Dependency Relations of Light Verb Constructions in Four Languages

Veronika Vincze
Hungarian Academy of Sciences,
Research Group on Artificial Intelligence
vinczev@inf.u-szeged.hu

István Nagy T.
Department of Informatics,
University of Szeged
nistvan@inf.u-szeged.hu

1 Introduction

Some of the most important questions concerning MWEs is how they should be represented at different levels of grammar and/or in different NLP resources. To answer these questions, a thorough examination of the features of MWEs seems indispensable. Here, we will focus on light verb constructions (LVCs) and examine their syntactic characteristics in four different languages. We will standardize the different representations of the various syntactic relations for the same grammatical relation across the languages and we will argue that the standardized representation may prove useful both in the automatic detection of MWEs and in treebank annotation.

2 Dependency Relations

As LVCs were manually annotated in the 4FX parallel corpus (Rácz et al., 2014) in four different languages, namely English, German, Spanish and Hungarian, we were able to examine the typical dependency relations of the LVCs in these four different languages.

Since we had no manual syntactic annotation for the 4FX corpus, we had to dependency parse the data in order to examine the syntactic relations among the LVCs' verbal and nominal components. 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 data (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. For Hungarian, the state-of-the-art dependency parser, `magyarlanc 2.0` (Zsibrita et al., 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. In order to compare the dependency labels of LVCs in different languages, 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. We also had to pay attention to the fact that prepositional phrases in English, Spanish and German usually correspond to the combination of a noun and an oblique case suffix in Hungarian, which were also lumped into one standardized category. Table 1 shows the distribution of standardized dependency label types on the four languages in the 4FX parallel corpus.

In some cases, the parser was not able to find a direct dependency edge between the verbal and nominal components. Some of these cases are due to parsing errors but in other cases, the nominal component was part of the object, preceded by a quantifying expression like *he **gained** much of his **fame*** so there is no direct link between the verb and the noun. In other cases, there was a rare and atypical syntactic relation between the noun and the verb (e.g. `dep` in *reach conform*), so these cases are merged into the *other* class in the table.

3 Extracting Potential LVCs

In the literature on the automatic detection of LVCs, several earlier approaches applied a syntactic restriction and treated only the `verb-object` pairs as potential LVCs, cf. Vincze et al. (2013a). But, as Table 1 shows, only 44.30% of English LVCs, 32.84% of German LVCs, 31.89% of Spanish LVCs and 32.19% in Hungarian LVCs have *verb-direct object* relations, respectively. On the contrary, manually annotated LVCs in the 4FX corpus have other LVC specific dependency labels like *verb-(passive) subject*, *verb-adpositional*, and

Type	English		German		Spanish		Hungarian	
	#	%	#	%	#	%	#	%
object	280	44.30%	213	32.87%	273	31.89%	272	32.19%
adpositional phrase	101	15.98%	35	5.40%	164	19.16%	263	31.12%
subject	81	12.82%	28	4.32%	45	5.26%	34	4.02%
participial modifier	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%**

Table 1: Edge types in the 4FX corpus.

noun-participial modifier and if they are also accounted for, we can cover 83.23% of English, 60.80% of German, 70.56% of Spanish and 87.10% of Hungarian LVCs, respectively. Thus, on the basis of empirical data from four languages, we argue that we should drop any syntactic restrictions while automatically extracting LVC candidates from texts and all the noun and verb combinations exhibiting the above mentioned dependency relations can be potential LVC candidates.

4 Constructing Treebanks

Our findings on the dependency labels of LVCs may be also useful from the annotation viewpoint. When aiming at annotating MWEs in treebanks, there are two basic options. First, a separate MWE (or LVC) label could be applied to mark the (syntactic) relationship among the members of the MWE, e.g. in *make a decision*, there would be an `LVC` label in between *decision* and *make*. This option might be viable when LVCs show a uniform behaviour, that is, most of them belong to one syntactic phrase type (e.g. verb-object). Second, a standard label could be used and the MWE-ness of the phrase could be marked at a separate layer of annotation: there would be an `object` label in the previous case at the syntactic layer and another LVC label at another layer (or a complex `object-LVC` label would also be plausible, as applied in the Hungarian treebank (Vincze et al., 2013b)). An advantage of this representation is that the inner structure of MWEs is also made transparent, which might be important when e.g. the modifiability of MWEs is under question.

As our data revealed, there are quite a number of dependency relation types among the members of LVCs in each language. Hence, we propose that this diversity should be preserved and both the inner syntactic structure and the MWE-ness of the phrase should be marked, either at separate layers or at the syntactic layer with a complex label.

References

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

Sabine Brants, Stefanie Dipper, Peter Eisenberg, Silvia Hansen-Schirra, Esther König, Wolfgang Lezius, Christian Rohrer, George Smith, and Hans Uszkoreit. 2004. TIGER: Linguistic interpretation of a German corpus. *Research on Language and Computation*, 2(4):597–620.

Montserrat Marimon, Beatríz Fisas, Núria Bel, Marta Villegas, Jorge Vivaldi, Sergi Torner, Mercè Lorente, Silvia Vázquez, and Marta Villegas. 2012. The IULA Treebank. In *Proceedings of LREC-2012*, pages 1920–1926, Istanbul, Turkey. ELRA.

Anita Rácz, István Nagy T., and Veronika Vincze. 2014. 4FX: Light Verb Constructions in a Multilingual Parallel Corpus. In *Proceedings of LREC'14*, Reykjavik, Iceland. ELRA.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the CoNLL-2008*, pages 159–177. ACL.

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

Veronika Vincze, István Nagy T., and János Zsibrita. 2013a. Learning to detect English and Hungarian light verb constructions. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(2).

Veronika Vincze, János Zsibrita, and István Nagy T. 2013b. Dependency Parsing for Identifying Hungarian Light Verb Constructions. In *Proceedings of IJCNLP 2013*, pages 207–215, Nagoya, Japan.

János Zsibrita, Veronika Vincze, and Richárd Farkas. 2013. magyarlanc: A toolkit for morphological and dependency parsing of Hungarian. In *Proceedings of RANLP*, pages 763–771.