Identifying English and Hungarian Light Verb Constructions: A Contrastive Approach

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

Abstract

Here, we introduce a machine learning-based approach that allows us to identify light verb constructions (LVCs) in Hungarian and English free texts. We also present the results of our experiments on the SzegedParalellFX English–Hungarian parallel corpus where LVCs were manually annotated in both languages. With our approach, we were able to contrast the performance of our method and define language-specific features for these typologically different languages. Our presented method proved to be sufficiently robust as it achieved approximately the same scores on the two typologically different languages.

1 Introduction

In natural language processing (NLP), a significant part of research is carried out on the English language. However, the investigation of languages that are typologically different from English is also essential since it can lead to innovations that might be usefully integrated into systems developed for English. Comparative approaches may also highlight some important differences among languages and the usefulness of techniques that are applied.

In this paper, we focus on the task of identifying light verb constructions (LVCs) in English and Hungarian free texts. Thus, the same task will be carried out for English and a morphologically rich language. We compare whether the same set of features can be used for both languages, we investigate the benefits of integrating language specific features into the systems and we explore how the systems could be further improved. For this purpose, we make use of the English–Hungarian parallel corpus SzegedParalellFX (Vincze, 2012), where LVCs have been manually annotated.

2 Light Verb Constructions

Light verb constructions (e.g. *to give advice*) are a subtype of multiword expressions (Sag et al., 2002). They consist of a nominal and a verbal component where the verb functions as the syntactic head, but the semantic head is the noun. The verbal component (also called a light verb) usually loses its original sense to some extent. Although it is the noun that conveys most of the meaning of the construction, the verb itself cannot be viewed as semantically bleached (Apresjan, 2004; Alonso Ramos, 2004; Sanromán Vilas, 2009) since it also adds important aspects to the meaning of the construction (for instance, the beginning of an action, such as *set on fire*, see Mel'čuk (2004)). The meaning of LVCs can be only partially computed on the basis of the meanings of their parts and the way they are related to each other, hence it is important to treat them in a special way in many NLP applications.

LVCs are usually distinguished from productive or literal verb + noun constructions on the one hand and idiomatic verb + noun expressions on the other (Fazly and Stevenson, 2007). Variativity and omitting the verb play the most significant role in distinguishing LVCs from productive constructions and idioms (Vincze, 2011). Variativity reflects the fact that LVCs can be often substituted by a verb derived from the same root as the nominal component within the construction: productive constructions and idioms can be rarely substituted by a single verb (like *make a decision – decide*). Omitting the verb exploits the fact that it is the nominal component that mostly bears the semantic content of the LVC, hence the event denoted by the construction can be determined even without the verb in most cases. Furthermore, the very same noun + verb combination may function as an LVC in certain contexts while it is just a productive construction in other ones, compare *He gave her a*

Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 255–261,
Sofia, Bulgaria, August 4-9 2013.

ring made of gold (non-LVC) and *He gave her a ring because he wanted to hear her voice* (LVC), hence it is important to identify them in context.

In theoretical linguistics, Kearns (2002) distinguishes between two subtypes of light verb constructions. True light verb constructions such as *to give a wipe* or *to have a laugh* and vague action verbs such as *to make an agreement* or *to do the ironing* differ in some syntactic and semantic features and can be separated by various tests, e.g. passivization, WH-movement, pronominalization etc. This distinction also manifests in natural language processing as several authors pay attention to the identification of just true light verb constructions, e.g. Tu and Roth (2011). However, here we do not make such a distinction and aim to identify all types of light verb constructions both in English and in Hungarian, in accordance with the annotation principles of SZPFX.

The canonical form of a Hungarian light verb construction is a bare noun + third person singular verb. However, they may occur in non-canonical versions as well: the verb may precede the noun, or the noun and the verb may be not adjacent due to the free word order. Moreover, as Hungarian is a morphologically rich language, the verb may occur in different surface forms inflected for tense, mood, person and number. These features will be paid attention to when implementing our system for detecting Hungarian LVCs.

