Natural Language Engineering (2020), **26**, pp. 319–348
doi:10.1017/S1351324919000330

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Detecting light verb constructions across languages

István Nagy T.[1], Anita Rácz[2] and Veronika Vincze[3,*]

[1]Black Swan Hungary, Budapest, Hungary, [2]Institute of Informatics, University of Szeged, Szeged, Hungary and
[3]MTA-SZTE Research Group on Artificial Intelligence, Szeged, Hungary
*Corresponding author. Email: vinczev@inf.u-szeged.hu

(Received 19 June 2018; revised 15 April 2019; accepted 15 April 2019; first published online 15 July 2019)

Abstract

Light verb constructions (LVCs) are verb and noun combinations in which the verb has lost its meaning to some degree and the noun is used in one of its original senses, typically denoting an event or an action. They exhibit special linguistic features, especially when regarded in a multilingual context. In this paper, we focus on the automatic detection of LVCs in raw text in four different languages, namely, English, German, Spanish, and Hungarian. First, we analyze the characteristics of LVCs from a linguistic point of view based on parallel corpus data. Then, we provide a standardized (i.e., language-independent) representation of LVCs that can be used in machine learning experiments. After, we experiment on identifying LVCs in different languages: we exploit language adaptation techniques which demonstrate that data from an additional language can be successfully employed in improving the performance of supervised LVC detection for a given language. As there are several annotated corpora from several domains in the case of English and Hungarian, we also investigate the effect of simple domain adaptation techniques to reduce the gap between domains. Furthermore, we combine domain adaptation techniques with language adaptation techniques for these two languages. Our results show that both out-domain and additional language data can improve performance. We believe that our language adaptation method may have practical implications in several fields of natural language processing, especially in machine translation.

Keywords: Semantics; Machine learning; Lexicography; Multilinguality

1. Introduction

In natural languages, there are many ways to express complex human thoughts and ideas. This can be achieved by exploiting compositionality, that is, concatenating simplex elements of language, and thus, yielding a more complex meaning that can be computed from the meaning of the original parts and the way they are combined. However, non-compositional phrases can also be found in languages, which are complex phrases that can be decomposed into single meaningful units, but the meaning of the whole phrase cannot (or can only partially) be computed from the meaning of its parts. Such phrases are often called multiword expressions (MWEs) and they display lexical, syntactic, semantic, pragmatic, and/or statistical idiosyncrasies (Sag *et al.* 2002; Calzolari *et al.* 2002; Kim 2008), which might pose problems for linguistic processing, especially in language learning and natural language processing (NLP).

From a multilingual perspective, the handling of MWEs is even more complicated. The very same meaning can be expressed by completely different words in different languages: for example, take the phrase *to make a decision*, which is *eine Entscheidung treffen* (a decision to meet) in German and *döntést hoz* (decision-ACC bring) in Hungarian, that is, three different verbs are used here whose original meanings are quite distinct. Hence, MWEs need special attention across languages as well.

In this paper, we focus on a specific type of MWEs, namely, light verb constructions (LVCs). They are verb and noun combinations in which the verb has lost its meaning to some degree[a] and the noun is used in one of its original senses, most typically denoting an event or state (e.g., *have lunch* or *have a walk*). They usually share their syntactic structures with those of compositional syntactic phrases (*have a walk* vs. *have a dog*), which may impede their automatic identification. However, in several NLP applications like information retrieval or machine translation it is important to identify LVCs in context since they require special care, especially because of their semantic features; however, the detection of LVCs is often hindered by the lack of annotated resources of sufficient size. Here, we will overcome these difficulties by showing that annotated data from other languages[b] can be fruitfully exploited in training supervised machine learning methods to detect LVCs in texts.

We will first analyze the characteristics of LVCs with special regard to four different languages (English, German, Spanish, and Hungarian) from a linguistic point of view based on parallel corpus data. Then, we will provide a standardized (i.e., language-independent) representation of LVCs that can be used in machine learning experiments. After, we will demonstrate that additional language data can be successfully employed in improving the performance of supervised LVC detection for a given language.

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

- We provide an *extended definition of LVCs* and a *linguistic characterization of LVCs in English, German, Spanish, and Hungarian* on the basis of data from the 4FX corpus (Rácz, Nagy T., and Vincze 2014).
- We offer a *standardized linguistic representation of LVCs across languages* with morphological, syntactic, and lexical features, which is deeply rooted in theoretical linguistics but we bear in mind its applicability in machine learning experiments too.
- We introduce our *machine learning-based method* to automatically identify all LVC occurrences in raw text *in four different languages*. Our machine learning method to automatically detect LVCs has already been implemented for English and Hungarian (Vincze, Nagy T., and Farkas 2013a), but here we *adapt this method to German and Spanish* as well.
- As a first step of our machine learning method, we provide a *data-driven candidate extraction method* for all the different languages, which is based on the above-mentioned *standardized syntactic relations*.
- As a second step of our machine learning method, we rely on the *classification approach* reported in Vincze *et al.* (2013a), but we extend the feature set with some *new language-specific features* too.
- We report *results for each language separately* concerning LVC detection with this machine learning method.
- We also examine *how the different languages can affect each other*, therefore, we carry out some *cross-language experiments* for each possible combination of applying a language as target language and a subset or union of the other three languages as source language.
- Here, we apply a domain adaptation-based cross-language adaptation technique and examine *how (domain) adaptation can enhance the results if we only have a limited amount of annotated target language data*.
- We also *examine the effectiveness of feature types* by applying an ablation analysis in several machine learning settings.

[a]Other linguistic studies assume that the verb is semantically underspecified here, and thus, can be used in many different situations (Belvin 1993; Butt and Lahiri 2013).

[b]We will henceforth call data coming from a language other than the target language *additional language data*.

- As in the case of English and Hungarian, we have access to manually annotated data from other domains, we also investigate *the effect of simple domain adaptation techniques to reduce the gap between domains*.
- We *combine domain adaptation techniques with language adaptation techniques* in the case of English and Hungarian and show that *both out-domain and additional language data can improve performance*.

The remainder of the paper is structured as follows: In Section 2, we examine the analyzed languages and their LVCs. In Section 3, we summarize related work, where we present the related corpora and discuss previous methods for identifying LVCs. Then, we discuss the 4FX parallel corpus, the annotation principles and the statistics on corpus data, which is followed by Section 5, where we present our method for the automatic detection of LVCs on different languages. Here, we introduce our approaches to cross-language adaptation and domain adaptation with additional language data too. The results are presented in detail in Section 6. Next, results are elaborated on in Section 7. Lastly, the contributions of the paper are summarized and some possible directions for future study are briefly mentioned.

2. Characteristics of LVCs

It has been earlier pointed out that especially in computational linguistics, different authors apply different names for the same phenomena or rather, they call different phenomena by the same name (see the example of hedges and speculation, discussed in Szarvas *et al.* 2012), making it difficult to compare results obtained by different authors on different datasets. The same is true for LVCs (both as a term and as a linguistic phenomenon). Thus, we think it necessary to give a short summary of how these verb + noun constructions are accounted for in the (computational) linguistic literature.

Traditionally, light verbs have a (direct or indirect) object which is a predicative noun denoting an eventuality. On the other hand, Krenn (2008) deals with just verbs + prepositional phrases. German theoretical linguists (see, e.g., Duden 2006) usually do not consider verb + subject combinations as *Funktionsverbgefüge* (a German term for fixed verb + noun constructions) but Hungarian linguists do (Vincze 2011). However, Mel'čuk (Mel'čuk 1974; Mel'čuk, Clas, and Polguère 1995; Mel'čuk 2005), among others, extended the definition to cover cases where the noun is the subject or an oblique/prepositional argument of the verb. Verbs that bear aspectual information may also occur in this group (Danlos 2010; Vincze 2011; Varga 2014).

As for their semantics, some verbal components are semantically more bleached than the others (Alonso Ramos 2004; Sanromán Vilas 2009). Meyers, Reeves, and Macleod (2004) distinguished support verbs and light verbs: semantically empty support verbs are called light verbs, that is, the term *support verb* is a hypernym of *light verb*. Kearns (2002) also made a distinction between two subtypes of what is traditionally called LVCs. True LVCs 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 properties and can be separated by various tests, for example, passivization, WH-movement, pronominalization, etc.

Thus, a wide range of fixed verb + noun constructions with peculiar semantic behavior have received intensive attention in many languages (for definitions, see, e.g., Langer 2005, Mel'čuk 2005, and Danlos 2010). Here, we summarize their most important general characteristics:

- they are lexically fixed expressions;
- they follow the base form: `verb prep? det? noun` or `prep? det? noun verb` (using the ? operator as in regular expressions);
- their meaning is not totally compositional;

- the noun typically refers to an action or event;
- the noun is a syntactic argument of the verb;
- the verb contributes to the meaning only to a small degree (e.g., aspectual information).

In this paper, we take a comprehensive approach and use the term *light verb construction* as widely as possible. That is, we do not make any morphosyntactic, lexical, semantic, and language-specific restrictions (see also Section 3.3), and we consider each verb + (preposition) + (article) + noun combination[c] as a LVC that fulfills the above criteria. Prepositions and determiners may be also part of the construction (e.g., *take into account*), as we do not restrict ourselves to the investigation of only verb + noun constructions. We believe that using an extended definition of LVCs is beneficial for computational linguistic applications as a wider range of constructions that require special treatment can be identified in this way.

In the following, we will focus on the four languages we are interested in, namely, English, German, Spanish, and Hungarian, and we will also briefly present their grammatical characteristics in general and describe the main characteristics of LVCs as well.

2.1 English

English is known as a prototypical configurational language, that is, it has strict word order and poor morphology. The syntactic roles of arguments are mostly determined by their position within the sentence. As for LVCs, the canonical order (i.e., the form that occurs in dictionaries) is verb + noun, but there can also be an article and/or a preposition in between and due to passivization, the two components may not be adjacent.

Earlier, several studies were carried out to investigate the syntactic features of English LVCs (Hale and Keyser 2002; Meyers *et al.* 2004). They usually share their syntactic structures with productive or literal verb + noun constructions on the one hand and idiomatic verb + noun expressions on the other, see, for example, Fazly and Stevenson (2007), as reflected in the examples *make a cake* (productive), *make a decision* (LVC), and *make a meal* (idiom) (Vincze, Nagy T., and Zsibrita 2013b). Hence, their identification cannot be based on solely syntactic patterns. Kearns (2002) distinguished between two subtypes of what is traditionally called LVCs: true LVCs and vague action verbs differ in some syntactic and semantic properties and can be separated by various tests, like passivization, WH-movement, and pronominalization. Vincze *et al.* (2013b) classify LVCs from morphological and semantic aspects, and they also take into account theoretical and computational linguistic considerations.

2.2 German

German forms part of the Germanic language family and it has a morphologically complex grammar system. Being a flective language, grammatical functions are expressed mostly by the alternation of word forms. As for verbs, they are conjugated for person, number, tense, mood, and voice, and in nominal declination, the number, gender, and case can be specified (Bußmann 2002).

The canonical form of German LVCs is a noun (eventually preceded by a preposition and/or an article) + infinitive verb (Fleischer, Helbig, and Lerchner 2001; Duden 2006). The verb can be conjugated thus can appear in many surface forms, besides, due to word order reasons, it may precede or follow the noun.

