained for the two languages, it is striking that the Hungarian results are generally better than the English results. This might be due to several factors. First, in Hungarian, datasets were much bigger than those in English, hence the training datasets contained more examples, which probably had a positive effect on the results. However, the general proportion of LVCs is not significantly different in the two languages as far as the LVC/verb ratio or LVC/token ratio is concerned. Hence we think that if we could have access to more domain-specific data in English, we could achieve better results on the English corpora as well. Second, shorter LVCs were easier to identify (see Table XV and XVI) and about 90% of the Hungarian LVCs are bigrams, which is true only for LVC lemmas in English (see Table IV). This is primarily due to language specific rules. On the one hand, in Hungarian, most LVCs do not have an article within the construction, whereas this is often the case with their English equivalents (cf., *döntést hoz* (decision-ACC brings) vs. *make a decision*). On the other hand, the canonical order of the Hungarian construction is noun + verb, hence modifiers of the noun do not go in between the noun and the verb, whereas in English, if the noun has premodifiers, they go in between the verb and the noun. Compare:

Example 5.1. make a very good decision
nagyon jó döntést hoz (very good decision-ACC brings)

In the Hungarian construction, the noun and the verb are adjacent, while in English they are not, which—given that CRF-based approaches are optimized for sequence labeling—results in an easier detectability of Hungarian LVCs. Third, our feature set included a lot of morphological features, which are especially effective for a morphologically rich language.

6. ERROR ANALYSIS

In order to gain a deeper understanding of the system's performance, we carried out an error analysis of the data. Besides annotation errors, in many cases, erroneous predictions were related to incorrect POS-tags. In Hungarian, a common error of the

POS-tagger was that past tense verbs were often tagged as adjectives (past participles—the word form of which coincides with past tense verbs—do not have a distinct code but are tagged as adjectives), and an adjective + noun sequence was not marked as an LVC. In English, participial occurrences of LVCs were also marked by the system, for example, *taking a decision* can be a participle form and a verbal form as well, depending on the context. However, we focused only on verbal occurrences and removed participial LVCs from the gold standard data before evaluation, thus if a participial occurrence of an LVC was marked, it was treated as a false positive.

An interesting source of error in Hungarian was related to lemmatization. Some word forms can be ambiguous between the derived forms of two verbal stems: for instance, *vetet* can be a causative form of *vesz* "buy" and *vet* "sow" as well. While *vesz* is a typical light verb in Hungarian, this is not true for *vet*, which rarely occurs in LVCs, hence a false lemma can easily lead to errors in LVC detection.

The length of LVCs can also have an impact on their detection: the longer the LVC, the worse the results are likely to be (see above). Constructions with nontypical nominal components (i.e., those not derived from a verb) are also harder to detect; furthermore, constructions with rare verbal components are difficult to recognize, which is especially true for Hungarian newspaper texts. There, we can find many verbal components which do not occur among the most frequent ones or they form a light verb construction with only one or two nouns (e.g., *tüzet nyit* (fire-ACC opens) "to open fire" or *búcsút int* (farewell-ACC waves) "to bid farewell"). To sum up, constructions with nontypical nominal or verbal components and infrequent LVCs are the most difficult to recognize.

7. CONCLUSIONS

In this article we presented our CRF-based system, which is able to identify LVCs in various domains and can be applied to both Hungarian and English texts. Our results highlight the domain-specificity of LVCs, but it was also shown that the gap between domains can be reduced by simple domain adaptation techniques. Based on our experiments, the legal domain seems to be the easiest for LVC detection in both languages. It would also be interesting to extend the scope of LVC detection to other domains like literature or science, which we hope to do in the future.

We presented a linguistics-oriented classification of LVC phenomena and showed that our system is able to effectively identify several types of LVCs; that is, a broader range of LVCs can be detected with our method than with those described in earlier studies. As previous studies focused only on certain subtypes of LVCs, it should be emphasized that, with slight modifications of the features, our system can be further refined to distinguish those subtypes as well—as justified by our results on detecting true light verb constructions in the Tu and Roth database. Later on, we would like to attempt to classify LVCs into the groups defined in this article.

Our performance scores were higher for Hungarian texts than for the English ones. This may be due to the fact that the corpora—hence the training databases—were bigger in the case of Hungarian. In the future, we would like to examine the effect of the size of the datasets by expanding the English corpora, and we would like to experiment with other languages as well.

REFERENCES

ALONSO, M. R. 2004. *Las construcciones con verbo de apoyo.* Visor Libros, Madrid.

APRESJAN, J. D. 2004. O semantičeskoj nepustote i motivirovannosti glagol'nyx leksičeskix funkcij. *Voprosy jazykoznanija 4,* 3–18.

APRESJAN, J. D. AND TSINMAN, L. L. 2002. Formal'naja model' perifrazirovanija predloženij dlja sistem pererabotki tekstkov na estestvennyx jazykax. *Russkij jazyk v naučnom osveščenii 2,* 4, 102–146.