3 Related Work

Recently, LVCs have received special interest in the NLP research community. They have been automatically identified in several languages such as English (Cook et al., 2007; Bannard, 2007; Vincze et al., 2011a; Tu and Roth, 2011), Dutch (Van de Cruys and Moirón, 2007), Basque (Gurrutxaga and Alegria, 2011) and German (Evert and Kermes, 2003).

Parallel corpora are of high importance in the automatic identification of multiword expressions: it is usually one-to-many correspondence that is exploited when designing methods for detecting multiword expressions. Caseli et al. (2010) developed an alignment-based method for extracting multiword expressions from Portuguese–English parallel corpora. Samardžić and Merlo (2010) analyzed English and German light verb constructions in parallel corpora: they pay special attention to their manual and automatic alignment. Zarrieß and Kuhn (2009) argued that multiword expressions can be reliably detected in parallel corpora by using dependency-parsed, word-aligned sentences. Sinha (2009) detected Hindi complex predicates (i.e. a combination of a light verb and a noun, a verb or an adjective) in a Hindi–English parallel corpus by identifying a mismatch of the Hindi light verb meaning in the aligned English sentence. Many-to-one correspondences were also exploited by Attia et al. (2010) when identifying Arabic multiword expressions relying on asymmetries between paralell entry titles of Wikipedia. Tsvetkov and Wintner (2010) identified Hebrew multiword expressions by searching for misalignments in an English–Hebrew parallel corpus.

To the best of our knowledge, parallel corpora have not been used for testing the efficiency of an MWE-detecting method for two languages at the same time. Here, we investigate the performance of our base LVC-detector on English and Hungarian and pay special attention to the added value of language-specific features.

4 Experiments

In our investigations we made use of the Szeged-ParalellFX English-Hungarian parallel corpus, which consists of 14,000 sentences and contains about 1370 LVCs for each language. In addition, we are aware of two other corpora – the Szeged Treebank (Vincze and Csirik, 2010) and Wiki50 (Vincze et al., 2011b) –, which were manually annotated for LVCs on the basis of similar principles as SZPFX, so we exploited these corpora when defining our features.

To automatically identify LVCs in running texts, a machine learning based approach was applied. This method first parsed each sentence and extracted potential LVCs. Afterwards, a binary classification method was utilized, which can automatically classify potential LVCs as an LVC or not. This binary classifier was based on a rich feature set described below.

The candidate extraction method investigated the dependency relation among the verbs and nouns. Verb-object, verb-subject, verb-prepositional object, verb-other argument (in the case of Hungarian) and noun-modifier pairs were collected from the texts. The dependency labels were provided by the Bohnet parser (Bohnet, 2010) for English and by `magyarlanc 2.0` (Zsibrita et al., 2013) for Hungarian.

The features used by the binary classifier can be categorised as follows:

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 investigates the POS-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*. The English **auxiliary** verbs, *do* and *have* often occur as light verbs, hence we defined a feature for the two verbs to denote whether or not they were auxiliary verbs in a given sentence.The POS code of the next word of LVC candidate was also applied as a feature. As Hungarian is a morphologically rich language, we were able to define various morphology-based features like the case of the noun or its number etc. Nouns which were historically derived from verbs but were not treated as derivation by the Hungarian morphological parser were also added as a feature.

Semantic features: This feature also exploited the fact that the nominal component is usually derived from verbs. Consequently, the `activity` or `event` semantic senses were looked for among the upper level hyperonyms of the head of the noun phrase in English WordNet 3.1[1] and in the Hungarian WordNet (Miháltz et al., 2008).

Orthographic features: The **suffix** feature is also based on the fact that many nominal components in LVCs are derived from verbs. This feature checks whether the lemma of the noun ended in a given character bi- or trigram. 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 when the candidate was one of the syntactic phrases collected.