LVCs have been studied by German theoretical linguistic research for over 50 years now (Daniels 1963; Häusermann 1977). However, as Heine demonstrates, in descriptive grammars, there is still no consensus about the main features of LVCs (Heine 2006). In Duden, prototypical LVCs are characterized by features like *the determined use of article*, *fixed way of negation* (i.e., it is

[c]The order of these components might vary, depending on the language and/or context.

lexically determined whether the given phrase can be negated with the determiner *kein* or with the adverb *nicht*), or *no free choice of attributes in the case of the nominal component* (Duden 2006). In certain cases, they can be formed with the copula *sein* "to be" and *bleiben* "to stay" as well (Duden 2006) and some verb–noun combinations in genitive case also count as LVCs if they express an opinion (Helbig and Buscha 2001).

2.3 Spanish

Spanish belongs to the Ibero-Romance languages, where verbs can be conjugated for person, number, tense, mood, and aspect and nouns, adjectives, and determiners also have an inflectional paradigm since they carry information about number and gender (RAE 2009).

The canonical form of Spanish LVCs is an `infinitive + preposition? + article? +` noun, hence, their canonical word order is just like in English. In some cases, however, these elements can also be separated, yielding non-canonical, non-adjacent forms.

Until recently, there were only few studies about Spanish LVCs, but nowadays they have come to the fore in the NLP community (Alonso Ramos 2000). The main features of Spanish LVCs are described among others in Ramos (2004), Blanco Escoda (2000), and Bosque (2001). In the field of applied linguistics, we would like to mention Buckingham's work, in which a corpus of written academic language is used in order to identify the most common LVCs (Buckingham 2009). The study of Leoni de León (2014) is also worth noting, as it gives an overview about the lexical status and syntactic features of MWEs and describes a system for their identification.

2.4 Hungarian

Hungarian is an agglutinative language, which means that a word can have hundreds of word forms due to inflectional or derivational morphology (É. Kiss 2002). Hungarian word order is related to information structure, thus, the position relative to the verb has no predictive force as regards the syntactic function of the given argument. In contrast, the grammatical function of words is determined by case suffixes.

The canonical form of a Hungarian LVC is a bare noun + third person singular verb, where the noun can function not only as the object of the verb but also another argument as well. However, the verb may precede the noun, or they may not be adjacent for word order reasons, moreover, the verb may occur in different surface forms inflected for tense, mood, person, and number.

Varga (2014) analyzed Hungarian LVCs of the form verb + object in the lexicon-grammar framework. Vincze *et al.* (2013b) provided a detailed analysis of English and Hungarian LVCs from a morphological viewpoint. Hungarian LVCs are described in detail from a syntactic and semantic viewpoint in Vincze (2011), and they are also contrasted with English as well.

3. Related work

Next, we will present the related corpora and discuss previous methods used to identify LVCs. We mostly focus on databases and methods developed for our languages investigated in this paper, however, we would like to emphasize that there is an extensive ongoing work on MWEs in other languages, for example, within the PARSEME international research network.[d]

3.1 Related corpora

In order to build supervised machine learning-based methods to automatically identify LVCs in texts, well-designed and tagged corpora are invaluable for training and testing statistical models.

[d] `https://typo.uni-konstanz.de/parseme/`.

Here, we present the most commonly used databases and corpora annotated for LVCs in several languages.

Krenn (2008) developed a database of German PP-verb combinations, while Kolesnikova and Gelbukh (2010) collected and classified Spanish verb–noun combinations. Vincze *et al.* (2011b) manually annotated several types of MWEs (including LVCs) in English Wikipedia texts, while Tu and Roth (2011) also published an English corpus, where true LVCs were manually marked. PropBank also contains annotation for LVCs (Hwang *et al.* 2010). Hungarian LVCs were manually marked in the Szeged Treebank (Vincze and Csirik 2010). An English–Hungarian annotated parallel corpus of LVCs was recently published (Vincze 2012). Moreover, English, German, Spanish, and Hungarian legal texts were recently annotated for LVCs in a parallel corpus (Rácz *et al.* 2014). As this corpus will be employed in most of our experiments, it will be described later on in Section 4.

Due to the PARSEME Shared Task that aimed at identifying verbal MWEs in raw texts, a new multilingual corpus has been published (Savary *et al.* 2017; Ramisch *et al.* 2018) in 2017 and 2018. This corpus contains texts for 18 different languages, all annotated for each occurrence of idioms, LVCs, and other types of verbal MWEs. The organizers of the shared task provided a standardized guideline for annotation, which was followed for all languages. Thus, the data in the different languages are easily comparable and provide space for multilingual experiments. However, we did not choose to experiment on this dataset as we first wanted to test our approach on a parallel corpus. In this case, differences in the results obtained cannot be attributed to differences in domain, genre, and the available quantity of annotated texts for different languages. If our method seems to be feasible on the 4FX dataset, it can be extended to other corpora in a future study.

3.2 Methods for identifying LVCs

Several applications have been developed for identifying LVCs in running texts. The methods applied may be divided into three categories, namely, statistical, rule-based, and machine learning-based approaches. It should be noted, however, that most of these studies applied restrictions as regards the set of LVCs to be identified (see Section 3.3 below), hence, their results are not reported here as they are not directly comparable either to each other or to our results.

Stevenson, Fazly, and North (2004) and Fazly and Stevenson (2007) built LVC detection systems with statistical features. Stevenson *et al.* (2004) focused on classifying true LVC candidates containing the verbs *make* and *take*. Fazly and Stevenson (2007) used linguistically motivated statistical measures to distinguish subtypes of verb + noun combinations.

Nagy T., Vincze, and Berend (2011), Vincze, Nagy T., and Berend (2011a), and Nagy T. and Vincze (2011) applied rule-based systems to automatically detect LVCs in running texts. Vincze *et al.* (2011a) exploited shallow morphological features to identify LVCs in English texts, while the domain specificity of the problem was highlighted in Nagy T. *et al.* (2011). Nagy T. and Vincze (2011) identified verb–particle constructions and LVCs in running texts using rule-based methods and examined how adding certain features could affect the overall performance.

Tan, Kan, and Cui (2006) and Tu and Roth (2011) attempted to identify true LVCs by applying machine learning techniques. Tan *et al.* (2006) combined statistical and linguistic features and trained a random forest classifier to separate LVC candidates, while Tu and Roth (2011) trained a Support Vector Machine with contextual and statistical features on their balanced dataset annotated for English LVCs. Bannard (2007) sought to identify verb and noun constructions in English on the basis of syntactic fixedness. Nagy T., Vincze, and Farkas (2013) focused on the full coverage identification of English LVCs in running texts. They classified LVC candidates using a decision tree algorithm, which was based on a rich feature set with new features like semantic and morphological features.

In the PARSEME Shared Tasks, several machine learning-based methods competed on the four languages under investigation, among others. Some of them relied on parsing (Al Saied, Constant,

and Candito 2017; Nerima, Foufi, and Wehrli 2017; Simkó, Kovács, and Vincze 2017; Waszczuk 2018), whereas others exploited sequence labeling using CRFs (Boroş *et al.* 2017; Maldonado *et al.* 2017; Moreau *et al.* 2018) and neural networks (Klyueva, Doucet, and Straka 2017; Berk *et al.* 2018; Boroş and Burtica 2018; Ehren, Lichte, and Samih 2018; Stodden, QasemiZadeh, and Kallmeyer 2018; Zampieri *et al.* 2018).

3.3 Restrictions on LVCs

As earlier studies applied some restrictions on LVCs to be detected, here we will classify them according to the restrictions they utilize.

Some methods previously used (Tu and Roth 2011; Fazly and Stevenson 2007) employed morphosyntactic restrictions and just treated the verb–object pairs as potential LVCs (Stevenson *et al.* 2004; Tan *et al.* 2006). English verb + prepositional constructions were mostly neglected in previous studies, although the annotated corpora mentioned in Section 3.1 contained several examples of such structures, for example, *take into consideration* or *come into contact*. Some researchers applied lexical restrictions and filtered LVC candidates by selecting only certain verbs that may be part of the construction. One such example is Tu and Roth (2011), where the authors chose six light verbs (*make*, *take*, *have*, *give*, *do*, *get*) to focus on. Furthermore, semantic restrictions were applied by some earlier researchers, so they just focused on the identification of true LVCs, neglecting vague action verbs (Stevenson *et al.* 2004; Tan *et al.* 2006; Tu and Roth 2011).

Here,—in accordance with Nagy T. *et al.* (2013)—we seek to identify each type of LVCs and do not restrict ourselves to certain types of LVCs as we think that they all are relevant for NLP (e.g., they must be treated as one complex predicate (Vincze 2012)).

4. The corpus

Now, we will describe the 4FX corpus (Rácz *et al.* 2014) in detail and analyze language-specific features on the characteristics and distribution of LVCs based on empirical data from the corpus.

4.1 Statistical data on the corpus

Texts from our 4FX corpus are based on the JRC-Acquis Multilingual Parallel Corpus, consisting of legal texts for a range of languages used in the European Union (Steinberger *et al.* 2006). For an earlier investigation on LVC detection (Vincze *et al.* 2013b), 60 documents from the English version of the corpus were randomly selected and LVCs in them were annotated. We annotated the Spanish, German, and Hungarian equivalents of those 60 documents, thus, yielding a quadrilingual parallel corpus named 4FX. The annotation was carried out by two native speakers of Hungarian who have an advanced level knowledge of English, German, and Spanish. Data on annotated texts can be seen in Table 1.

Table 1. Statistical data on the 4FX parallel corpus

	English	German	Spanish	Hungarian	All
Sentences	5143	5527	5675	4568	20,913
Tokens	94,747	89,523	107,851	92,707	384,828
Average token/sentence	18.42	16.19	19.01	20.29	18.40

4.2 Annotation principles

In order to make the annotation in the different languages as uniform as possible, we adapted the guidelines applied while constructing the SzegedParalellFX corpus.

During annotation, we decided to mark only the verb that belonged to the LVC (i.e., passive or perfect auxiliaries were not marked as part of an LVC) but verbal prefixes that were separated from their main verb due to word order reasons were marked in German and Hungarian.

We should also mention that we marked not only the prototypical constructions (VERB, e.g., *make contracts*), but also their nominal (marked as NOM, e.g., *making of contracts*) and participial forms (PART, e.g., *contracts made*) were included in the annotation. Constructions marked with PART and NOM were marked regardless of whether they consisted of one or more elements (cf. *szerződéskötés* "making a contract" or *Durchführung einer Untersuchung* "carrying out an investigation"). Also, we took into account the non-adjacent forms of LVCs (SPLIT), like in ***contracts** that have been **made***. The corpus contains annotation for verbal, participial, nominal, and split occurrences of LVCs as well. However, we will neglect participial constructions spelt as one word and also the nominal occurrences of LVCs in our experiments since their identification would require methods radically different from syntax-based methods applicable to verbal and non-compound participial forms.

4.3 Analysis of corpus data

A comparison of the data on the four languages reveals some interesting facts. The statistics on the subtypes of LVCs presented in Table 2 displays intriguing tendencies if we take into account the fact that the parallel versions of the same corpus were the basis of our manual annotation. As can be seen, Spanish constructions marked as VERB occurred more than twice as frequently as those in English. This fact accords with our belief that there is a fundamental difference among LVC types of the languages as well as among the languages regarding their LVCs.

One of the most salient results is that, in German, the number of non-adjacent, SPLIT LVCs exceeds the frequency of all the rest of languages. This may be attributed to the strict rules of German word order. The verb has to occupy the first or second place in a main clause, but its arguments have a more flexible position, for example,

*Diese Verordnung **tritt** am 31. März 2006 **in Kraft**.*
this-FEM directive step-3SG on-MASC-DAT 31 March 2006 in force
"This Directive shall **enter into force** on the 31 March 2006."

