Light Verb Constructions in Universal Dependencies

Joakim Nivre
Dept. of Linguistics and Philology
Uppsala University
joakim.nivre@lingfil.uu.se

Veronika Vincze
Department of Informatics
University of Szeged
vinczev@inf.u-szeged.hu

1 Introduction

Universal Dependencies (UD) is an initiative to develop cross-linguistically consistent treebank annotation for many languages, with the goal of facilitating multilingual parser development, cross-lingual learning, and research on parsing from a language typology perspective (Zeman, 2008; Petrov et al., 2012; de Marneffe et al., 2014; Nivre, 2015). The latest UD release (v1.1) contains treebanks for 18 languages.

Light verb constructions (LVCs) pose interesting challenges for linguistic annotation, especially from a cross-linguistic perspective. The goal of this paper is to make a survey of the different ways in which LVCs are analyzed in UD v1.1. Our hope is that this will lead to a better understanding of the role of LVCs in different languages and ultimately lead to better guidelines for their analysis.

2 Annotation of LVCs in UD

Our initial survey focuses on constructions of the type illustrated by the sentence *She took a photo of the cathedral*. The LVC in this sentence consists of the light verb *take* and the noun *photo*, which is semantically equivalent to the transitive verb *photograph*. Syntactically, *photo* is the direct object of *take*, and *photo* is modified by the prepositional phrase *of the cathedral*. Semantically, however, *cathedral* is rather an argument of the complex predicate *take a photo* (corresponding to the direct object of the verb *photograph*). Our survey shows that the 18 treebanks can be divided into three groups:

1. Treebanks that do not distinguish LVCs.
2. Treebanks that distinguish LVCs only by their structure.
3. Treebanks that distinguish LVCs (also) by special labels.

Group 1: Figure 1 (top) shows the analysis of the English sentence *Take a photo of a very light plain subject close to the lens*. Here, *take a photo* is an LVC and *of a very light plain subject close to the lens* is a semantic argument of the complex predicate. However, the annotation treats the object and the prepositional argument as ordinary syntactic arguments, *photo* being attached to the verb as a direct object (`dobj`) and *subject* to *photo* as a nominal modifier (`nmod`). This analysis is used in most UD treebanks.

Group 2: Figure 1 (middle) exhibits the Swedish sentence *Den här broschyren vill ge dig en bild av din militärutbildning* (this here brochure wants give you a picture of your military-education) “This brochure is meant to give you a picture of your military education”. The LVC *ge en bild* “give a picture” is analyzed as an ordinary transitive verb with a direct object, but the semantic argument *av din militärutbildning* “of your military education”

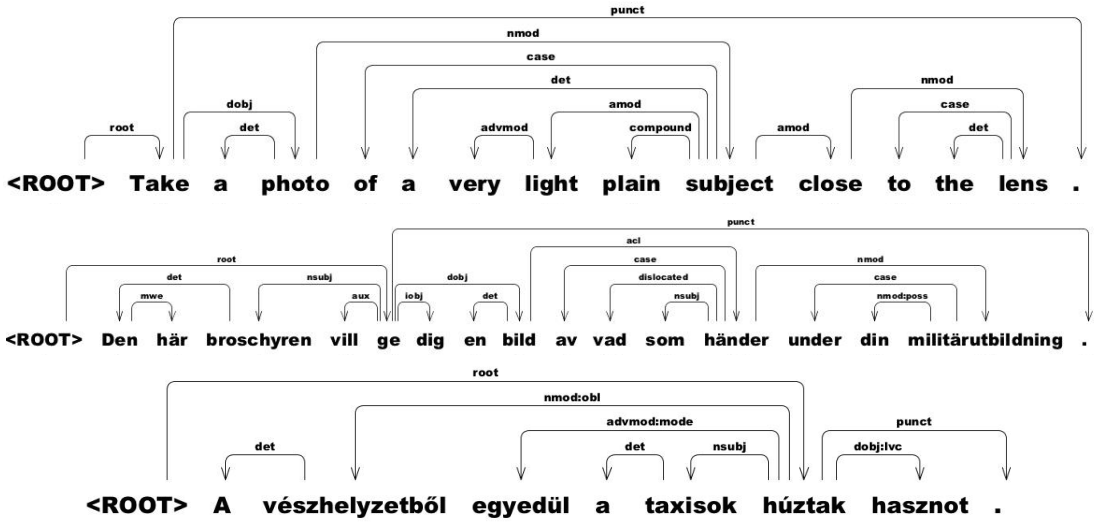

Figure 1: Annotation of LVCs in Universal Dependencies v1.1.

is attached to the verb, not to the object, to indicate that it is a dependent of the entire LVC. This analysis is used in Swedish, German, Irish and (sometimes) in French.

Group 3: In Figure 1 (bottom), we see the Hungarian sentence *A vészhelyzetből egyedül a taxisok húztak hasznot* (the emergency-ELA only the taxi.driver-PL draw-PAST-3PL advantage-ACC) "Only the taxi drivers took advantage of the emergency". The LVC *hasznot húz* "take advantage" consists of the verb *húz* "draw" and its object *hasznot* "advantage-ACC", and the dependency label `dobj:lvc` denotes that the two form an LVC where the nominal component is a direct object. The argument *vészhelyzetből* "of emergency" is, however, attached to the verb as a nominal modifier (`nmod:obl`). Thus, the structure is the same as in the Swedish example, but the label `dobj:lvc` explicitates the status of the LVC. This analysis is used in Hungarian and Persian.

3 Conclusion

We have presented a survey of how LVCs are annotated in Universal Dependencies v1.1, focusing on LVCs consisting of a transitive verb and a direct object, complemented by a prepositional phrase. In the final version, we will include data from all 18 languages/treebanks.

References

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. Universal Stanford Dependencies: A cross-linguistic typology. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 4585–4592.

Joakim Nivre. 2015. Towards a universal grammar for natural language processing. In *Computational Linguistics and Intelligent Text Processing*. Springer.

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*.

Daniel Zeman. 2008. Reusable tagset conversion using tagset drivers. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC)*, pages 213–218.