Lexical features: We exploit the fact that the **most common verbs** are typically light verbs. Therefore, fifteen typical light verbs were selected from the list of the most frequent verbs taken from the Wiki50 (Vincze et al., 2011b) in the case of English and from the Szeged Treebank (Vincze and Csirik, 2010) in the case of Hungarian. Then, we investigated whether the lemmatised verbal component of the candidate was one of these fifteen verbs. The **lemma of the noun** was also applied as a lexical feature. The nouns found in LVCs were collected from the above-mentioned corpora. Afterwards, we constructed **lists of lemmatised LVCs** got from the other corpora.

Syntactic features: As the candidate extraction methods basically depended on the **dependency relation** between the noun and the verb, they could also be utilised in identifying LVCs. Though the `dobj`, `prep`, `rcmod`, `partmod` or `nsubjpass` dependency labels were used in candidate extraction in the case of English, these syntactic relations were defined as features, while the `att`, `obj`, `obl`, `subj` dependency relations were used in the case of Hungarian. When the noun had a **determiner** in the candidate LVC, it was also encoded as another syntactic feature.

Our feature set includes language-independent and language-specific features as well. Language-independent features seek to acquire general features of LVCs while language-specific features can be applied due to the different grammatical characteristics of the two languages or due to the availability of different resources. Table 1 shows which features were applied for which language.

We experimented with several learning algorithms and decision trees have been proven performing best. This is probably due to the fact that our feature set consists of compact – i.e. high-level – features. We trained the J48 classifier of the WEKA package (Hall et al., 2009). This machine learning approach implements the decision trees algorithm C4.5 (Quinlan, 1993). The J48 classifier was trained with the above-mentioned features and we evaluated it in a 10-fold cross validation.

The potential LVCs which are extracted by the candidate extraction method but not marked as positive in the gold standard were classed as negative. As just the positive LVCs were annotated on the SZPFX corpus, the $F_{\beta=1}$ score interpreted on the positive class was employed as an evaluation metric. The candidate extraction methods could not detect all LVCs in the corpus data, so some positive elements in the corpora were not covered. Hence, we regarded the omitted LVCs as false negatives in our evaluation.

[1] `http://wordnet.princeton.edu`

Features	Base	English	Hungarian
Orthographical	•	–	–
VerbalStem	•	–	–
POS pattern	•	–	–
LVC list	•	–	–
Light verb list	•	–	–
Semantic features	•	–	–
Syntactic features	•	–	–
Auxiliary verb	–	•	–
Determiner	–	•	–
Noun list	–	•	–
POS After	–	•	–
LVC freq. stat.	–	•	–
Agglutinative morph.	–	–	•
Historical derivation	–	–	•

Table 1: The basic feature set and language-specific features.

	English	Hungarian
ML	63.29/56.91/59.93	66.1/50.04/56.96
DM	73.71/29.22/41.67	63.24/34.46/44.59

Table 2: Results obtained in terms of precision, recall and F-score. ML: machine learning approach DM: dictionary matching method.

5 Results

As a baseline, a context free dictionary matching method was applied. For this, the gold-standard LVC lemmas were gathered from Wiki50 and the Szeged Treebank. Texts were lemmatized and if an item on the list was found in the text, it was treated as an LVC.

Table 2 lists the results got on the two different parts of SZPFX using the machine learning-based approach and the baseline dictionary matching. The dictionary matching approach yielded the highest precision on the English part of SZPFX, namely 73.71%. However, the machine learning-based approach proved to be the most successful as it achieved an F-score that was 18.26 higher than that with dictionary matching. Hence, this method turned out to be more effective regarding recall. At the same time, the machine learning and dictionary matching methods got roughly the same precision score on the Hungarian part of SZPFX, but again the machine learning-based approach achieved the best F-score. While in the case of English the dictionary matching method got a higher precision score, the machine learning approach proved to be more effective.