Table 2. Types of LVCs in the 4FX corpus

	en	de	es	hu	Total
NOM	37	151	82	199	469
	5.50%	18.73%	8.74%	6.91%	13.49%
VERB	245	265	519	384	1413
	36.40%	32.88%	55.33%	51.51%	40.65%
SPLIT	127	278	132	94	631
	18.87%	34.49%	14.07%	8.88%	18.15%
PART	264	112	205	382	963
	39.23%	13.90%	21.86%	36.07%	27.70%
All	673	806	938	1059	3476
	100.00	100.00	100.00	100.00	100.00

NOM, nominal LVCs; VERB, verbal occurrences; SPLIT, split LVCs; PART, participial LVCs.

Yet another peculiarity of this language is the outstanding frequency of nominal constructions, which is noticeable in Hungarian to some extent as well. This feature is most probably due to the so-called *Nominalstil*, the preferred use of nominal forms in technical jargon. Consequently, our results seem to confirm statistically a fact often referred to in related German literature too, namely, that legal texts abound in nominalized constructions and LVCs in general (Duden 2006). Concerning this feature, German and Hungarian are quite close to each other. In the latter case, however, the wide-ranging rules of word building make it possible to create three different participial forms as well, which are less frequently used in the other languages. This fact might explain the dominance of Hungarian PART constructions in the corpus.

As for a qualitative analysis of the data, Table 3 shows that there are some common LVCs that are frequent in each of the four languages, for instance:

to enter into force—in Kraft treten—entrar en vigor—hatályba lép
to grant aid—Hilfe gewähren—conceder ayuda—támogatást nyújt.

Table 3. LVC occurrences in the 4FX corpus

English		German	
LVC	#	LVC	#
have regard	91	Hilfe gewähren “grant aid”	50
grant aid	38	in Kraft treten “enter into force”	36
take into account	32	Flug durchführen “operate a flight”	33
enter into force	31	Rechnung tragen “take account”	20
receive aid	20	Antrag stellen “hand in an application”	17
Spanish		**Hungarian**	
LVC	#	LVC	#
tener en cuenta “take into account”	79	támogatást nyújt “grant support”	70
conceder ayuda “grant aid”	65	figyelembe vesz “take into account”	49
entrar en vigor “enter into force”	36	részt vesz “take part”	46
adoptar una medida “adopt measures”	24	hatályba lép “enter into force”	34
beneficiarse de una ayuda “receive aid”	20	rendelkezésre áll “be at his disposal”	24

Table 4. Light verb occurrences in the 4FX corpus

English		German		Spanish		Hungarian	
Light verb	#	Light verb	#	Light verb	#	Light verb	#
have	105	gewähren "guarantee"	81	tener "have"	150	vesz "take"	110
take	104	durchführen "execute"	65	conceder "grant"	87	nyújt "offer"	88
make	65	treten "step"	39	efectuar "effect"	50	kerül "get done"	54
grant	42	haben "have"	31	llevar "hold"	45	tesz "make, put"	47
carry out	41	tragen "hold"	30	adoptar "adopt"	43	lép "step"	41

Table 5. Statistics on light verb occurrences in the 4FX corpus

	English	German	Spanish	Hungarian
LVCs	641	655	860	863
LVC lemmas	194	258	336	286
Average occurrence of lemmas	3.30	2.54	2.56	3.02
LVC verbs	42	93	75	80
Average occurrence in lemmas	4.62	2.77	4.48	3.58
Hapax LVCs	101	163	212	175
%	52.06	63.18	63.10	61.19
Hapax LVC verbs	11	35	24	16
%	26.19	37.63	32.00	20.00

Some other constructions are found in the top five LVCs in three of the languages (but not in German), such as

to take into account—tener en cuenta—figyelembe vesz
to receive aid—beneficiarse de una ayuda—támogatásban részesül.

These constructions are the most frequent in the corpus and they are also characteristic of the legal language. On the other hand, we also found language-specific LVCs in the 4FX corpus that do not have an equivalent in all or any of the other languages, just like the English phrase *having regard to* corresponds to the Hungarian phrase *tekintettel* regard-INS "with regard to."

Next, as is seen in Table 5, concerning the number of LVC lemmas, Spanish comes first but concerning the sum of light verbs, German comes top. With regard to these values, English comes quite low, which suggests that LVCs here are less diverse than in the other languages, at least in the legal domain. Analyzing the number of hapax LVCs (i.e., those that occur only once in the dataset), a similar picture is seen, since the frequency of LVCs with only one occurrence in the corpus is the lowest in English.

5. Automatic detection of LVCs

In order to see how the different languages can affect each other, an implementation of the language-independent representation of LVCs is required. Later, we can apply a domain adaptation-based cross-language adaptation technique and discover how (domain) adaptation can enhance the results if we only have a limited amount of annotated target data.

As mentioned in Section 3.1, different types of texts may contain different types of LVCs. Therefore, we also investigate the effect of simple domain adaptation techniques to reduce the gap between the legal text domain and other domains, as there are other manually annotated corpora available for English and Hungarian, where LVCs are manually marked.

5.1 Candidate extraction

As LVCs were manually annotated in the 4FX parallel corpus in four different languages, we were able to define a language-independent syntax-based candidate extraction method to automatically extract potential LVCs from raw text, based on Nagy T. *et al.* (2013). Hence, we examined the syntactic relations among the LVCs' verbal and nominal components in the four different languages. To parse the English, German, and Spanish parts of the 4FX corpus, the Bohnet parser (Bohnet 2010) was applied, which was trained on the English part of the CoNLL shared task (Surdeanu *et al.* 2008), on the TIGER treebank (Brants *et al.* 2004) in German and on the IULA treebank (Marimon *et al.* 2012) in Spanish. In the case of Hungarian, the state-of-the art Hungarian dependency parser, `magyarlanc` (Zsibrita, Vincze, and Farkas 2013), was used, which was trained on the Szeged Dependency Treebank (Vincze *et al.* 2010).

As the different models on different languages were trained on different treebanks, the parsed texts have different dependency representations. For instance, the *verb–direct object* dependency relation is presented by `dobj` in English, `OA` in German, `DO` in Spanish, and `OBJ` in Hungarian. Hence, it was necessary to standardize the different representations of the various syntactic relations and apply the same dependency label for the same grammatical relation across the languages.[e] Table 6 shows the distribution of standardized dependency label types on the four

Table 6. Syntactic relations of LVCs annotated in the 4FX parallel corpus

	English		German		Spanish		Hungarian	
Type	#	%	#	%	#	%	#	%
dobj	280	44.30%	213	32.87%	273	31.89%	272	32.19%
pobj	101	15.98%	35	5.40%	164	19.16%	263	31.12%
nsubjpass	81	12.82%	28	4.32%	45	5.26%	34	4.02%
partmod	64	10.13%	118	18.21%	122	14.25%	167	19.76%
sum	526	83.23%	394	60.80%	604	70.56%	736	**87.10%**
none	68	10.76%	188	29.01%	214	25.00%	107	12.66%
other	38	6.01%	66	10.19%	38	4.44%	2	0.24%
sum	632	100.00	648	100.00%	801	100.00%	638	**100.00%**

dobj, direct object; pobj, adpositional object; nsubjpass, subject of passive verb; partmod, participial modifier; other, a syntactic relation other than the above; none, no syntactic relation between the noun and the verb within the LVC.

[e]Our approach is similar in vein to the universal dependency annotation described in McDonald *et al.* (McDonald *et al.* 2013), but here we only focus on LVC-specific relations. We did not make use of the Universal Dependency treebanks in this phase as some treebanks—including the Hungarian one—explicitly mark LVCs with a specific dependency label, which might affect our results.

languages in the 4FX parallel corpus. In the candidate extraction phase, we treated as a potential LVC each verb and noun combination, which are connected with *verb–direct object* (dobj), the *verb–(passive) subject* (subj/nsubjpass), the *verb–adpositional* (pobj), and *noun–participial* modifier (partmod) syntactic relations.

5.2 Standardized feature representation on different languages

For the automatic classification of the candidate LVCs a machine learning-based approach (Vincze *et al.* 2013a) was utilized, which was successfully applied to English and Hungarian. This method is based on a rich feature set with the following categories: statistical, lexical, morphological, syntactic, and orthographic. In order to detect Spanish and German LVCs in text, we also implemented this feature set in both new languages and we defined some new language-specific features too. A standardized representation of the feature set is also required, as we also would like to investigate how the different languages influence each other. Therefore, the same features in different languages were associated with each other. This feature set includes language-independent and language-specific features as well. Mainly the language-independent features aim to grab general features of LVCs while language-specific features can be applied due to the different grammatical characteristics of the four languages. Table 7 shows which features were applied to which language and a detailed description of the features follows below.

Morphological features: As the nominal component of LVCs is typically derived from a verbal stem (*make a decision*) or coincides with a verb (*have a walk*), the *VerbalStem* binary feature focuses on the stem of the noun; if it had a verbal nature, the candidates were marked as *true*. The *POS-pattern* feature examines the part-of-speech tag sequence of the potential LVC. If it matched one pattern typical of LVCs (e.g., `verb + noun`) the candidate was marked as *true*; otherwise as *false*.

Here, we also defined some language-specific features. In the case of English, the morphological analyzer we applied was able to distinguish between main verbs and *auxiliary verbs*, therefore, we defined a feature for *do* and *have* to denote whether or not they were auxiliary verbs in a

Table 7. The basic feature set and language-specific features

Features	Base	English	German	Spanish	Hungarian
Orthographical	•	•	•	•	•
Syntactic features	•	•	•	•	•
LVC list	•	•	•	•	•
Light verb list	•	•	•	•	•
Noun list	•	•	•	•	•
POS pattern	•	•	•	•	•
VerbalStem	•	•	•	•	•
isSuffix	•	•	•	•	•
LVC freq. stat.	–	•	–	–	–
Auxiliary verb	–	•	–	–	–
Noun compound	–	–	•	–	–
Gender	–	–	•	•	–
Participle	–	–	–	•	–
Agglutinative morph.	–	–	–	–	•
Historical derivation	–	–	–	–	•

given sentence as they often occur as light verbs as well. As Hungarian is a morphologically rich language, we were able to define various morphology-based features. For this, we basically used the morphological analyses of words. In this way, we defined some *agglutinative features* like the mood and the type of the verbs, the case, the number of possessor, the person of the possessor, and the number of the possessed of the noun. Nouns which were historically derived from verbs but were not treated as derivations by the Hungarian morphological parser were also added as a feature. The *gender of nouns* was also utilized as feature in the case of German and Spanish, where the nominal components of LVCs often belong to the feminine gender, due to their derivational suffixes. Moreover, we also defined a German-specific feature which checked if the nominal part of the candidate was a *compound noun* or not. We also defined a special feature for encoding Spanish participles as the morphological analysis did not make any distinction between adjectives and participles, and while participles can be parts of LVCs, adjectives cannot be.

Orthographic features: The *suffix* feature is also based on the fact that many nominal components in LVCs are derived from verbs. This feature checks if the lemma of the noun ends in a given character bi- or trigram. We defined these specific bi- and trigrams for all of the four languages. The *number of words* of the candidate LVC was also noted and applied as a feature.

Statistical features: Potential English LVCs and their *occurrences* were collected from 10,000 English Wikipedia pages by the candidate extraction method. The number of occurrences was used as a feature in case the candidate instantiated one of the four syntactic relations identified in Section 5.1.

