Verb-Noun Idiomatic Combinations in a Czech-English Dependency Corpus

Extended poster proposal abstract based on a paper presented at WMWE, at NAACL 2013

Zdeňka Urešová, Jana Šindlerová, Eva Fučíková and Jan Hajič
Charles University in Prague[1]
{uresova,sindlerova,fucikova,hajic}@ufal.mff.cuni.cz

Introduction

While working on valency lexicons for Czech and English, it was necessary to define treatment of multiword entities (MWEs) with the verb as the central lexical unit. Morphological, syntactic and semantic properties of such MWEs had to be formally specified to fit into lexicon entries and be usable in treebank annotation. Such a formal specification has also been used for automated quality control of the annotation vs. the lexicon entries. We present a corpus-based study, concentrating on multilayer specification of verbal MWEs, their properties in Czech and English, and a comparison between the two languages using the parallel Czech-English Dependency Treebank (PCEDT, Hajic et al., 2012). This comparison revealed interesting differences in the use of verbal MWEs in translation (discovering that such MWEs are actually rarely translated as MWEs, at least between Czech and English) as well as some inconsistencies in their annotation. Since Czech and English are typologically different languages, we believe that our findings will also contribute to a better understanding of verbal MWEs and possibly their more unified treatment across languages.

Valency and MWEs

Valency as a linguistic phenomenon plays a crucial role in the majority of today's linguistic theories and may be considered a base for both lexicographical and grammatical work. Our approach to valency is based on the theoretical framework of Functional Generative Description (Sgall et al.,1986), and is adapted for data-oriented tasks (Urešová 2011a; 2011b). In general, valency is understood as a specific ability of certain lexical units - primarily of verbs - to open "slots" to be filled in by other lexical units. MWEs are expressions which consist of more than a single word while having non-compositional meaning. They can be described (Sag et al., 2002) as "idiosyncratic interpretations that cross word boundaries." For the purpose of our work, we have focused on Verb-Noun Idiomatic Constructions (VNICs).

Verb-noun idiomatic combinations (VNICs) in Czech and English

To compare the annotation and use of VNICs in Czech and English, we have used the PCEDT. We found a total of 92890 occurrences of aligned (non-auxiliary) verbs. Interestingly, the annotation contained only 88 occurences of VNICs marked[2] on both sides of the cross-lingual alignment. Czech VNICs were aligned with English counterparts not annotated as a VNIC in 570 cases, and there were 278 occurrences of English VNICs aligned with Czech non-VNICs. We concentrated on cases where the verb in the English original has been annotated as VNIC, but the

[1] Authors' full address: Institute of Formal and Applied Linguistics, Charles University in Prague, Faculty of Mathematics and Physics, Malostranske nam. 25, 11800 Prague 1,Czech Republic
[2] In the PCEDT, VNICs are marked by the virtual functor DPHR, both in the data and in the valency lexicons.

Czech translation has been marked as a non-VNIC, and on cases where Czech VNICs were linked to non-VNIC in English.

Main findings and conclusions/future work

We have explored the PCEDT to find interesting correspondences between the annotation and lexicon entries in the English and Czech annotation schemes and the corresponding lexicons. A translation of a VNIC as a VNIC is rare, even if we take into account the annotation errors. By far the most common case of translating a VNIC in both directions is the usage of a completely non-MWE phrase. While the low overall number of VNICs found in the parallel corpus may be explained by not including phrasal verbs proper in our study, we can only speculate why only a small proportion of VNICs are translated as VNICs in(to) the other language: it seems that an explanation may be grounded in historical linguistic perspective.

Exploring the whole corpus in the future, we should be able to get a more reliable material for a thorough study of the use of MWEs in translation, with the aim of improving identification and analysis of MWEs, e.g., by enriching the approach taken by and described in (Bejcek et al., 2013). We would also like to improve machine translation results by identifying relevant features of MWEs (including but not limited to VNICs) and using the associated information stored in the valency lexicons in order to learn translation correspondences involving MWEs.

References

Bejček Eduard, Straňák Pavel, Pecina Pavel: Syntactic Identification of Occurrences of Multiword Expressions in Text using a Lexicon with Dependency Structures. In: *The 9th Workshop on Multiword Expressions (MWE 2013)*, Copyright © Association for Computational Linguistics, Atlanta, Georgia, USA, ISBN 978-1-937284-47-3, pp. 106-115, 2013

Hajič Jan, Hajičová Eva, Panevová Jarmila, Sgall Petr, Bojar Ondřej, Cinková Silvie, Fučíková Eva, Mikulová Marie, Pajas Petr, Popelka Jan, Semecký Jiří, Šindlerová Jana, Štěpánek Jan, Toman Josef, Urešová Zdeňka, Žabokrtský Zdeněk: Announcing Prague Czech-English Dependency Treebank 2.0. In: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, Copyright © European Language Resources Association, İstanbul, Turkey, ISBN 978-2-9517408-7-7, pp. 3153-3160, 2012

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

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. The Meaning of the Sentence in Its Semantic and Pragmatic Aspects. Dordrecht, Reidel, and Prague, Academia.

Zdeňka Urešová. 2011a. Valence sloves v Pražském závislostním korpusu. Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Prague.

Zdeňka Urešová. 2011b. Valenční slovník Pražského závislostního korpusu (PDT-Vallex). Studies in Computational and Theoretical Linguistics. Ústav formální a aplikované lingvistiky, Prague.