An ablation analysis was carried out to examine the effectiveness of each individual feature of the machine learning-based candidate classification. For each feature type, a J48 classifier was trained with all of the features except that one. We also investigated how language-specific features improved the performance compared to the base feature set. We then compared the performance to that got with all the features. Table 3 shows the contribution of each individual feature type on the SZPFX corpus. For each of the two languages, each type of feature contributed to the overall performance. Lexical features were very effective in both languages.

Feature	English	Hungarian
All	59.93	56.96
Lexical	-19.11	-14.05
Morphological	-1.68	-1.75
Orthographic	-0.43	-3.31
Syntactic	-1.84	-1.28
Semantic	-2.17	-0.34
Statistical	-2.23	–
Language-specific	-1.83	-1.05

Table 3: The usefulness of individual features in terms of F-score using the SZPFX corpus.

6 Discussion

According to the results, our base system is robust enough to achieve approximately the same results on two typologically different languages. Language-specific features further contribute to the performance as shown by the ablation analysis. It should be also mentioned that some of the base features (e.g. POS-patterns, which we thought would be useful for English due to the fixed word order) were originally inspired by one of the languages and later expanded to the other one (i.e. they were included in the base feature set) since it was also effective in the case of the other language. Thus, a multilingual approach may be also beneficial in the case of monolingual applications as well.

The most obvious difference between the performances on the two languages is the recall scores (the difference being 6.87 percentage points between the two languages). This may be related to the fact that the distribution of light verbs is quite different in the two languages. While the top 15 verbs covers more than 80% of the English LVCs, in Hungarian, this number is only 63% (and in order to reach the same coverage, 38 verbs should be included). Another difference is that there are 102

different verbs in English, which follow the Zipf distribution, on the other hand, there are 157 Hungarian verbs with a more balanced distributional pattern. Thus, fewer verbs cover a greater part of LVCs in English than in Hungarian and this also explains why lexical features contribute more to the overall performance in English. This fact also indicates that if verb lists are further extended, still better recall scores may be achieved for both languages.

As for the effectiveness of morphological and syntactic features, morphological features perform better on a language with a rich morphological representation (Hungarian). However, syntax plays a more important role in LVC detection in English: the added value of syntax is higher for the English corpora than for the Hungarian one, where syntactic features are also encoded in suffixes, i.e. morphological information.

We carried out an error analysis in order to see how our system could be further improved and the errors reduced. We concluded that there were some general and language-specific errors as well. Among the general errors, we found that LVCs with a rare light verb were difficult to recognize (e.g. *to utter a lie*). In other cases, an originally deverbal noun was used in a lexicalised sense together with a typical light verb ((e.g. *buildings are given (something)*) and these candidates were falsely classed as LVCs. Also, some errors in POS-tagging or dependency parsing also led to some erroneous predictions.

As for language-specific errors, English verb-particle combinations (VPCs) followed by a noun were often labeled as LVCs such as *make up his mind* or *give in his notice*. In Hungarian, verb + proper noun constructions (*Hamletet játsszák* (Hamlet-ACC play-3PL.DEF) "they are playing Hamlet") were sometimes regarded as LVCs since the morphological analysis does not make a distinction between proper and common nouns. These language-specific errors may be eliminated by integrating a VPC detector and a named entity recognition system into the English and Hungarian systems, respectively.

Although there has been a considerable amount of literature on English LVC identification (see Section 3), our results are not directly comparable to them. This may be explained by the fact that different authors aimed to identify a different scope of linguistic phenomena and thus interpreted the concept of "light verb construction" slightly differently. For instance, Tu and Roth (2011) and Tan et al. (2006) focused only on true light verb constructions while only object–verb pairs are considered in other studies (Stevenson et al., 2004; Tan et al., 2006; Fazly and Stevenson, 2007; Cook et al., 2007; Bannard, 2007; Tu and Roth, 2011). Several other studies report results only on light verb constructions formed with certain light verbs (Stevenson et al., 2004; Tan et al., 2006; Tu and Roth, 2011). In contrast, we aimed to identify all kinds of LVCs, i.e. we did not apply any restrictions on the nature of LVCs to be detected. In other words, our task was somewhat more difficult than those found in earlier literature. Although our results are somewhat lower on English LVC detection than those attained by previous studies, we think that despite the difficulty of the task, our method could offer promising results for identifying all types of LVCs both in English and in Hungarian.