Lexical features: Here, we exploit the fact that the *most common verbs* are typically light verbs. Therefore, 15 typical light verbs were selected for all of the four languages (cf. Table 4) and we investigated whether the lemmatized verbal component of the candidate was one of these 15 verbs. Due to the language differences, the typical light verbs in various languages are not the same. But for the standardized feature representation, a uniform treatment of these verbs was required. Therefore, we unified the lists of the most common light verbs in all of the four languages and translated them to all the languages. For instance, if *make* was among the most frequent English light verbs, then its German, Spanish, and Hungarian counterparts were added to the list as well, regardless of whether they were among the most frequent light verbs in the specific language. In this way, quadruples like *make—machen—hacer—tesz* were formed. The list got in this way contains (four times) 29 verbs. The *lemma of the noun* was also applied as a lexical feature. The nouns found in LVCs were collected from the treebanks on which the parsers were trained.

Moreover, *lists of lemmatized LVCs* were also applied as a feature. In the case of English, manually annotated LVCs were collected and lemmatized from the English part of the SzegedParalellFX (Vincze 2012), while the filtered version of the German PP-Verb Collocations list (Krenn 2008) was applied in German. The filtered version of Spanish verb–noun lexical function dictionary (Kolesnikova and Gelbukh 2010) was utilized in Spanish and the manually annotated LVCs from the Hungarian part of the above-mentioned parallel corpus were collected for Hungarian. In German and in Spanish, filtering involved the selection of items that conformed to our definition of LVCs. These lists were also used in our baseline experiments.

Syntactic features: As the candidate extraction methods essentially depended on the *dependency relation* between the nominal and the verbal part of the candidate, they could also be utilized in identifying LVCs. Here, the dependency labels which were used in candidate extraction and introduced in Section 5.1 were defined as features. If the noun had a *determiner* in the candidate LVC, it was also encoded as another syntactic feature.

5.3 Machine learning-based candidate classification

As Nagy T. *et al.* (2013) found that decision trees performed the best in this task, we also applied this machine learning algorithm. Therefore, the J48 classifier of the WEKA package (Hall *et al.* 2009) was trained on the above-mentioned feature set, which implements the C4.5 (Quinlan 1993) decision tree algorithm.

As the different parts of the 4FX corpus were not big enough to split them into training and test sets of appropriate size, we evaluated our models in a 10-fold cross-validation manner at the instance level on each part of the corpus. However, in the 4FX corpus, only LVCs were marked in the text, that is, no negative examples were annotated. Therefore, all potential LVCs were treated as negative which were extracted by the candidate extraction method but were not marked as positive in the gold standard. Thus, in the dataset negative examples were overrepresented.

As Table 6 shows, the syntax-based candidate extraction methods did not cover all manually annotated LVCs in the different parts on the 4FX corpus. Hence, we treated the omitted LVCs as false negatives in our evaluation.

As a baseline, a context-free dictionary lookup method was applied in the four languages. Here, we applied the same LVC lists which were described among the lexical features in Section 5.2. Then, we marked candidates of the syntax-based candidate extraction methods as LVCs if they were found in the list.

In order to examine the effectiveness of each individual feature type, we carried out an ablation analysis in the four different languages. Tables 12–15 tell us how useful the individual features proved to be for the four languages.

5.4 Language adaptation

Here, we introduce our language adaptation technique, which is very similar to domain adaptation. In most cases, domain adaptation can enhance the results if we only have a limited amount of annotated target data. When we only have a limited set of annotated data from one domain, but there are a plenty of data available in another domain, we can apply domain adaptation methods to achieve better results on the target 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.

Here, we treated the different parts of the 4FX corpus with different languages as different domains. Furthermore, the above-mentioned standardized feature representation of LVCs on different languages allowed us to focus on the portability of models trained on different parts of the 4FX corpus. We were also able to investigate the effect of the language adaptation techniques so as to reduce the gap between the language pairs.

Therefore, we were able to apply the adapted machine learning-based method with language-specific features to each language to examine how much the different language models affected each other.

In the pure in-language setting, a 10-fold cross-validation was performed at the candidate level on each subcorpus of the 4FX corpus. To compare the different languages, a pure cross-language setting (CROSS) was utilized, where our model was trained on the source language and evaluated on the target (i.e., no labeled target language datasets were used for training); for example, we trained the model on the English part of corpus and tested it on the Hungarian part.

For the cross-experiments, a unified representation of the candidate LVCs on different languages was required. However, as Table 7 shows, language-specific features were also defined in each language. Therefore, the basic feature set was extended with all of the language-specific features on each language.

A very simple approach was used for language adaptation (ADAPT): we applied 10-fold cross-validation and for each fold, we used 90% of the target language as training data and the other 10% was held out for test. The source language training dataset was extended with instances from the target language training dataset. First, 10% of the target language training dataset was added to the source language training dataset, and then we kept adding 25%, 50%, and finally all of the target training instances. As the source language, we applied all possible combinations of the non-target languages, that is, we added only just one of the languages, two languages, or all the three languages.

Table 8. Experimental results when the English part of the corpus was used as target and source of different language pairs in terms of precision, recall, and F-score

Languages	ADAPT			CROSS		
	Precision	Recall	F-score	Precision	Recall	F-score
Dictionary lookup	83.85	19.71	31.92	83.85	19.71	31.92
In-language EN	**78.81**	**55.82**	**65.35**	–	–	–
EN + HU	77.93	**56.60**	65.58	75.65	27.36	40.18
EN + ES	79.18	56.13	65.69	67.84	21.23	32.34
EN + DE	78.9	55.82	65.38	64.87	**36.01**	**46.31**
EN + HU + ES	81.65	54.72	65.52	81.7	28.77	42.56
EN + HU + DE	79.15	56.13	65.68	57.97	36.01	44.42
EN + ES + DE	**81.97**	55.66	**66.30**	**90.3**	23.43	37.2
EN + HU + ES + DE	80.56	55.03	65.39	80.98	23.43	36.34
Average	79.91	55.73	65.65	74.19	28.03	39.91

EN, English; DE, German; ES, Spanish; HU, Hungarian; ADAPT, language adaptation setting; CROSS: cross-language setting.

To evaluate language adaptation, our machine learning-based method was trained on the union of the source data and the candidates selected from the target language and we performed a 10-fold cross-validation at the candidate level. Each time, 90% of target data was utilized for training and the other instances were held out for testing.

The results of the experiments on language adaptation can be seen in Table 8 for English, Table 9 for German, Table 10 for Spanish, and Table 11 for Hungarian. The highest values are marked with bold.

5.5 Domain adaptation with additional language data

Here, we also investigated how we could reduce the distance between the legal domain and other domains if only limited annotated data were available for the other domains and we also exploited additional language data here. We applied a similar domain adaptation technique as in the case of the language adaptation presented in Section 5.4. We report results only for English and Hungarian, as there are other manually annotated corpora available where just English and Hungarian LVCs are manually marked.

In the case of English, the Wiki50 Corpus (Vincze *et al.* 2011b) was applied as the English target domain and the German and the English parts of the 4FX corpus were employed as the source domain as it proved the best combination during the language adaptation setting in Table 8. On the other hand, the Hungarian target domain was the short business news domain of the Szeged Corpus (Vincze and Csirik 2010). Based on the language adaptation results in Table 11, the union of the English and Hungarian parts of the 4FX corpus was selected as the source. In both settings, data from the 4FX corpus were always used as out-domain source data.

Here, we also examined how language adaptation could help the overall performance on the other domain. Therefore, we applied a two-step training. First, our model was trained on the union of the whole set of the additional language data of the 4FX corpus and the randomly selected instances of the in-language data from the 4FX corpus. First, we extended the training dataset with 10% target instances, than kept adding 25%, 50%, and 100% of the target instances. Then, the remaining part of the in-language but out-domain data was treated as test set, and we evaluated our model on it, to produce an automatically labeled (silver standard) dataset. The union of the automatically labeled test set (including in-language but out-domain data) and the additional language data (with gold standard annotation) were applied as the training dataset where our model

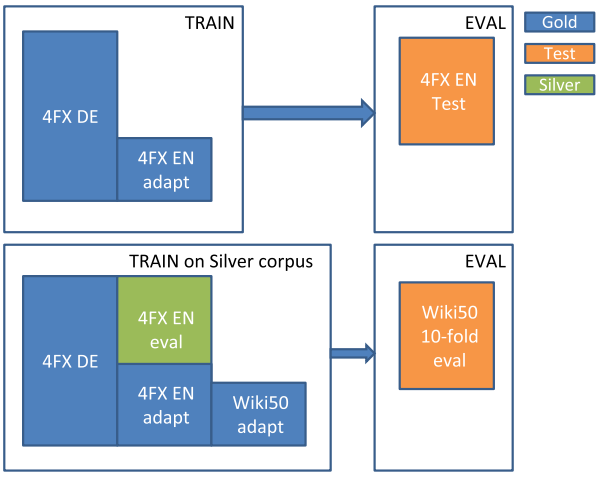

Figure 1. System architecture of the domain adaptation with additional language data.

Table 9. Experimental results when the German part of the corpus was used as target and source of different language pairs in terms of precision, recall, and F-score

Languages	ADAPT			CROSS		
	Precision	Recall	F-score	Precision	Recall	F-score
Dictionary lookup	85.71	7.45	13.71	85.71	7.45	13.71
In-language DE	**64.52**	**41.68**	**50.64**	–	–	–
DE + HU	64.4	42.44	51.17	**85.37**	**5.34**	10.06
DE + ES	**65.68**	41.98	**51.23**	24.14	13.89	17.64
DE + EN	64.48	41.83	50.74	23.26	**25.04**	24.12
DE + HU + ES	64.48	41.68	50.63	23.36	8.7	12.68
DE + HU + EN	62.22	41.83	50.03	37.62	23.66	**29.05**
DE + ES + EN	63.78	42.44	50.97	23.83	10.84	14.9
DE + HU + ES + EN	59.93	**43.36**	50.31	22.93	10.99	14.86
Average	63.57	42.22	50.73	34.36	14.07	17.62

EN, English; DE, German; ES, Spanish; HU, Hungarian; ADAPT, language adaptation setting; CROSS, cross-language setting.

was trained in the second step. Afterwards, this model was evaluated on the target dataset in a 10-fold cross-validation at the instance level.

Later on, the training dataset was extended with instances from the target domain. As we applied 10-fold cross-validation, we used 90% of target data for training for each fold and additional instances were held out for testing. Similar to the language adaptation setting, we kept adding 10%, 25%, 50%, and finally all of the training instances from the target domain, and our model was trained on the union of the additional language data, selected instances from the out-domain data and in-domain data, and was evaluated on the test set of the in-domain data in a 10-fold cross-validation manner.

We also investigated what results 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 at the instance level.

Figure 1 outlines the process how we applied additional language data to identify each individual LVC in a running text.

Table 10. Experimental results when the Spanish part of the corpus was used as target and source of different language pairs in terms of precision, recall, and F-score.

Languages	ADAPT			CROSS		
	Precision	Recall	F-score	Precision	Recall	F-score
Dictionary lookup	54.99	31.78	40.28	54.99	31.78	40.28
In-language ES	**62.99**	**45.59**	**52.90**	–	–	–
ES + HU	62.01	44.56	51.86	**51.41**	31.39	**38.98**
ES + DE	65.32	45.36	**53.54**	32.47	30.13	31.25
ES + EN	66.42	44.22	53.09	37.48	27.95	32.02
ES + HU + DE	65.21	45.02	53.26	34.62	31.84	33.17
ES + HU + EN	**67.59**	43.87	53.21	42.33	27.84	33.59
ES + DE + EN	66.77	43.87	52.95	36.32	**32.07**	34.06
ES + HU + DE + EN	66.86	43.64	52.81	37.28	31.73	34.28
Average	65.74	44.36	52.96	38.84	30.42	33.91

EN, English; DE, German; ES, Spanish; HU, Hungarian; ADAPT, language adaptation setting; CROSS, cross-language setting.