BANNARD, C. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions (MWE'07),* Association for Computational Linguistics, 1–8.

BEJCEK, E. AND STRANÁK, P. 2010. Annotation of multiword expressions in the Prague Dependency Treebank. *Lang. Resources Eval. 44*, 1-2, 7–21.

BOUMA, G. 2010. Collocation extraction beyond the independence assumption. In *Proceedings of the ACL Conference* (Short Papers). Association for Computational Linguistics, 109–114.

CALZOLARI, N., FILLMORE, C., GRISHMAN, R., IDE, N. LENCI, A., MACLEOD, C., AND ZAMPOLLI, A. 2002. Towards best practice for multiword expressions in computational lexicons. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC'02)*. 1934–1940.

CINKOVÁ S. AND KOLÁŘOVÁ, V. 2005. Nouns as components of support verb constructions in the Prague Dependency Treebank. In *Insight into Slovak and Czech Corpus Linguistics*, M. Šimková, Ed., Veda Bratislava, Slovakia, 113–139.

COOK, P., FAZLY, A. AND STEVENSON, S. 2007. Pulling their weight: Exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions (MWE'07).* Association for Computational Linguistics, 41–48.

COOK, P., FAZLY, A., AND STEVENSON, S. 2008. The VNC-tokens dataset. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE'08).* 19–22.

DAUMÉ III, H. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.* Association for Computational Linguistics, 256–263.

DIAB, M. AND BHUTADA, P. 2009. Verb noun construction MWE token classification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications.* Association for Computational Linguistics, 17–22.

DIAS, G. 2003. Multiword unit hybrid extraction. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment.* Vol. 18, Association for Computational Linguistics, 41–48.

É. KISS, K. 2002. *The Syntax of Hungarian*. Cambridge University Press, Cambridge, UK.

FAZLY, A. AND STEVENSON, S. 2007. Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions.* Association for Computational Linguistics, 9–16.

FINKEL, J. R., GRENAGER, T., AND MANNING, C. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL'05).* Association for Computational Linguistics, 363–370.

GURRUTXAGA, A. AND ALEGRIA, I. N. 2011. Automatic extraction of NV Expressions in Basque: Basic issues on co-occurrence techniques. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World.* Association for Computational Linguistics, 2–7.

HENDRICKX, L., MENDES, A., PEREIRA, S., GONÇALVES, A., AND DUARTE, I. 2010. Complex predicates annotation in a corpus of Portuguese. In *Proceedings of the 4th Linguistic Annotation Workshop.* Association for Computational Linguistics, 100–108.

KAALEP, H.-J. AND MUISCHNEK, K. 2006. Multi-word verbs in a flective language: The case of Estonian. In *Proceedings of the EACL Workshop on Multi-Word Expressions in a Multilingual Contexts.* Association for Computational Linguistics, 57–64.

KAALEP, H.-J. AND MUISCHNEK, K. 2008. Multi-word verbs of Estonian: A database and a corpus. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE'08)*. 23–26.

KEARNS, K. 2002. Light verbs in English. Manuscript.

KIM, S. N. 2008. Statistical modeling of multiword expressions. Ph.D. dissertation, University of Melbourne.

KLEIN D. AND MANNING, C. D. 2003. Accurate unlexicalized parsing. In *Proceedings of the Annual Meeting of the ACL.* Vol. 41, 423–430.

KRENN, B. 2008. Description of evaluation resource—German PP-verb data. In *Proceedings of the LREC Workshop Towards a Shared Task for Multiword Expressions (MWE'08)*. 7–10.

LAFFERTY, J. D., MCCALLUM, A. K., AND PEREIRA, F. C. N. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the 18th International Conference on Machine Learning (ICML'01).* Morgan Kaufmann, San Francisco, CA, 282–289.

MCCALLUM, A. K. 2002. MALLET: A machine learning for language toolkit. http://mallet.cs.umass.edu.

MEYERS, A., REEVES, R., MACLEOD, C., SZEKELY, R., ZIELINSKA, V., YOUNG, B., AND GRISHMAN, R. 2004. The NomBank project: An interim report. In *Proceedings of the HLT-NAACL Workshop: Frontiers in Corpus Annotation.* A. Meyers, Ed., Association for Computational Linguistics, 24–31.

Muischnek, K. and Kaalep, H. J. 2010. The variability of multi-word verbal expressions in Estonian. *Lang. Resources Eval. 44*, 1–2, 115–135.

Nagy T., I., Vincze, V., and Berend, G. 2011. Domain-dependent identification of multiword expressions. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'11)*. 622–627.

Pecina, P. 2010. Lexical association measures and collocation extraction. *Lang. Resources Eval. 44*, 1-2, 137–158.

Piao, S. S. L., Rayson, P., Archer, D., Wilson, A., and McEnery, T. 2003. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*. Vol. 18, Association for Computational Linguistics, 49–56.