7 Conclusions

In this paper, we introduced our machine learning-based approach for identifying LVCs in Hungarian and English free texts. The method proved to be sufficiently robust as it achieved approximately the same scores on two typologically different languages. The language-specific features further contributed to the performance in both languages. In addition, some language-independent features were inspired by one of the languages, so a multilingual approach proved to be fruitful in the case of monolingual LVC detection as well.

In the future, we would like to improve our system by conducting a detailed analysis of the effect of each feature on the results. Later, we also plan to adapt the tool to other types of multiword expressions and conduct further experiments on languages other than English and Hungarian, the results of which may further lead to a more robust, general LVC system. Moreover, we can improve the method applied in each language by implementing other language-specific features as well.

Acknowledgments

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

References

Margarita Alonso Ramos. 2004. *Las construcciones con verbo de apoyo*. Visor Libros, Madrid.

Jurij D. Apresjan. 2004. O semantičeskoj nepustote i motivirovannosti glagol'nyx leksičeskix funkcij. *Voprosy jazykoznanija*, (4):3–18.

Mohammed Attia, Antonio Toral, Lamia Tounsi, Pavel Pecina, and Josef van Genabith. 2010. Automatic Extraction of Arabic Multiword Expressions. In *Proceedings of the 2010 Workshop on Multiword Expressions: from Theory to Applications*, pages 19–27, Beijing, China, August. Coling 2010 Organizing Committee.

Colin Bannard. 2007. A measure of syntactic flexibility for automatically identifying multiword expressions in corpora. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 1–8, Morristown, NJ, USA. Association for Computational Linguistics.

Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 89–97.

Helena de Medeiros Caseli, Carlos Ramisch, Maria das Graças Volpe Nunes, and Aline Villavicencio. 2010. Alignment-based extraction of multiword expressions. *Language Resources and Evaluation*, 44(1-2):59–77.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2007. Pulling their weight: exploiting syntactic forms for the automatic identification of idiomatic expressions in context. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 41–48, Morristown, NJ, USA. Association for Computational Linguistics.

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

Afsaneh Fazly and Suzanne Stevenson. 2007. Distinguishing Subtypes of Multiword Expressions Using Linguistically-Motivated Statistical Measures. In *Proceedings of the Workshop on A Broader Perspective on Multiword Expressions*, pages 9–16, Prague, Czech Republic, June. Association for Computational Linguistics.

Antton Gurrutxaga and Iñaki Alegria. 2011. Automatic Extraction of NV Expressions in Basque: Basic Issues on Cooccurrence Techniques. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 2–7, Portland, Oregon, USA, June. Association for Computational Linguistics.

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

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

Igor Mel'čuk. 2004. Verbes supports sans peine. *Lingvisticae Investigationes*, 27(2):203–217.

Márton Miháltz, Csaba Hatvani, Judit Kuti, György Szarvas, János Csirik, Gábor Prószéky, and Tamás Váradi. 2008. Methods and Results of the Hungarian WordNet Project. In Attila Tanács, Dóra Csendes, Veronika Vincze, Christiane Fellbaum, and Piek Vossen, editors, *Proceedings of the Fourth Global WordNet Conference (GWC 2008)*, pages 311–320, Szeged. University of Szeged.

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

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002*, pages 1–15, Mexico City, Mexico.