Table 11. Experimental results when the Hungarian part of the corpus was used as target and source of different language pairs in terms of precision, recall, and F-score

Languages	ADAPT			CROSS		
	Precision	Recall	F-score	Precision	Recall	F-score
Dictionary lookup	85.86	22.25	35.34	85.86	22.25	35.34
In-language HU	**80.06**	**54.32**	**64.72**	–	–	–
HU + ES	80.21	54.2	64.69	44.74	21.66	29.19
HU + DE	79.38	54.44	64.58	45.67	**51.12**	48.24
HU + EN	**80.64**	54.79	**65.25**	55.36	44.62	**49.41**
HU + ES + DE	80.13	**55.03**	**65.25**	46.12	22.49	30.23
HU + ES + EN	80.27	54.79	65.13	**67.17**	**21.07**	32.07
HU + DE + EN	80.16	54.91	65.18	49.48	44.73	46.99
HU + ES + DE + EN	80.05	54.79	65.05	39.13	28.76	33.15
Average	80.12	54.71	65.02	49.67	33.49	38.47

EN, English; DE, German; ES, Spanish; HU, Hungarian; ADAPT, language adaptation setting; CROSS, cross-language setting.

6. Results

In this section, we will present the results of the methods applied on different languages in the 4FX corpus. We will also report the results of language adaptation and domain adaptation with additional language data.

6.1 Results of language adaptation

Tables 8–11 list the results for the language adaptation. Based on the in-domain 10-fold cross-validation results, our method can achieve relatively the same results on the Hungarian and English parts of the corpus with F-scores of about 65, and the same results on the German and Spanish parts of the corpus with F-scores of about 50. On each corpus, the language adaptation

achieved an F-score that was higher than that of the dictionary lookup approach. Moreover, except for the Spanish case, the cross-experiments also yielded higher F-scores than in the dictionary lookup case.

Table 8 shows the results on the English corpus, where all the other three languages can contribute to the overall performance with the language adaptation. This method was most effective with an F-score of 66.30 with the highest precision score (81.97%), when we applied the union of the Spanish and German parts of the 4FX corpus as the source. In the case of the cross-experiments, the best results we got were when we trained our model on German. The difference between the average results of the CROSS and those of the dictionary lookup experiments was 7.99. The results of the baseline dictionary lookup method were noticeably exceeded by the ADAPT results.

We can see the results of German experiments in Table 9, where the largest difference is between the average language adaptation and cross-language experimental settings as the difference was 33.11 in terms of F-score. Here, our approach proved the most effective when we extended the German training set with the Spanish part of the 4FX corpus, since in this case, the F-score was 0.59 higher than the in-domain German results. The average difference between dictionary lookup and cross-experiments was 3.91, but with the best combination the gap was 15.34.

Table 10 lists the results for Spanish experiments. Here, our method was the most effective when the Spanish training set was extended with German. In this case, primarily the precision was affected by the language adaptation as the average precision score was 2.75 higher than the in-domain precision. Moreover, this difference in the precision score is 4.60 when the training set was extended with the union of the English and Hungarian corpora. The language adaptation results exceeded the cross-experiments by 19.05. But in this case, the dictionary lookup outperformed the cross-experiments as it achieved an F-score of 40.28.

Table 11 shows the results got on the Hungarian part of the 4FX corpus by using baseline dictionary lookup and our language adaptation. Language adaptation achieved the same F-score (65.25) with different precision and recall scores, when we extended the Hungarian training set with the English corpus and the union of the German and Spanish parts of the 4FX corpus. The difference between the average results of the language adaptation and the Hungarian in-domain experiments was 0.30, but the method generally worked better when the training set was not just extended with one language. The dictionary lookup approach got the highest precision, namely 85.86% but with the average cross-experimental setting we achieved an F-score that was 3.13 points higher.

The size of the target language data added to the source datasets greatly influenced the results as shown in Figure 2 with four typical settings in the four different languages. As we can see, language adaptation is effective when there is only a limited amount of target data available, since the target data contribute to the training process and the models produced increasing results.

Tables 12–15 show the usefulness of each individual feature type in four languages in our in-domain and cross-experiments. The performance scores of the features were compared with those obtained by applying all features described in Section 5.2. Here, we trained a J48 classifier with all of the features except that one, for each feature type. Then, we compared the performance to that got with all the features.

As our ablation analysis shows, each type of feature contributed to the overall performance in each language. We found only two exceptions. First, in English cross-experimental results, the statistical features had a detrimental effect and second, in the case of Spanish, the machine learning-based method could achieve a slightly better result (+0.03) without syntactic features.

From the tables for ablation analysis, we ascertain that the lexical features were the most important in all languages, as here it has the largest difference. For example, in English, the language adaptation ablation results obtained an F-score that was 39.2 points lower when the lexical features were ignored. However, these features also play an important role in German and Spanish but to a lesser extent. For instance, in the case of German we obtained an F-score that was 20 points lower

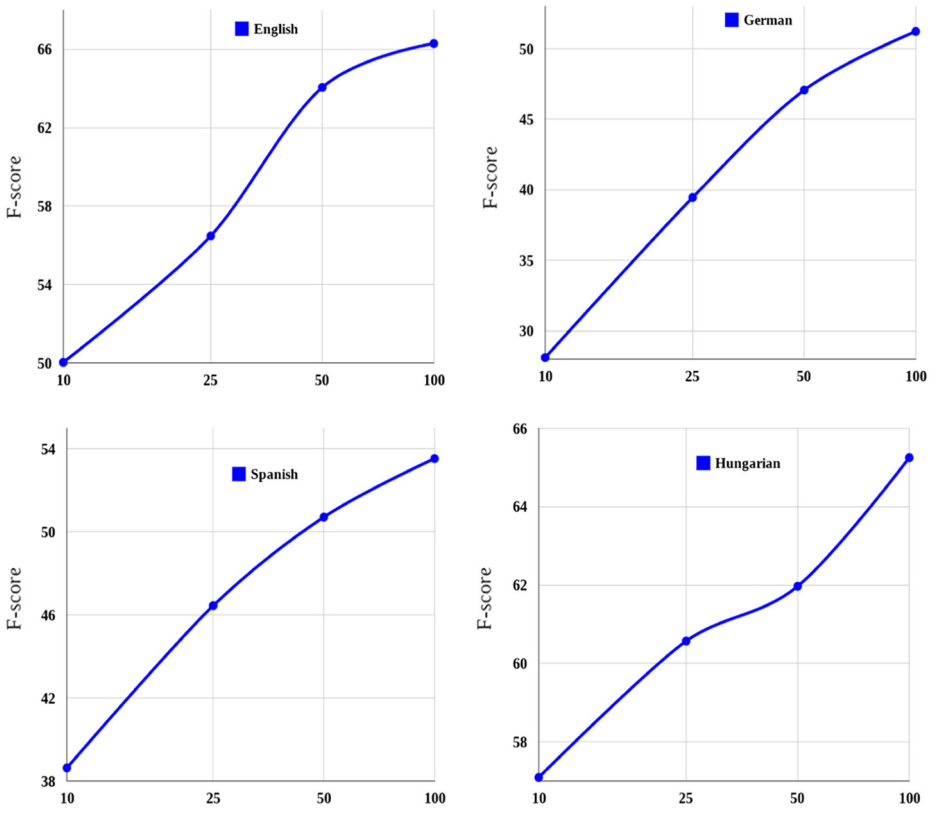

Figure 2. The effect of the size of the target language data on detecting LVCs in four different languages. English: the union of the German and Spanish parts of 4FX was the source. German: the Spanish part of 4FX was the source. Spanish: the German part of 4FX was the source. Hungarian: the English part of 4FX was the source.

when we omitted lexical-based features, but in this case, the orthographic and syntactic features had a bigger influence than they did for the other languages.

In English, the precision scores were less affected by the cross-experimental ablation analysis compared to the in-domain experiments. In contrast, the omission of features had a greater effect on the precision scores in Hungarian cross-experiments than they did on the in-domain precision scores. Moreover, the language-specific features proved more useful in German and Spanish than they did in English and Hungarian.

6.2 Results of domain adaptation with additional language data

Table 16 lists the domain adaptation results got using additional language data on the English corpora, while Table 17 shows those for the Hungarian corpora. In the case of English, Wiki50 was the target, while the union of the English, German, and Spanish parts of the 4FX corpus was the source.

As Table 16 shows, when the gold annotated target language instances were increased in the source domain, the average difference between the in-domain and domain adaptation results consistently increased. When only 10% of target language was used in the source, the average difference in F-score was 0.09, and when all of the source data were used, it was 1.76. Moreover, in the latter case, we can outperform the in-domain results by 0.47. Here, the cross-domain experimental settings can achieve an F-score of 41.05, while the in-domain setting yielded an F-score of 53.89.

Table 12. The usefulness of individual features in terms of precision, recall, and F-score using the English part of 4FX

	In-domain			ADAPT DE+ES		
	Precision	Recall	F-score	Precision	Recall	F-score
All	78.81	55.82	65.35	81.97	55.66	66.3
Orthographic	−3.42	1.26	−0.38	−0.33	−0.01	−0.12
Lexical	−33.53	−26.57	−29.81	−17.28	−38.52	−39.2
Morphological	−2.41	0.94	−0.22	−1.39	0.47	−0.13
Syntactic	−5.74	0.63	−1.66	−2.00	−1.41	−1.66
Statistical	−4.26	1.41	−0.6	−0.21	0.94	**0.60**
Language spec.	−4.22	1.57	−0.48	−0.43	−0.44	−0.45

Table 13. The usefulness of individual features in terms of precision, recall, and F-score using the German part of 4FX

	In-domain			ADAPT ES		
	Precision	Recall	F-score	Precision	Recall	F-score
All	64.52	41.68	50.64	50.64	41.98	51.23
Orthographic	−10.45	−8.55	−9.55	−13.42	−7.48	−9.66
Lexical	−3.16	−20.46	−19.1	−4.32	−20.76	−19.69
Morphological	−1.17	−0.31	−0.58	−0.24	0	−0.09
Syntactic	−7.66	−6.26	−6.99	−4.93	−6.1	−6.12
Language spec.	−4.19	−2.14	−2.87	−0.31	−8.24	−6.72

Table 14. The usefulness of individual features in terms of precision, recall, and F-score using the Spanish part of 4FX

	In-domain			ADAPT DE		
	Precision	Recall	F-score	Precision	Recall	F-score
All	62.99	45.59	52.9	65.32	45.36	53.54
Orthographic	0.01	−6.64	−4.76	−1.14	−6.18	−4.89
Lexical	−17.57	−23.14	−22.85	−17.98	−27.95	−28.08
Morphological	−3.58	−3.44	−3.59	−2.87	−4.7	−4.29
Syntactic	0.31	−0.11	0.03	−1.14	−0.13	−0.48
Language spec.	−0.12	−1.95	−1.38	−1.38	−3.09	−2.65

In the case of Hungarian, the short business news domain from the Szeged Corpus was applied as the target corpus, and the union of the English and Hungarian parts of the 4FX corpus was the source. The DAAL column of Table 17 lists the results obtained for the Hungarian domain adaptation with the additional language data task. Here, the average differences also consistently increased between in-domain results and domain adaptation with additional language data results when we extended the gold standard target language instances in the source. However, the most obvious difference was demonstrated when the whole of the source data was applied as the gold standard. Here, the average F-score difference is 2.72. Moreover, the in-domain setting achieved