Ramisch, C., Villavicencio, A., and Boitet, C. 2010a. Multiword expressions in the wild? The `MWEToolkit` comes in handy. In *Proceedings of COLING'10* (Demonstrations). 57–60.

Ramisch, C., Villavicencio, A., and Boitet, C. 2010b. `MWEToolkit`: A framework for multiword expression identification. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*. N. Calzolari et al., Eds., European Language Resources Association, 19–21.

Rayson, P., Piao, S. S., Sharoff, S., Evert, S. and Moirón, B. V. 2010. Multiword expressions: Hard going or plain sailing? *Lang. Resources Eval. 44*, 1-2, 1–5.

Sag, I. A., Baldwin, T., Bond, F., Copestake, A., and Flickinger, D. 2002. Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing'02)*. 1–15.

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 Workshop on NLP and Linguistics: Finding the Common Ground*. Association for Computational Linguistics, 52–60.

Sanches, M. D., Ramisch, C., Aluísio, S. M., and Villavicencio, A. 2011. Identifying and analyzing Brazilian Portuguese complex predicates. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World.* Association for Computational Linguistics, 74–82.

Sanromán Vilas, B. N. 2009. Towards a semantically oriented selection of the values of Oper_1: The case of *golpe* 'blow' in Spanish. In *Proceedings of the 4th International Conference on Meaning-Text Theory (MTT'09).* D. Beck et al., Eds., 327–337.

Sass, B. 2010. Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból [Extracting parallel multiword verbs from parallel corpora]. In *VII. Magyar Számítógépes Nyelvészeti Konferencia,* A. Tanács and V. Vincze, Eds., Szegedi Tudományegyetem, Szeged, 102–110.

Sinha, R. M. 2011. Stepwise mining of multi-word expressions in Hindi. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World.* Association for Computational Linguistics, 110–115.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., and Varga, D. 2006. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*. 2142–2147.

Stevenson, S., Fazly, A., and North, R. 2004. Statistical measures of the semi-productivity of light verb constructions. In *Proceedings of the 2nd ACL Workshop on Multiword Expressions: Integrating Processing*. T. Tanaka et al., Eds., Association for Computational Linguistics, 1–8.

Szarvas, Gy., Farkas, R., and Kocsor, A. 2006. A multilingual named entity recognition system using boosting and C4.5 decision tree learning algorithms. In *Discovery Science*, 267–278.

Szarvas, Gy., Vincze, V., Farkas, R., Móra, Gy., and Gurevych, I. 2012. Cross-genre and cross-domain detection of semantic uncertainty. *Computat. Ling.* (Special Issue on Modality and Negation) *38*, 2, 335–367.

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*. Association for Computational Linguistics, 49–56.

Tjong Kim Sang, E. F., and De Meulder, F. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of CoNLL-03*. W. Daelemans and M. Osborne, Eds., 142–147.

Toutanova, K. and Manning, C. D. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In *Proceedings of EMNLP'00*. Association for Computational Linguistics, 63–70.

Tu, Y. and Roth, D. 2011. Learning English light verb constructions: Contextual or statistical. In *Proceedings of the Workshop on Multiword Expressions: From Parsing and Generation to the Real World*. Association for Computational Linguistics, 31–39.

Van De Cruys, T. and Moirón, B. V. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions (MWE'07)*, Association for Computational Linguistics, 25–32.

Villavicencio, A., Kordoni, V., Zhang, Y., Idiart, M., and Ramisch, C. 2007. Validation and evaluation of automatically acquired multiword expressions for grammar engineering. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*. Association for Computational Linguistics, 1034–1043.

Vincze, V. 2011. Semi-compositional noun + verb constructions: Theoretical questions and computational linguistic analyses. Ph.D. dissertation, University of Szeged, Szeged, Hungary.

Vincze, V. 2012. Light verb constructions in the SzegedParalellFX English–Hungarian parallel corpus. In *Proceedings of LREC'12*.

Vincze, V. and Csirik, J. 2010. Hungarian corpus of light verb constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling'10)*. Coling 2010 Organizing Committee, 1110–1118.

Vincze, V., Nagy T., I., and Berend, G. 2011a. 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*. ACL, 116–121.

Vincze, V., Nagy T., I., and Berend, G. 2011b. Multiword expressions and named entities in theWiki50 corpus. In *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'11)*. 289–295.

Vincze, V., Szauter, D., Almási, A., Móra, Gy., Alexin, Z., and Csirik, J. 2010. Hungarian dependency treebank. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC'10)*.

Zsibrita, J., Vincze, V., and Farkas, R. 2010. Ismeretlen kifejezések és a szófaji egyértelműsítés [Unknown expressions and POS-tagging]. In MSzNy 2010 – VII. Magyar Számítógépes Nyelvészeti Konferencia, A. Tanács and V. Vincze, Eds., University of Szeged, Szeged, Hungary, 275–283.

Received June 2012; revised October 2012; accepted February 2013