Tanja Samardžić and Paola Merlo. 2010. Cross-lingual variation of light verb constructions: Using parallel corpora and automatic alignment for linguistic research. In *Proceedings of the 2010 Workshop on NLP and Linguistics: Finding the Common Ground*, pages 52–60, Uppsala, Sweden, July. Association for Computational Linguistics.

Begoña Sanromán Vilas. 2009. Towards a semantically oriented selection of the values of Oper_1. The case of *golpe* 'blow' in Spanish. In David Beck, Kim Gerdes, Jasmina Milićević, and Alain Polguère, editors, *Proceedings of the Fourth International Conference on Meaning-Text Theory – MTT'09*, pages 327–337, Montreal, Canada. Université de Montréal.

R. Mahesh K. Sinha. 2009. Mining Complex Predicates In Hindi Using A Parallel Hindi-English Corpus. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 40–46, Singapore, August. Association for Computational Linguistics.

Suzanne Stevenson, Afsaneh Fazly, and Ryan North. 2004. Statistical Measures of the Semi-Productivity of Light Verb Constructions. In Takaaki Tanaka, Aline Villavicencio, Francis Bond, and Anna Korhonen, editors, *Second ACL Workshop on Multiword Expressions: Integrating Processing*, pages 1–8, Barcelona, Spain, July. Association for Computational Linguistics.

Yee Fan Tan, Min-Yen Kan, and Hang Cui. 2006. Extending corpus-based identification of light verb constructions using a supervised learning framework. In *Proceedings of the EACL Workshop on*

Multi-Word Expressions in a Multilingual Contexts, pages 49–56, Trento, Italy, April. Association for Computational Linguistics.

Yulia Tsvetkov and Shuly Wintner. 2010. Extraction of multi-word expressions from small parallel corpora. In *Coling 2010: Posters*, pages 1256–1264, Beijing, China, August. Coling 2010 Organizing Committee.

Yuancheng Tu and Dan Roth. 2011. Learning English Light Verb Constructions: Contextual or Statistical. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 31–39, Portland, Oregon, USA, June. Association for Computational Linguistics.

Tim Van de Cruys and Begoña Villada Moirón. 2007. Semantics-based multiword expression extraction. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions*, MWE '07, pages 25–32, Morristown, NJ, USA. Association for Computational Linguistics.

Veronika Vincze and János Csirik. 2010. Hungarian corpus of light verb constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1110–1118, Beijing, China, August. Coling 2010 Organizing Committee.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011a. Detecting Noun Compounds and Light Verb Constructions: a Contrastive Study. In *Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World*, pages 116–121, Portland, Oregon, USA, June. ACL.

Veronika Vincze, István Nagy T., and Gábor Berend. 2011b. Multiword expressions and named entities in the Wiki50 corpus. In *Proceedings of RANLP 2011*, Hissar, Bulgaria.

Veronika Vincze. 2011. *Semi-Compositional Noun + Verb Constructions: Theoretical Questions and Computational Linguistic Analyses*. Ph.D. thesis, University of Szeged, Szeged, Hungary.

Veronika Vincze. 2012. Light Verb Constructions in the SzegedParalellFX English–Hungarian Parallel Corpus. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, May. European Language Resources Association (ELRA).

Sina Zarrieß and Jonas Kuhn. 2009. Exploiting Translational Correspondences for Pattern-Independent MWE Identification. In *Proceedings of the Workshop on Multiword Expressions: Identification, Interpretation, Disambiguation and Applications*, pages 23–30, Singapore, August. Association for Computational Linguistics.

János Zsibrita, Veronika Vincze, and Richárd Farkas. 2013. magyarlanc 2.0: szintaktikai elemzés és felgyorsított szófaji egyértelműsítés [magyarlanc 2.0: Syntactic parsing and accelerated POS-tagging]. In Attila Tanács and Veronika Vincze, editors, *MSzNy 2013 – IX. Magyar Számítógépes Nyelvészeti Konferencia*, pages 368–374, Szeged. Szegedi Tudományegyetem.