Table 15. The usefulness of individual features in terms of precision, recall and F-score using the Hungarian part of 4FX

	In-domain			ADAPT EN		
	Precision	Recall	F-score	Precision	Recall	F-score
All	80.06	54.32	64.72	80.64	54.79	65.25
Orthographic	−4.51	−0.12	−1.6	−7.47	1.42	−1.67
Lexical	−27.31	−27.22	−28.91	−36.37	−17.28	−24.64
Morphological	−0.99	−0.24	−0.49	−4.53	1.73	−0.38
Syntactic	−1.19	−0.71	−0.89	−5.15	1.78	−0.58
Language spec.	0.57	−1.54	−0.92	−3.31	0.36	−0.87

Table 16. Results of domain adaptation with additional language data on the English Wiki50 corpus in terms of precision, recall, and F-score

EN (%)	DE+ES (%)	Wiki50 (%)	ID			DAAL		
			P	R	F	P	R	F
10	100	100	56.69	51.36	53.89	55.41	50.54	52.86
10	100	50	54.74	48.91	51.66	50.69	52.72	51.69
10	100	25	51.33	44.29	47.55	48.39	47.55	47.97
10	100	10	46.10	40.49	43.11	44.94	43.21	44.05
25	100	100	56.69	51.36	53.89	55.21	50.82	52.92
25	100	50	54.74	48.91	51.66	52.56	51.90	52.23
25	100	25	51.33	44.29	47.55	48.95	47.74	48.34
25	100	10	46.10	40.49	43.11	46.93	44.02	45.43
50	100	100	56.69	51.36	53.89	57.38	50.54	53.75
50	100	50	54.74	48.91	51.66	52.90	51.63	52.26
50	100	25	51.33	44.29	47.55	52.04	46.47	49.10
50	100	10	46.10	40.49	43.11	46.60	45.92	46.26
100	100	100	56.69	51.36	53.89	57.40	51.63	**54.36**
100	100	50	54.74	48.91	51.66	54.30	50.82	52.50
100	100	25	51.33	44.29	47.55	53.70	46.74	49.98
100	100	10	46.10	40.49	43.11	50.82	42.66	46.39
100	100	0	–	–	–	44.20	38.32	**41.05**
0	0	100	–	–	–	56.69	51.36	**53.89**

EN, English part of 4FX; ES+DE, the union of the Spanish and German parts of 4FX; Wiki50, Wiki50 corpus; ID, in-domain results; DAAL, domain adaptation with additional language data

an F-score of 57.32, which was exceeded by the DAAL results of 1.83 more, while the cross-domain experiments yielded an F-score of 42.37.

The size of the target data added to the source data sets greatly influenced the results as shown using two settings in Figures 3 and 4. Figure 3 shows the Wiki50 target, cross-experiments and domain adaptation with additional language data results when the source was the whole set of union of the English, German, and Spanish parts of the 4FX corpus.

Table 17. Results of domain adaptation with additional language data on the Hungarian short business news corpus in terms of precision, recall, and F-score

			ID			DAAL		
HU (%)	EN (%)	newsml (%)	P	R	F	P	R	F
10	100	100	67.79	49.66	57.32	67.79	50	57.55
10	100	50	65.15	46.33	54.15	63.51	47.36	54.26
10	100	25	60.39	44.15	51.01	60.68	46.33	52.54
10	100	10	59.41	35.21	44.21	53.87	40.48	46.23
25	100	100	67.79	49.66	57.32	67.50	48.39	56.37
25	100	50	65.15	46.33	54.15	64.93	47.71	55.00
25	100	25	60.39	44.15	51.01	61.14	48.74	54.24
25	100	10	59.41	35.21	44.21	54.09	41.28	46.83
50	100	100	67.79	49.66	57.32	65.21	51.03	57.26
50	100	50	65.15	46.33	54.15	62.66	50.69	56.04
50	100	25	60.39	44.15	51.01	60.83	49.66	54.68
50	100	10	59.41	35.21	44.21	53.26	42.09	47.02
100	100	100	67.79	49.66	57.32	61.89	56.65	**59.15**
100	100	50	65.15	46.33	54.15	59.53	53.33	56.26
100	100	25	60.39	44.15	51.01	55.63	51.61	53.54
100	100	10	59.41	35.21	44.21	50.32	47.02	48.62
100	100	0	–	–	–	50.56	36.47	**42.37**
0	0	100	–	–	–	67.79	49.66	**57.32**

HU, Hungarian; EN, English; newsml, Hungarian short business news corpus; ID, in-domain results; DAAL, domain adaptation with additional language data.

The gap between the in-domain results and domain adaptation with additional language data results progressively decreases as the size of the dataset increases. But the gap between the results of CROSS and those of DAAL progressively increases with the amount of the data added.

Figure 4 lists the results by applying the Hungarian short business news domain as the target, cross-experiments and domain adaptation when the source was the union of the all of Hungarian and English part of the 4FX corpus. The results got from this model also exceeded the TARGET results, when all of the target instances were added to the training data as it yielded an F-score 1.83 that was higher than the in-domain results.

7. Discussion

Here we discuss our results achieved in different settings, namely, in-domain, cross-language, language adaptation and domain adaptation with additional language data settings.

7.1 In-domain experiments

Our machine learning method extensively outperformed the dictionary lookup baseline method in all the four languages in the 4FX corpus, which demonstrates that our syntax-based machine-learning method can be readily applied to automatic LVC detection in different languages.

Interestingly, our method can perform better in the case of English and Hungarian than German and Spanish. As Table 6 shows, the syntax-based candidate extraction method can extract only 60.80% of German LVCs, however, it extracted 87.10% of the Hungarian LVCs. In those

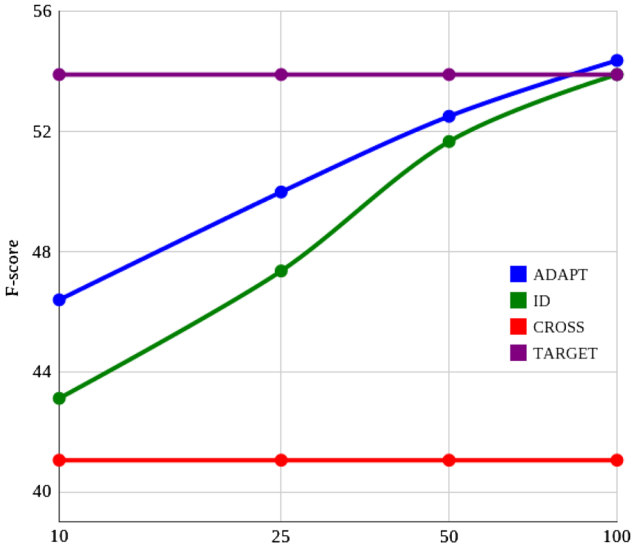

Figure 3. The effect of the size of the target data on detecting LVCs in Wiki50. ADAPT, domain adaptation with additional language data setting; ID, training on a limited set of target data; CROSS, cross-domain setting; TARGET, in-domain setting.

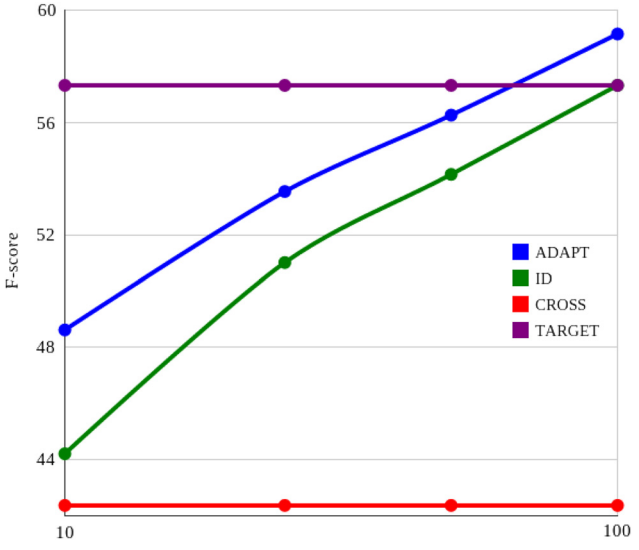

Figure 4. The effect of the size of the target data on detecting LVCs in the short business news domain from Szeged Corpus. ADAPT, domain adaptation with additional language data setting; ID, training on a limited set of target data; CROSS, cross-domain setting; TARGET, in-domain setting.

languages, where the candidate extraction method proved to be less effective, the classification method reached lower recall scores and achieved modest F-score results. We think that the less effective dependency parser is probably responsible for the less impressive candidate extraction results on German and Spanish: the parser did not find any direct dependency relations in 29.01% of the German LVCs and 25.00% of Spanish LVCs. In contrast, in the case of English and Hungarian, the dependency parsers did not detect any syntactic relations for only 10% of LVCs. However, the automatic detection of German LVCs proved to be the most difficult task, as the dictionary lookup and the cross-language experiments also yielded an F-score of less than 20. It is mainly due to the relatively high ratio of the hapax LVCs, which is also a possible limit for the dictionary lookup method to achieve a higher recall score. Besides, the German light verbs were the most diverse, as we found 93 unique light verbs among the manually annotated German LVCs. This also prevents the machine learning-based method from providing good results as the lexical features are the most useful ones here.

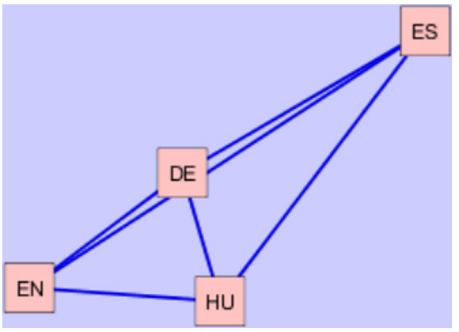

Figure 5. Language similarity measure based on Kendall's coefficients.

Another possible reason for the modest performance on German and Spanish is that the average occurrences of LVCs are fewer than in English and Hungarian. So, the machine learning-based method encounters fewer examples of the particular LVCs in these languages, which may hinder the learning process.

7.2 Cross-language and language adaptation experiments

The results of our cross-language experiments exceeded those of the dictionary lookup method. This highlights the fact that a machine learning-based model trained on a different language can be more effective than a target language dictionary lookup method. Moreover, to achieve a higher recall score, the dictionary lookup approach was primarily limited by the size of the available dictionaries, but this method still managed to yield a fairly good precision score. There was one notable exception: in the case of Spanish, the dictionary lookup method performed better than the cross-language experiments as it achieved a relatively high recall score. This may be attributed to the big size of the dictionary on the one hand and to the building principles behind the dictionary on the other, which were based on the theory of lexical functions (see, e.g., Melčuk 1974), which was also taken into account when designing the annotation principles for 4FX.

Moreover, the cross-language training by itself did not prove sufficient in many cases. Therefore, the inclusion of annotated target language data into the training dataset was necessary to reduce the gap among different languages. As Figure 2 shows, by adding some target language data to the training dataset, it is possible to achieve similar results to those for in-domain settings.

The cross-experiments and language adaptation results showed that the proper selection of the source language largely influences the performance of the automatic detection of LVCs. The Kendall's coefficients of the union of the light verbs were calculated to demonstrate the differences or similarities of languages, which is shown in Figure 5. This figure can help us to select the proper source language. Based on the Kendall's coefficients, the German LVCs are the most similar to English, which accords with the cross-language experimental results in Table 8. Here, the model trained on the German part of the 4FX corpus achieved the best cross-language results with an F-score of 46.31.

With Figure 2, we would expect Hungarian cross-language experiment settings to perform the best on German since Hungarian proved to be the most similar based on the Kendall's coefficient of the light verbs. But the results in Table 9 suggest that the English LVCs are the most similar to German, since the model trained on the English part of 4FX yielded higher cross-language results with an F-score of 24.12. Still, the precision scores of the German cross-language experiments demonstrate a clear relationship between Hungarian and German: the Hungarian setting achieved a much higher precision score with 85.37% than any of the other languages. So, some German examples are very similar to Hungarian LVCs—they are literal translations, such as *in Verkehr bringen* and *forgalomba hoz* (into traffic bring) "put into circulation" or *in Kraft treten* and *hatályba lép* (into force step) "enter into force"—and the model trained on the Hungarian

examples was able to detect these with a relatively high precision. However, there were only a handful of such cases, which resulted in a relatively low recall score.

In the case of the Spanish cross-language experimental results, the Hungarian model proved to be the most successful. But, we were able to achieve the best results in the language adaptation experiments when German was selected as the source language, which was the most similar to Spanish according to the Kendall's coefficients. In the Hungarian cross-language experiments, the two most similar languages, English and German yielded the best results.

As illustrated by our cross-language experiments, the model trained on the Hungarian part of 4FX mostly had a positive effect on the precision scores while the German model improved the recall scores. This may be due to the fact that German light verbs are the most diverse, hence, German data can provide many different examples of possible LVCs and also data sparsity for hapax LVCs in the target language may be resolved to little extent.

As Table 8 indicates, language adaptation outperformed the in-domain English results in every case. As the English part of the 4FX corpus contains the fewest gold annotated LVCs in the corpus, the machine learning-based method achieved a better result when we extended the training set with instances from other languages.

As regards German, the cross-language experiments performed the weakest, which is mainly due to the weak recall scores. This may be due to the facts that were mentioned in Section 7.1: the ratio of the hapax LVCs is relatively high in the German corpus and the light verbs were the most diverse here concerning the four languages examined. Based on the data in Table 5, the distribution of German LVCs is rather balanced and this language seemed to contain the most heterogeneous LVCs. Thereby, the machine learning-based model trained on the additional language model was not able to provide acceptable recall scores, as there were no LVCs with high frequency that may cover the majority of LVC occurrences. It is also seen that the cross-language experiment setting is the most successful when we trained our approach on the union of the English and Hungarian corpora. This is due to the specific characteristics of the two languages: when we applied the Hungarian cross-settings, our method yielded the highest precision score (85.37%), and achieved the best recall score with 25.04% when we trained it just on the English corpus.

It is well demonstrated how difficult the automatic detection of the Spanish LVCs is, as the best recall score was achieved in the in-domain setting. Here, the cross-language experiments only had a positive effect on the precision scores. This is probably due to the same fact as in German: namely, for Spanish the hapax LVCs and light verbs occurred in a relatively high ratio, so the additional language data could not assist the machine learning-based detection of rare examples. Also, the average precision score of the language adaptation results is just 2.75 points higher than in the in-domain case.

As for Hungarian, the most successful cross-language experimental results were achieved with English as the source language, and the union of the English and Spanish datasets had the most beneficial effect on the precision score. This may be explained by the fact that the English in-domain model could also achieve a very high precision score, so from the English dataset, the model may learn how to select genuine LVCs from the candidates. Still, applying German as the training set again had a beneficial effect on the recall score, most likely because of the heterogeneity of its LVCs.

7.3 Ablation analysis

As our ablation analysis revealed, the most important feature type was the lexical feature type. This is not surprising as most typical verbs in LVCs belong to a well-defined set, which are present in a high percentage of data in each language.

Syntactic features proved to be the most effective in the case of German. This may be attributed to the fact that the ratio of split LVCs is the highest in German, which means that components of LVCs are often not adjacent and syntactic information helps us a lot to match these parts.

In the case of Hungarian, we observe that for the language adaptation setting, recall improves with the omission of features, except for the lexical ones. This is again probably due to the fact that the English model mainly improves precision and its effect on recall is less clear.

As for language-specific features, it is noticeable that in the case of English, they improve precision but harm recall whereas they behave in the opposite manner in the case of Hungarian. This may be because in English, one of the language-specific features is whether the given verb is used as an auxiliary, which relies on POS-tags. Thus, selecting only those verbs that are not auxiliaries can improve precision because the set of candidate verbs is filtered. On the other hand, in Hungarian, the feature *historical derivation* yielded some false positive cases that included nouns of verbal origin but they are now used in a lexicalized meaning and do not form part of LVCs. These cases had a harmful effect on precision, however, this feature has a beneficial effect on recall as many nouns of verbal origin could be identified in this way.

7.4 Domain adaptation with additional language data

Our English and Hungarian cross-domain experimental results highlighted the fact that domain dependence is also important for the automatic detection of LVCs since the cross-domain results were always worse than the corresponding in-domain results. However, when there is only a limited amount of target data available, domain adaptation is more effective since the out-domain dataset also contributes to the training process. Moreover, training only on the amount of target data cannot achieve such outstanding results. On both languages, the domain adaptation experiments extended with additional language data exceeded the in-domain results when the machine learning-based model was trained on the entire set of gold standard source data with 100% of the target data. However, when we trained our model on the whole set of target domain, it did not seem to benefit from the silver standard out-domain data: it just confused learning since the in-domain results exceeded the domain adaptations. But, in other cases, the beneficial effect of the domain adaptation could be observed when we applied the silver standard as source domain.

8. Conclusions

In this paper, we presented a machine learning-based method that allows the full coverage identification of LVCs in raw texts of different languages. The method applied essentially depends on an English-based approach which was adapted to German, Spanish, and Hungarian with language-specific features. In order to see how much different languages affected each other, we also implemented a language-independent representation of LVCs and features.

Our method solved the problem in a two-step approach, so we implemented a standardized candidate extraction method on different languages in the first step. We found that the efficiency of the machine learning-based method mainly depends on the dependency parsers applied as the candidate extraction method was based on the dependency labels. The somewhat lower quality of parsing was responsible for our modest F-score results on German and Spanish, compared to those got for English and Hungarian, where we were able to achieve better results in terms of F-score. Furthermore, the automatic detection of German LVCs proved to be the most difficult task as the dictionary lookup, the cross-language experiments, and the in-domain machine learning-based method were the weakest here. This may be due to the facts that German light verbs were the most diverse, the ratio of the hapax LVCs is relatively high, the average occurrences of LVCs were fewer than these in English and Hungarian, and the ratio of hapax light verbs was much higher in German than other languages.

We also found that the proper selection of the source language largely influenced the performance of the automatic detection of LVCs in the target language. We saw that the similarity of the different languages based on the Kendall's coefficient got by comparing the light verbs used in different languages can help us in our choice.

As our cross-language experimental results outperformed the dictionary lookup results, it highlights the fact that a machine learning-based model trained on data from (an)other language(s) can indeed achieve better results than a target language dictionary lookup method. However, the cross-language training by itself did not prove sufficient in many cases. This was why we applied a simple language adaptation technique to reduce the gap between different language pairs and we found that it was possible to achieve similar results as in the case of in-domain settings.

For English and Hungarian, we also experimented with cross-domain settings, which revealed that the domain dependence is also essential for the automatic detection of LVCs since the cross-domain results were always worse than the corresponding in-domain results. Nevertheless, when training on the combination of out-domain and additional language data, we were able to achieve practically the same results as in the case of in-domain training, which again demonstrates the portability of our model and usefulness of our standardized feature set across languages and domains. Thus, our approach can prove useful in cases where there is a small amount of annotated data for the given language or domain but there is an ample amount of annotated data for another language and domain. As for LVC detection, this means that if we would like to develop an LVC detector for another language, for example, Italian, we only have to annotate a few hundreds of Italian LVCs and apply language adaptation with one or more of the previously annotated corpora.

From a linguistic point of view, we argued that the very same feature representation of a given phenomenon might be suitable for typologically different languages as well. In this way, we were also able to make interesting comparisons across languages concerning the distribution of LVCs and light verbs in the 4FX corpus, which could assist theoretical research in contrastive linguistics and lexicology. The annotated corpus also allows one to study legal language in its entirety, or just LVCs in the legal domain.

In the future, we would like to improve our methods, with special regard to the candidate extraction phase for Spanish and German. we intend to test our approach on the PARSEME shared task corpora, which were prepared on the basis of the same annotation guidelines, hence provide a perfect testbed for cross-lingual comparisons. We also would like to extend our approach to other languages as well and we would also like to apply language adaptation techniques to other types of MWEs and even test our approach in other NLP tasks such as information extraction. We also think that our results may be incorporated in machine translation systems as well and improve their overall performance.

Acknowledgements. This research was supported by the UNKP-17-4 New National Excellence Program of the Ministry of Human Capacities.

References

Alonso Ramos M. (2000). Verbos de apoyo, funciones léxicas y traducción automática. *Revista de lexicografía* **6**, 155–177.

Alonso Ramos M. (2004). *Las construcciones con verbo de apoyo*. Madrid: Visor Libros.

Al Saied H., Constant M. and Candito M. (2017). The ATILF-LLF system for parseme shared task: A transition-based verbal multiword expression tagger. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain: Association for Computational Linguistics, pp. 127–132.

Bannard C. (2007). A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of MWE 2007*, Morristown, NJ, USA: Association for Computational Linguistics, pp. 1–8.

Belvin R. S. (1993). The two causative haves are the two possessive haves. In *Papers from the Fifth Student Conference in Linguistics*, vol. 20, Cambridge: MITWPL, pp. 19–34.

Berk G., Erden B. and Güngör T. (2018). Deep-BGT at PARSEME shared task 2018: Bidirectional LSTM-CRF model for verbal multiword expression identification. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 248–253.

Blanco Escoda X. (2000). Verbos soporte y clases de predicados en español. *LEA* **22**, 99–117.

Bohnet B. (2010). Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of Coling 2010*, Beijing, China: Coling 2010 Organizing Committee, pp. 89–97.

Boroș T. and Burtica R. (2018). GBD-NER at PARSEME shared task 2018: Multi-word expression detection using bidirectional long-short-term memory networks and graph-based decoding. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 254–260.

Boroș T., Pipa S., Barbu Mititelu V. and Tufiș D. (2017). A data-driven approach to verbal multiword expression detection. PARSEME Shared Task system description paper. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain: Association for Computational Linguistics, pp. 121–126.

Bosque I. (2001). On the weight of light verb predicates. In Zagona K., Maléln E. and Herschenson J. (eds), *Features and Interfaces in Romance*, Amsterdam: Benjamins, pp. 23–38.

Brants S., Dipper S., Eisenberg P., Hansen-Schirra S., König E., Lezius W., Rohrer C., Smith G. and Uszkoreit H. (2004). TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation* **2**(4), 597–620.

Buckingham L. (2009). *Las construcciones con verbo soporte en un corpus de especialidad*. Frankfurt am Main – Bern – Bruxelles – New York – Wien: Peter Lang.

Bußmann H. (2002). *Lexikon der Sprachwissenschaft*. Stuttgart: Alfred Kröner.

Butt M. and Lahiri A. (2013). Diachronic Pertinacity of Light Verbs. *Lingua* **135**, 7–29.

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 LREC 2002*, Las Palmas, Spain: European Language Resources Association (ELRA), pp. 1934–1940.

Daniels K. (1963). *Substantivierungstendenzen in der deutschen Gegenwartssprache: Nominaler Ausbau des verbalen Denkkreises*. Düsseldorf: Schwann.

Danlos L. (2010). Extension de la notion de verbe support. In Nakamura T., Laporte E., Dister A. and Fairon C. (eds), *Les Tables, La grammaire par le menu, Volume d'hommage à Christian Leclère*, Louvain: Presses Universitaires de Louvain, pp. 81–90.

Duden. (2006). *Der Duden in 12 Bänden. Das Standardwerk zur deutschen Sprache: Duden 06. Das Aussprachewörterbuch: Unerlässlich für die richtige Aussprache. Betonung ... Namen: Bd 6 (Duden Series Volume 6): Band 6*. Gebundene Ausgabe, Mannheim: Bibliographisches Institut (F.A. Brockhaus).

É. Kiss K. (2002). *The Syntax of Hungarian*. Cambridge: Cambridge University Press.

Ehren R., Lichte T. and Samih, Y. (2018). Mumpitz at PARSEME shared task 2018: A bidirectional LSTM for the identification of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 261–267.

Fazly A. and Stevenson S. (2007). Distinguishing subtypes of multiword expressions using linguistically-motivated statistical measures. In *Proceedings of MWE 2007*, Prague, Czech Republic: Association for Computational Linguistics, pp. 9–16.

Fleischer W., Helbig G. and Lerchner G. (2001). *Kleine Einzyklopädie. Deutsche Sprache*. Frankfurt am Main – Berlin – Bruxelles – New York – Wien: Peter Lang.

Hale K. and Keyser S.J. (2002). *Prolegomenon to a Theory of Argument Structure*. Cambridge: MIT Press.

Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P. and Witten I.H. (2009). The WEKA data mining software: An update. *SIGKDD Explorations* **11**(1), 10–18.

Häusermann J. (1977). *Hauptprobleme der deutschen Phraseologie auf der Basis sowjetischer Forschungsergebnisse*. Tübingen: M. Niemeyer.

Heine A. (2006). *Funktionsverbgefüge in System, Text und korpusbasierter (Lerner-)Lexikographie*. Frankfurt am Main: Peter Lang.

Helbig G. and Buscha J. (2001). *Deutsche Grammatik. Ein Handbuch für den Ausländerunterricht*. Berlin and München: Langenscheidt.

Hwang J.D., Bhatia A., Bonial C., Mansouri A., Vaidya A., Xue N. and Palmer M. (2010). PropBank annotation of multilingual light verb constructions. In *Proceedings of the Fourth Linguistic Annotation Workshop*, Uppsala, Sweden: Association for Computational Linguistics, pp. 82–90.

Kearns K. (2002). *Light verbs in English*. Manuscript.

Kim S.N. (2008). *Statistical Modeling of Multiword Expressions*. PhD thesis, Melbourne: University of Melbourne.

Klyueva N., Doucet A. and Straka M. (2017). Neural networks for multi-word expression detection. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain: Association for Computational Linguistics, pp. 60–65.

Kolesnikova O. and Gelbukh A. (2010). Supervised machine learning for predicting the meaning of verb-noun combinations in Spanish. In *Advances in Soft Computing*, Berlin – Heidelberg: Springer, pp. 196–207.

Krenn B. (2008). Description of evaluation resource – German PP-verb data. In *Proceedings of MWE 2008*, Marrakech, Morocco: European Language Resources Association (ELRA), pp. 7–10.

Langer S. (2005). A formal specification of support verb constructions. In Langer S. and Schnorbusch D. (eds), *Semantik im Lexikon*, Tübingen: Gunter Narr Verlag, pp. 179–202.

Leoni de León, J.A. (2014). Lexical-syntactic analysis model of Spanish multi-word expressions. In Nolan B. and Periñán-Pascual C. (eds), *Language Processing and Grammars. The role of functionally oriented computational models*, Amsterdam: Benjamins, pp. 39–77.

Maldonado A., Han L., Moreau E., Alsulaimani A., Chowdhury K.D., Vogel C. and Liu Q. (2017). Detection of verbal multi-word expressions via conditional random fields with syntactic dependency features and semantic re-ranking. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain: Association for Computational Linguistics, pp. 114–120.

Marimon M., Fisas B., Bel N., Arias B., Vázquez S., Vivaldi J., Torner S., Villegas M. and Lorente M. (2012). The IULA Treebank. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA), pp. 1920–1926.

McDonald R., Nivre J., Quirmbach-Brundage Y., Goldberg Y., Das D., Ganchev K., Hall K., Petrov S., Zhang H., Täckström O., Bedini C., Bertomeu Castelló N. and Lee J. (2013). Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Sofia, Bulgaria: Association for Computational Linguistics, pp. 92–97.

Mel'čuk I. (2005). Verbes supports sans peine. *Lingvisticae Investigationes* **27**(2), 203–217.

Mel'čuk I. (1974). Esquisse d'un modèle linguistique du type "Sens<->Texte". In *Problèmes actuels en psycholinguistique*. Colloques inter. du CNRS, no. 206, Paris: CNRS, pp. 291–317.

Mel'čuk I., Clas A. and Polguère A. (1995). *Introduction à lexicologie explicative et combinatoire*. Louvain-la-Neuve, France: Duculot.

Meyers A., Reeves R. and Macleod C. (2004). NP-External arguments: A study of argument sharing in English. In *Proceedings of MWE 2004*, Barcelona, Spain: Association for Computational Linguistics, pp. 96–103.

Moreau E., Alsulaimani A., Maldonado A. and Vogel C. (2018). CRF-Seq and CRF-DepTree at PARSEME shared task 2018: Detecting verbal MWEs using sequential and dependency-based approaches. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 241–247.

Nagy T. I. and Vincze V. (2011). Identifying verbal collocations in Wikipedia articles. In *Proceedings of the 14th International Conference on Text, Speech and Dialogue*, Berlin, Heidelberg: Springer-Verlag, pp. 179–186.

Nagy T. I., Vincze V. and Berend G. (2011). Domain-dependent identification of multiword expressions. In *Proceedings of RANLP 2011*, Hissar, Bulgaria: RANLP 2011 Organising Committee, pp. 622–627.

Nagy T. I., Vincze V. and Farkas R. (2013). Full-coverage identification of English light verb constructions. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, Nagoya, Japan: Asian Federation of Natural Language Processing, pp. 329–337.

Nerima L., Foufi V. and Wehrli E. (2017). Parsing and MWE detection: Fips at the PARSEME shared Tas. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain: Association for Computational Linguistics, pp. 54–59.

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

Rácz A., Nagy T. I. and Vincze V. (2014). 4FX: Light verb constructions in a multilingual parallel corpus. In Calzolari N., Choukri K., Declerck T., Loftsson H., Maegaard B., Mariani J., Moreno A., Odijk J. and Piperidis S. (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland: European Language Resources Association (ELRA).

Ramisch C., Cordeiro S.R., Savary A., Vincze V., Barbu Mititelu V., Bhatia A., Buljan M., Candito M., Gantar P., Giouli V., Güngör T., Hawwari A., Iñurrieta U., Kovalevskaitė J., Krek S., Lichte T., Liebeskind C., Monti J., Parra Escartín C., QasemiZadeh B., Ramisch R., Schneider N., Stoyanova I., Vaidya A. and Walsh A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 222–240.

Real Academia Española, Asociación de Academias de la Lengua Española. (2009). *Nueva Gramática de la Lengua Española*. Madrid: Espasa Libros.

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 CICLing 2002*, Berlin – Heidelberg – New York: Springer, pp. 1–15.

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

Savary A., Ramisch C., Cordeiro S., Sangati F., Vincze V., QasemiZadeh B., Candito M., Cap F., Giouli V., Stoyanova I. and Doucet A. (2017). The PARSEME shared task on automatic identification of verbal multiword expressions. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain: Association for Computational Linguistics, pp. 31–47.

Simkó K.I., Kovács V. and Vincze V. (2017). USzeged: Identifying verbal multiword expressions with POS tagging and parsing techniques. In *Proceedings of the 13th Workshop on Multiword Expressions (MWE 2017)*, Valencia, Spain: Association for Computational Linguistics, pp. 48–53.

Steinberger R., Pouliquen B., Widiger A., Ignat C., Erjavec T. and Tufiş D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of LREC 2006*, Genova, Italy: European Language Resources Association (ELRA), pp. 2142–2147.

Stevenson S., Fazly A. and North R. (2004). Statistical measures of the semi-productivity of light verb constructions. In *MWE 2004*, Barcelona, Spain: Association for Computational Linguistics, pp. 1–8.

Stodden R., QasemiZadeh B. and Kallmeyer L. (2018). TRAPACC and TRAPACCS at PARSEME shared task 2018: Neural transition tagging of verbal multiword expressions. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 268–274.

Surdeanu M., Johansson R., Meyers A., Màrquez L. and Nivre J. (2008). The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning*, Association for Computational Linguistics, pp. 159–177.

Szarvas G., Vincze V., Farkas R., Móra G. and Gurevych I. (2012). Cross-genre and cross-domain detection of semantic uncertainty. *Computational Linguistics – 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 MWE 2006*, Trento, Italy: ACL, pp. 49–56.

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

Varga L. (2014). Verbe support et noms prédicatifs à l'accusatif du hongrois. In Kakoyianni-Doa F. (ed), *Penser le Lexique-Grammaire; Perspectives actuelles*, Paris: Honoré Champion, pp. 249–261.

Vincze V. (2011). *Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses*. PhD thesis, Szeged, Hungary: University of Szeged.

Vincze V. (2012). Light verb constructions in the SzegedParalellFX English–Hungarian parallel corpus. In *Proceedings of LREC 2012*, Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2381–2388.

Vincze V. and Csirik J. (2010). Hungarian corpus of light verb constructions. In *Proceedings of Coling 2010*, Beijing, China: Coling 2010 Organizing Committee, pp. 1110–1118.

Vincze V., Szauter D., Almási A., Móra G., Alexin Z. and Csirik J. (2010). Hungarian dependency Treebank. In *Proceedings of LREC 2010*, Valletta, Malta: European Language Resources Association (ELRA), pp. 1855–1862.

Vincze V., Nagy T. I. and Berend G. (2011a). Detecting noun compounds and light verb constructions: A contrastive study. In *Proceedings of MWE 2011*, Portland, Oregon, USA: Association for Computational Linguistics, pp. 116–121.

Vincze V., Nagy T. I. and Berend G. (2011b). Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of RANLP 2011*, Hissar, Bulgaria: RANLP 2011 Organising Committee, pp. 289–295.

Vincze V., Nagy T. I. and Farkas R. (2013a). Identifying English and Hungarian light verb constructions: A contrastive approach. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Association for Computational Linguistics, Sofia, Bulgaria, pp. 255–261.

Vincze V., Nagy T. I. and Zsibrita J. (2013b). Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing (TSLP)* **10**(2). https://protect-eu.mimecast.com/s/xJMuC5747UM2yDwTzuXzz?domain=dl.acm.org

Waszczuk J. (2018). TRAVERSAL at PARSEME shared task 2018: Identification of verbal multiword expressions using a discriminative tree-structured model. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 275–282.

Zampieri N., Scholivet M., Ramisch C. and Favre B. (2018). Veyn at PARSEME shared task 2018: Recurrent neural networks for VMWE identification. In *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, Santa Fe, New Mexico, USA: Association for Computational Linguistics, pp. 290–296.

Zsibrita J., Vincze V. and Farkas R. (2013). magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In *Proceedings of RANLP*, Hissar, Bulgaria: RANLP 2013 Organizing Committee, pp. 763–771.

Cite this article: Nagy TI, Rácz A and Vincze V (2020). Detecting light verb constructions across languages. *Natural Language Engineering* **26**, 319–348. https://doi.org/10.1017/S1351324919000330